Information and Inference: A Journal of the IMA (2023) **12**, 2423–2502
https://doi.org/10.1093/imaiai/iaad034
Advance Access publication on 9 October 2023

Minimax optimal regression over Sobolev spaces via Laplacian Eigenmaps on neighbourhood graphs

Alden Green*
Department of Statistics, Stanford University, Stanford, CA 94305, USA
*Corresponding author: aldenjg@stanford.edu

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

and

Ryan J. Tibshirani
Department of Statistics, University of California, Berkeley, Berkeley, CA 94720, USA

[Received on 31 December 2021; revised on 21 April 2023; accepted on 7 June 2023]

In this paper, we study the statistical properties of Principal Components Regression with Laplacian Eigenmaps (PCR-LE), a method for non-parametric regression based on Laplacian Eigenmaps (LE). PCR-LE works by projecting a vector of observed responses $\mathbf{Y} = (Y_1, \ldots, Y_n)$ onto a subspace spanned by certain eigenvectors of a neighbourhood graph Laplacian. We show that PCR-LE achieves minimax rates of convergence for random design regression over Sobolev spaces. Under sufficient smoothness conditions on the design density p, PCR-LE achieves the optimal rates for both estimation (where the optimal rate in squared L^2 norm is known to be $n^{-2s/(2s+d)}$) and goodness-of-fit testing ($n^{-4s/(4s+d)}$). We also consider the situation where the design is supported on a manifold of small intrinsic dimension m, and give upper bounds establishing that PCR-LE achieves the faster minimax estimation ($n^{-2s/(2s+m)}$) and testing ($n^{-4s/(4s+m)}$) rates of convergence. Interestingly, these rates are almost always much faster than the known rates of convergence of graph Laplacian eigenvectors to their population-level limits; in other words, for this problem regression with estimated features appears to be much easier, statistically speaking, than estimating the features itself. We support these theoretical results with empirical evidence.

Keywords: Laplacian Eigenmaps; non-parametric regression; principal components regression; Sobolev space; minimax rates.

1. Introduction

Laplacian Eigenmaps (LE) [8] is a method for nonlinear dimensionality reduction and data representation. Given data points $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$, LE maps each X_i to a vector $(v_{1,i}, \ldots, v_{K,i})$ according to the following steps.

1. First, LE forms a *neighbourhood graph* $G = (V, W)$ over the points $\{X_1, \ldots, X_n\}$. The graph G is an undirected, weighted graph, with vertices $V = \{X_1, \ldots, X_n\}$, and weighted edges W_{ij} which correspond to the proximity between points X_i and X_j.

2. Next, LE forms an (unweighted) *graph Laplacian* matrix $L \in \mathbb{R}^{n\times n}$, a symmetric and diagonally dominant matrix with diagonal elements $L_{ii} = \sum_{j=1}^n W_{ij}$, and off-diagonal elements $L_{ij} = -W_{ij}$.

© The Author(s) 2023. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

3. Finally, LE takes the eigendecomposition $L = \sum_{k=1}^{n} \lambda_k v_k v_k^\top$, and outputs the vectors $(v_{1,i}, \ldots, v_{K,i}) \in \mathbb{R}^K$ for each $i = 1, \ldots, n$.

A natural way to use LE is by taking the collection of vectors $\{(v_{1,i}, \ldots, v_{K,i})\}_{i=1}^n$ to be features in a downstream regression algorithm. In this paper, we study a simple method along these lines: Principal Components Regression with Laplacian Eigenmaps (PCR-LE), a method for non-parametric regression which operates by running ordinary least squares (OLS) using the features output by LE. Given pairs of design points and responses $(X_1, Y_1), \ldots, (X_n, Y_n)$, PCR-LE computes an estimate $\widehat{f} \in \mathbb{R}^n$,

$$\widehat{f} := \underset{f \in \mathrm{span}\{v_1, \ldots, v_K\}}{\mathrm{argmin}} \|\mathbf{Y} - f\|_2^2, \tag{1.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ is the vector of responses and $\|\cdot\|_2$ denotes the usual Euclidean norm in $\mathbb{R}^n$. (For a formal definition of LE and PCR-LE, see Section 2.2.)

LE has been practically very successful, and by now has been used for various statistical tasks such as spectral clustering, manifold learning, level-set estimation, semi-supervised learning, etc. At this point there exists a rich literature [9, 15, 16, 19, 23, 27, 28, 44, 59, 60, 69] explaining this practical success from a theoretical perspective. Loosely speaking, these works model the design points as being independent samples from a distribution P with density p, and show that in this case the eigenvectors of the graph Laplacian L are good empirical approximations of population-level objects. These population-level objects are eigenfunctions ψ_k—meaning solutions, along with eigenvalues ρ_k, to the equation $\Delta_P \psi_k = \rho_k \psi_k$—of a density-weighted Laplacian operator defined via

$$\Delta_P f := -\frac{1}{p} \operatorname{div}(p^2 \nabla f). \tag{1.2}$$

(Here div stands for the divergence operator, and ∇ for the gradient. See (2.4) for the formal definition of (ρ_k, ψ_k).) These eigenfunctions in turn characterize various interesting structural aspects of p, such as the location and number of high- and low-density regions, the shape and intrinsic dimension of its support and so forth.

These aforementioned works justify LE as method for data representation, by establishing that each feature vector $(v_{1,i}, \ldots, v_{K,i})$ serves an empirical approximation to an idealized representation $(\psi_1(X_i), \ldots, \psi_K(X_i))$. They also provide quantitative guarantees for the accuracy with which LE approximates this ideal representation. However, this theory does not focus on the statistical properties of PCR-LE for classical regression problems such as estimation and testing. That is the major question we address in this paper. We adopt the usual model of non-parametric regression with random design, where one observes independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ of design points and responses. We assume the design points $\{X_1, \ldots, X_n\}$ are sampled from an unknown distribution P supported on $\mathcal{X} \subseteq \mathbb{R}^d$, and the responses follow a signal plus Gaussian noise model,

$$Y_i = f_0(X_i) + w_i, \quad w_i \sim N(0, 1), \tag{1.3}$$

with noise variables w_i independent of design points X_i. The regression function f_0 is unknown but assumed to belong to a Sobolev space $H^s(\mathcal{X})$. We consider two settings: one where $\mathcal{X}$ is a full-dimensional domain, and the other where $\mathcal{X}$ is a low-dimensional submanifold of $\mathbb{R}^d$. In each setting, we derive upper bounds which imply that the PCR-LE estimate $\widehat{f}$ and a test using the statistic $\widehat{T} = \|\widehat{f}\|_2^2$ are statistically optimal methods for two classical problems in non-parametric regression: estimation and goodness-of-fit testing.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Sobolev spaces and spectral series regression. To analyse PCR-LE, we work in a classical situation where the regression function is assumed to belong to a (Hilbert-)Sobolev space. For an open domain $\mathscr{X} \subseteq \mathbb{R}^d$, the Sobolev space $H^s(\mathscr{X})$ consists of all functions $f \in L^2(\mathscr{X})$ which are s-times weakly differentiable, with all order-s partial derivatives $D^\alpha f \in L^2(\mathscr{X})$. We study regression over Sobolev spaces in part because, generally speaking, the minimax rates are well understood; as mentioned before, when the domain $\mathscr{X}$ is full-dimensional they are $n^{-2s/(2s+d)}$ for estimation, and $n^{-4s/(4s+d)}$ for testing. For this reason, regression over Sobolev spaces is a good setting to see whether PCR-LE measures up to more standard minimax optimal approaches, which have strong theoretical guarantees but are less often used in practice. We give a more specific comparison between PCR-LE and some of these more classical methods in Section 6.

We also view PCR-LE as being particularly well suited for regression over Sobolev spaces due to their close connection with *spectral series regression*. Spectral series regression computes generalized empirical Fourier coefficients $\widetilde{a}_k := \frac{1}{n}\sum_{i=1}^n Y_i \psi_k(X_i)$, and truncates to the K-lowest frequency eigenfunctions of $\varDelta_P$, producing the estimate

$$\widetilde{f}(x) = \sum_{k=1}^{K} \widetilde{a}_k \psi_k(x). \tag{1.4}$$

Spectral series regression is intrinsically linked with Sobolev spaces. That is because under appropriate boundary conditions, a ball in the order-s Sobolev space consists of functions $f = \sum_k a_k \psi_k \in L^2(\mathscr{X})$ for which the generalized Fourier coefficients $\{a_k\}_{k=1}^\infty$ satisfy the decay condition $\sum_k a_k^2 \rho_k^s \le C$ (See Section 2.3 for more details). This decay condition justifies the truncated series estimator (1.4), since it means the truncation will incur only a limited amount of bias for any $f_0 \in H^s(\mathscr{X})$. For this reason spectral series regression over Sobolev spaces has been well-studied—at least when $\mathscr{X} = [0,1]^d$—since at least Rice [51],[1] and its optimality properties are by this point generally well understood.

PCR-LE serves as an empirical approximation to spectral series regression, since as already mentioned the eigenvectors v_k are empirical approximations to the eigenfunctions ψ_k of $\varDelta_P$. Viewed in this light, a major advantage of PCR-LE is that it operates without needing knowledge of the design distribution P. This is an advantage because in our context P is an unknown and potentially complex distribution: for example, it can be highly non-uniform, have a complicated support which may be a submanifold of $\mathbb{R}^d$ or both. In contrast, spectral series regression relies on diagonalizing the density-weighted Laplacian $\varDelta_P$, and in our context must be viewed as an oracle method; to emphasize this we henceforth refer to the estimator defined in (1.4) as *population-level spectral series regression*. On the other hand, intuitively PCR-LE incurs some extra error using an empirical approximation to the underlying basis $\{\psi_k\}_{k=1}^\infty$: our work shows that in many cases, this extra error is not enough to change the overall rate of convergence.

1.1 *Main contributions*

Summarized succinctly, our main contribution is to theoretically analyse non-parametric regression with PCR-LE and establish upper bounds which imply that this method often achieves optimal rates of convergence over Sobolev spaces.

[1] And proposed much earlier in the context of density estimation by Čencov [67].

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Rates of convergence: population-level spectral series regression. As we have already mentioned, the minimax optimal rates over Sobolev spaces are generally well known, as are upper bounds for population-level spectral series methods which match these rates. However, we could not find precisely stated results applying to our setting, which is quite general in the following respects.

- We consider Sobolev spaces $H^s(\mathscr{X})$ for all combinations of s and d. This includes the subcritical regime where the smoothness parameter s satisfies $s < d/2$; in this regime $H^s(\mathscr{X})$ does not continuously embed into the space of continuous functions $C^0(\mathscr{X})$.
- We consider general design distributions P, which may satisfy certain regularity conditions but are not limited to being, say, the uniform distribution over $[0, 1]^d$.

For completeness, we analyse population-level spectral series methods in this general setting, and establish upper bounds showing that such methods converge at the ‘usual’ rates of $n^{-2s/(2s+d)}$ for estimation and $n^{-4s/(4s+d)}$ for testing. This analysis relies heavily on certain asymptotic properties of the continuum eigenfunctions ψ_k and eigenvalues ρ_k, which hold for quite general second-order differential operators $\mathscr{L}$ including the density-weighted Laplacian $\mathscr{L} = \Delta_P$.

Rates of convergence: PCR-LE. The rest of our results consist of various upper bounds on the rates of convergence for the PCR-LE estimator $\widehat{f}$, and a test using the statistic $\|\widehat{f}\|_2^2$. These upper bounds quantify how PCR-LE can take advantage of either smooth higher order derivatives, low intrinsic dimension of the design distribution or both, in an optimal manner. We consider two kinds of assumptions for the design distribution P, which we refer to as the flat Euclidean and manifold settings. In the first case we suppose the design distribution P has support $\mathscr{X}$ which is a full-dimensional set in $\mathbb{R}^d$, and that the true signal f_0 lies in a ball in the Sobolev space $H^s(\mathscr{X})$, (See Section 2.1 for the formal assumptions.) In this case our main contributions are as follows:

- We show that the PCR-LE estimator $\widehat{f}$ has in-sample mean-squared error on the order of $n^{-2s/(2s+d)}$, for any number of derivatives $s \in \mathbb{N}$ and dimension d (Theorems 1 and 3).
- We show that a test based on the statistic $\|\widehat{f}\|_2^2$, calibrated to have controlled type I error, has non-trivial power so long as the squared L^2 norm of f_0 is at least $n^{-4s/(4s+d)}$, for any number of derivatives $s \in \mathbb{N}$ and dimension $d \in \{1, 2, 3, 4\}$ (Theorems 2 and 4).

We then consider the behaviour of PCR-LE in the manifold setting, where the design distribution is supported on an (unknown) domain $\mathscr{X}$ which is a submanifold of $\mathbb{R}^d$ of intrinsic dimension $m \in \mathbb{N}, m < d$. (Again see Section 2.1 for the formal assumptions.) In this case, our main contributions are as follows:

- We show that the PCR-LE estimator $\widehat{f}$ has in-sample mean squared error of at most $n^{-2s/(2s+m)}$, when $s \in \{1, 2, 3\}$ and for any $m \in \mathbb{N}$ (Theorem 7).
- We show that a test based on the statistic $\|\widehat{f}\|_2^2$, calibrated to have controlled type I error, has non-trivial power so long as the squared L^2 norm of f_0 is at least $n^{-4s/(4s+m)}$, when $s \in \{1, 2, 3\}$ and $m \in \{1, 2, 3, 4\}$ (Theorem 8).

To the best of our knowledge, the minimax rates for non-parametric regression with random design over unknown manifolds have only been worked out for Hölder classes, and even in this case the calculations are only for $s \leq 2$ bounded derivatives [12, 73]. Our upper bounds confirm that these

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

TABLE 1 *Summary of PCR-LE estimation rates over Sobolev balls. The bold font marks minimax optimal rates. In each case, rates hold for all $d \in \mathbb{N}$ (in the flat Euclidean setting), and for all $m \in \mathbb{N}, 1 < m < d$ (in the manifold setting). Although we suppress it for simplicity, in all cases when the PCR-LE estimator is optimal, the dependence of the error rate on the radius M of the Sobolev ball is also optimal.*

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

TABLE 2 *Summary of PCR-LE testing rates over Sobolev balls. The bold font marks minimax optimal rates. Rates when $d > 4s$ assume that $f_0 \in L^4(\mathscr{X})$, and depend on $\|f_0\|_{L^4(\mathscr{X})}$. Although we suppress it for simplicity, in all cases when other PCR-LE test is optimal, the dependence of the error rate on the radius M of the Sobolev ball is also optimal.*

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

rates are the same for Sobolev spaces—in estimation, when loss is measured in empirical norm—for the values of s and m mentioned above.

In all these cases, our bounds also depend optimally on the radius M of the Sobolev ball under consideration. However, for some values of s (number of derivatives) and d (dimension), there do exist gaps between our upper bounds on the error of PCR-LE and the minimax rates. Although we do not give corresponding lower bounds verifying the tightness of our analysis, we believe these gaps reflect the true behaviour of the method rather than some looseness in our analysis, and we comment more on this at relevant parts in the text. For completeness, we summarize all of our upper bounds—those which match the minimax rates, and those which do not—in Tables 1 and 2.

Perspective: regression error versus feature reconstruction. We now pause for a moment, to emphasize that in a certain respect the aforementioned rates of convergence for PCR-LE are quite surprising. Remember that PCR-LE is a regression method using features (eigenvectors v_k of the graph Laplacian L) which are themselves empirical estimates of population-level quantities (eigenfunctions ψ_k of the density-weighted Laplacian Δ_P). It seems reasonable to expect that the error of PCR-LE should be decomposed into two parts: first, the error with which these empirically derived features estimate their continuum limits; second, the error with which, given ideal population-level features, the regression function is estimated.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Crucially, our analysis *does not* work in this way. This is important because all known upper bounds on the rates at which $v_k \to \psi_k$ as $n \to \infty$ are much slower than the minimax rates for regression over Sobolev classes. For instance, the best currently known upper bound on the empirical L^2 error $\frac{1}{n}\sum_{i=1}^{n}(\sqrt{n}v_{k,i} - \psi_k(X_i))^2$ is only of the order of $n^{-2/(4+d)}$ [19], which is slower than the minimax estimation rate over $H^s(\mathscr{X})$ for any $s \in \mathbb{N}, s \geq 1$.[2] Although this upper bound may not reflect the true rate of convergence of graph Laplacian eigenvectors—this is still an active area of research, and no lower bounds are known—it seems very unlikely that the true rate matches the minimax estimation rate $n^{-2s/(2s+d)}$, which after all approaches the dimension-free rate $1/n$ for large values of s. The bottom line is that the rate at which graph Laplacian eigenvectors are known to converge to density-weighted Laplacian eigenfunctions is too slow to explain the upper bounds we establish for PCR-LE.

Instead of relying on convergence of eigenvectors to eigenfunctions, our analysis proceeds via a bias-variance decomposition at the level of the graph. As usual for OLS estimates, the variance term depends only on the degrees of freedom $\mathrm{df}(\widehat{f}) = \mathrm{tr}(V_K V_K^\top) = K$. More surprisingly, the bias can also be upper bounded without appealing to concentration of eigenvectors $v_1, \ldots, v_K$ around eigenfunctions $\psi_1, \ldots, \psi_K$; for instance, we show in Lemma C.1 that for estimation the squared bias is at most of the order of $f_0^\top L^s f_0/(n\lambda_{K+1}^s)$.

Ultimately, our upper bound on the error of PCR-LE is determined entirely by a pair of graph functionals: the quadratic form $f_0^\top L^s f_0$, and the graph Laplacian eigenvalue λ_{K+1}. This brings a couple of advantages:

- First, it eliminates the need to analyse convergence of eigenvectors to eigenfunctions, which is critical in order to get sufficiently fast rates of convergence for PCR-LE, as we have already explained. Instead, we only have to consider these two graph functionals, both of which are known to converge at faster rates than graph Laplacian eigenvectors.
- Second, in order to obtain upper bounds on $\|\widehat{f} - f_0\|_n^2$ we do not require that these graph functionals themselves converge to population-level limits, but only that they be stochastically bounded on the right order. The latter is a much weaker requirement.

To derive our upper bounds on the error of PCR-LE, we directly analyse the quadratic form $f_0^\top L^s f_0$ and the eigenvalue λ_{K+1}, using some existing results as well as deriving some new ones which may be of independent interest.

To summarize, our work demonstrates, broadly speaking, that regression using estimated features can be analysed independently from the estimation error of the features themselves. Regression using learned features—that is, a feature representation derived from the data itself—is a general and widely applied paradigm, and we believe this observation may have consequences outside of its application to PCR-LE in this work.

1.2 *Related work*

Laplacian smoothing. In a previous paper [31], we (the authors) considered an alternative method for non-parametric regression via neighbourhood graphs: *Laplacian smoothing*, defined as the solution to

[2] To make matters worse, PCR-LE, when deployed optimally, does not use a single eigenvector v_k for a fixed index $k \in \mathbb{N}$, but rather many eigenvectors $v_1, \ldots, v_K$ with K growing in n. As K grows larger, the rate at which $v_K \to \psi_K$ gets slower, since the population-level object being estimated is less regular; see [15, 28].

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

the following optimization problem,

$$\underset{f\in\mathbb{R}^n}{\text{minimize}}\ \|\mathbf{Y}-f\|_2^2+\lambda f^\top Lf. \tag{1.5}$$

Laplacian smoothing is a penalized method for regression, where the penalty functional $f^\top Lf$ serves as a discrete approximation to the continuum functional $J(f) := \int \|\nabla f(x)\|^2 p^2(x)\,dx$ [13]. In the univariate setting ($d = 1$), this casts Laplacian smoothing as a discrete and density-weighted alternative to a first-order thin-plate spline estimator, which is defined as the solution to

$$\underset{f\in H^1(\mathbb{R})}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n}(Y_i-f(X_i))^2+J(f). \tag{1.6}$$

When $d = 1$ the first-order thin-plate spline estimator enjoys excellent theoretical properties, such as being minimax optimal over the first-order Sobolev space $H^1(\mathbb{R})$. However, when $d \geq 2$ the story changes dramatically: the problem (1.6) is in fact not even well-posed.[3] In contrast, in this previous paper, we showed that Laplacian smoothing was a well-posed and consistent estimator for any (fixed) dimension d, and achieved minimax optimal rates for estimation and testing so long as $d \in \{1, 2, 3, 4\}$.

However, Laplacian smoothing neither takes advantage of smooth higher order derivatives, nor is it provably optimal over $H^1(\mathscr{X})$ for dimensions $d \geq 5$. One of our motivations for considering PCR-LE was to find an estimator which addressed these deficiencies. In this work we indeed establish that PCR-LE has much stronger optimality properties than those we derived for Laplacian smoothing, or indeed those known for any other method of regression using neighbourhood graphs.

One way to interpret this difference between PCR-LE and Laplacian smoothing is to view the latter as a ridge regression problem. This follows from writing the Laplacian smoothing penalty as a (weighted) ridge penalty in the spectral domain, $f^\top Lf = \sum_{k=1}^{n}\lambda_k(v_k^\top f)^2$. Dhillon et al. [20] establish conditions under which principal components regression can have smaller risk than ridge regression using the same set of features. Viewed in this light, our work shows this phenomenon occurs when the features are eigenvectors of a neighbourhood graph Laplacian and the estimand is a function in Sobolev space. It also establishes that principal components regression can obtain the minimax rate of convergence even when ridge fails to do so. Interestingly, this is not the case if the function class in question is an RKHS [21], and further motivates the study of regression over Sobolev spaces in the subcritical regime, where surprising new phenomena emerge.

*Other related work.*Much of the work regarding regression using neighbourhood graph Laplacians deals with *semi-supervised learning*, where in addition to the labelled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ one observes unlabelled points $(X_{n+1}, \ldots, X_N)$, and the task is to produce an estimate at labelled and unlabelled points alike. To this end, the landmark paper of [75] proposed to interpolate the observed values by *harmonic extension*, i.e. compute the Laplacian matrix L_N corresponding to a graph formed over all design points $X_1, \ldots, X_N$, and then solve the constrained problem

$$\underset{f\in\mathbb{R}^N}{\text{minimize}}\ f^\top L_N f \quad \text{subject to} \quad f_i = Y_i \ \text{ for } i = 1, \ldots, n.$$

[3] This can be explained by reference to the Sobolev Embedding Theorem, since it is an implication of this theorem that convergence of a sequence of functions $\{f_N\}_{N\in\mathbb{N}} \to f$ in first-order Sobolev norm implies pointwise convergence only when $d = 1$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Conventional wisdom says that harmonic extension is sensible only when the responses are noiseless, $Y_i = f_0(X_i)$, and that in the noisy setting one should instead solve the penalized formulation

$$\underset{f \in \mathbb{R}^N}{\text{minimize}} \sum_{i=1}^{n} (Y_i - f_i)^2 + \lambda f^\top L_N f. \tag{1.7}$$

Notwithstanding their intuitive appeal, both the constrained and penalized problems have issues when $d > 1$ and $n/N \to 0$: the estimates tend towards degeneracy, meaning they are 'spiky' at labelled data points and close to constant everywhere else [17, 18, 48]. One solution to this problem is to instead use LE for semi-supervised learning (SSL-LE), i.e. compute the eigendecomposition $L_N = \sum_{k=1}^{N} \lambda_k u_k u_k^\top$ and, letting $U \in \mathbb{R}^{n \times K}$ be the matrix with entries $U_{ik} = u_{k,i}$ and columns $U_1, \ldots, U_K$, solve the problem.

$$\underset{f \in \text{span}\{U_1, \ldots, U_K\}}{\text{minimize}} \sum_{i=1}^{n} (Y_i - f_i)^2. \tag{1.8}$$

[46, 74] analyse SSL-LE in a particular asymptotic regime where the number of labelled points n is held fixed while the number of unlabelled points $N - n \to \infty$. They show that the SSL-LE estimator achieves minimax optimal rates—as a function of the number of labelled points n—over Sobolev spaces. However, in the particular asymptotic regime when n is fixed and $N - n \to \infty$, the n lowest frequency eigenvectors of the graph Laplacian L_N all converge to their continuum limits. Consequently, the SSL-LE estimator converges to the population-level spectral series estimator, and the analysis of SSL-LE reduces to that of the population-level method. As we have already explained, the supervised setting (where $N = n$) we consider in this work is very different, and analysing PCR-LE necessitates an entirely different approach,

In this supervised setting, there has been relatively little work regarding *random design* regression with neighbourhood graph Laplacians. Aside from our own work on Laplacian smoothing, summarized above, we highlight two other related papers: Lee et al. [46], who analyse a variant of PCR-LE, but derive suboptimal rates of convergence, and García Trillos and Murray [26], who study Laplacian smoothing and establish the uniform upper bound $\max_{i=1,\ldots,n} |\check{f}(X_i) - f_0(X_i)| \leq Cn^{-1/(2+d)}$ under the assumption $f_0 \in C^2(\mathscr{X})$, which is slower than the minimax rate $(\log n/n)^{-2/(4+d)}$ for this function class [65].

Most work on supervised learning using graphs adopts a *fixed design* perspective, treating the design points $X_1 = x_1, \ldots, X_n = x_n$ as vertices of a fixed graph, and carrying out inference with respect to the conditional mean vector $(f_0(x_1), \ldots, f_0(x_n))$. In this setting, matching upper and lower bounds have been established that certify the optimality of graph-based methods for estimation [38, 42, 43, 53, 54, 71]) and testing [55–58] over different 'function' classes (in quotes because these classes really model the n-dimensional vector of evaluations). 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, depending on the data collection process, it may be unnatural to model the design points as being a priori fixed, and the estimand as being a vector which exhibits a discrete notion of 'smoothness' over this fixed design. Instead, we adopt the *random design* perspective, and seek to estimate a function that we assume exhibits a more classical notion of smoothness.

Roadmap. We now outline the structure of the rest of this paper. In Section 2, we give our formal modelling assumptions, and precisely define the PCR-LE estimator and test we study. Propositions 1 and 2, in Section 2.3, show that under rather general (non-parametric) conditions on the design distribution, population-level spectral series methods achieve minimax rates of convergence over Sobolev classes. Then in Sections 3 and 4 we give our main upper bounds on the error of PCR-LE.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

These upper bounds (summarized above) hold under similarly general conditions, and imply that the PCR-LE estimator and test are also minimax rate-optimal. In Section 5 we examine the empirical behaviour of PCR-LE, and show that even at moderate sample sizes PCR-LE is competitive with population-level spectral series regression. We conclude with some discussion in Section 6.

Notation. We now introduce some notation; for ease of reference, we include a table summarizing notation in Appendix A.

We frequently refer to various classical function classes, starting with the Lebesgue space $L^2(\mathscr{X})$, defined differently depending on whether $\mathscr{X} \subseteq \mathbb{R}^d$ is a full-dimensional open set or a compact Riemannian manifold. When $\mathscr{X} \subseteq \mathbb{R}^d$ is a full-dimensional open set, letting $d\nu$ denote the Lebesgue measure, the space $L^2(\mathscr{X})$ refers to the set of ν-measurable functions f for which $\|f\|_{L^2(\mathscr{X})}^2 := \int f^2 \, d\nu < \infty$. When $\mathscr{X}$ is a compact Riemannian manifold, letting $d\mu$ denote the volume form induced by the embedding of $\mathscr{X}$ into $\mathbb{R}^d$, the space $L^2(\mathscr{X})$ refers to the set of μ-measurable functions f for which $\|f\|_{L^2(\mathscr{X})}^2 := \int f^2 \, d\mu < \infty$. We also define an inner-product over these spaces: for a measure P which admits a density p with respect to ν, we define $\langle f, g\rangle_P := \int f(x)g(x)p(x)\, d\nu(x)$; likewise, if P admits a density p with respect to μ, $\langle f, g\rangle_P := \int f(x)g(x)p(x)\, d\mu(x)$. We refer to the norm $\|f\|_P^2 := \langle f, f\rangle_P$ as the squared $L^2(P)$-norm. The inner product $\langle f, g\rangle_P$ and squared norm $\|f\|_P^2$ have empirical counterparts $\langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$ and $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n [f(X_i)]^2$.

We use $C^k(\mathscr{X})$ to refer to functions which are k times continuously differentiable in $\mathscr{X}$, either for some integer $k \geq 1$ or for $k = \infty$. We let $C_c^\infty(\mathscr{X})$ represent those functions in $C^\infty(\mathscr{X})$ with support V compactly contained in $\mathscr{X}$, meaning $\overline{V}$ is compact and $\overline{V} \subseteq \mathscr{X}$. 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 $D^\alpha f := \partial^{|\alpha|} f / \partial^{\alpha_1} x_1 \ldots \partial^{\alpha_d} x_d$ for multi-indices $\alpha \in \mathbb{N}^d$. Recall that for a given multi-index $\alpha \in \mathbb{N}^d$, a function f is *α-weakly differentiable* if there exists some $h \in L^1(\mathscr{X})$ such that

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

If such a function h exists, it is the αth weak partial derivative of f, and denoted by $D^\alpha f := h$. For functions f which are $|\alpha|$-times classically differentiable, this coincides with the classical definition of derivative, and so we use the same notation for both.

For a vector $v \in \mathbb{R}^n$, we write $\|v\| = \|v\|_2$ for Euclidean norm and $|v| = \|v\|_1$ for ℓ_1 norm; we will sometimes abuse notation by taking $\|v\|_n^2 = \frac{1}{n}\|v\|^2$. We let $d_{\mathscr{X}}(x', x)$ be the geodesic distance between points x and x' on a manifold $\mathscr{X}$. Then for a given $\delta > 0$, $B(x, \delta)$ is the radius-δ ball with respect to Euclidean distance, whereas $B_{\mathscr{X}}(x, \delta)$ is the radius-δ ball with respect to geodesic distance. Letting $T_x(\mathscr{X})$ be the tangent space at a point $x \in \mathscr{X}$, we write $B_m(v, \delta) \subset T_x(\mathscr{X})$ for the radius-δ ball centred at $v \in T_x(\mathscr{X})$.

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 $a \wedge b := \min\{a, b\}$.

2. Preliminaries

We begin in Sections 2.1–2.2 by precisely defining the models (random design points, Sobolev regression functions) and methods (PCR-LE) under consideration. Then in Section 2.3, we analyse the behaviour of population-level spectral series regression.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

2.1 *Non-parametric regression over Sobolev spaces*

As mentioned, we will always operate in the usual setting of non-parametric regression with random design. We observe independent random samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the design points $X_1, \ldots, X_n$ are sampled from a distribution P with support $\mathcal{X} \subseteq \mathbb{R}^d$, and the responses follow (1.3). We now formulate two sets of assumptions on the design distribution P, in which the support $\mathcal{X}$ is either a *flat Euclidean* or *manifold* domain. We also review the definition of L^2-Sobolev spaces in both cases.

Flat Euclidean setting. We will use the phrase *flat Euclidean* to refer to a design P satisfying the following pair of conditions:

- 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 ∞,

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

Note that at various points we will also assume that the density $p \in C^k(\mathcal{X})$ for some integer $k \geq 1$. For a flat Euclidean domain and an integer $s \geq 1$, the order-s Sobolev space is defined as follows.

Definition 1. (Sobolev space on a flat Euclidean domain). 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 $\alpha \in \mathbb{N}^d, |\alpha| \leq s$, the weak derivatives $D^\alpha f$ exist and satisfy $D^\alpha f \in L^2(\mathcal{X})$. The jth order semi-norm for $f \in H^s(\mathcal{X})$ is $|f|_{H^j(\mathcal{X})} := \sum_{|\alpha|=j} \|D^\alpha f\|_{L^2(\mathcal{X})}$, and the corresponding squared norm

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

induces the Sobolev ball

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

When $s > 1$ our results will also require that f_0 satisfy a zero-trace boundary condition. Recall that $H^s(\mathcal{X})$ can alternatively be defined as the completion of $C^\infty(\mathcal{X})$ in the Sobolev norm $\| \cdot \|_{H^s(\mathcal{X})}$. The zero-trace Sobolev spaces are defined in a similar fashion, as the completion of $C_c^\infty(\mathcal{X})$ in the same norm.

Definition 2. (Zero-trace Sobolev space). A function $f \in H^s(\mathcal{X})$ belongs to the zero-trace Sobolev space $H_0^s(\mathcal{X})$ if 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 $H_0^s(\mathcal{X}; M) := H_0^s(\mathcal{X}) \cap H^s(\mathcal{X}; M)$.

Boundary conditions play an important role in the analysis of spectral methods, as we explain further in Section 2.3. For now, we limit ourselves to pointing out that for functions $f \in C^\infty(\mathcal{X})$, the zero-trace

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

condition can be stated more concretely, as implying that $\partial^k f/\partial \mathbf{n}^k(x) = 0$ for each $k = 0, \dots, s-1$, and for all $x \in \partial \mathscr{X}$. (Here $\partial/(\partial \mathbf{n})$ is the partial derivative operator in the direction of the normal vector $\mathbf{n}$.)

Manifold setting. As in the flat Euclidean case, we start with some regularity conditions on the design. For the second condition, recall that the reach R is the largest radius of a ball which can be rolled around the manifold $\mathscr{X}$; mathematically,

$$R := \left\{ \sup_{r>0} : \forall z \in \mathbb{R}^d,\ \inf_{x \in \mathscr{X}} \|z - x\| \le r,\ \exists!\ y \in \mathscr{X} \text{ s.t. } \|z - y\| = \inf_{x \in \mathscr{X}} \|z - x\| \right\}.$$

We will use the phrase *manifold design* to refer to a design P satisfying the following conditions:

- The support $\mathscr{X}$ of the design distribution P is a closed, connected and smooth Riemannian manifold (without boundary) embedded in $\mathbb{R}^d$, of intrinsic dimension $1 \le m < d$.
- The manifold $\mathscr{X}$ has positive reach $R > 0$.
- The design distribution P admits a Lipschitz density p with respect to the volume form $d\mu$ induced by the Riemannian structure of $\mathscr{X}$, which is bounded away from 0 and ∞,

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

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

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

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

Definition 3. (Sobolev space on a manifold). A function $f \in L^2(\mathscr{X})$ belongs to the Sobolev space $H^s(\mathscr{X})$ if for all $\alpha \in \mathbb{N}^d, |\alpha| \le s$, the weak derivatives $D^\alpha f$ exist and satisfy $D^\alpha f \in L^2(\mathscr{X})$. The jth order semi-norm $|f|_{H^j(\mathscr{X})}$, the norm $\|f\|_{H^s(\mathscr{X})}$, and the ball $H^s(\mathscr{X}; M)$ are all defined as in Definition 1.

As we defined them, the norm $\|f\|_{H^s(\mathscr{X})}$ and ball $H^s(\mathscr{X}; M)$ are highly non-intrinsic, in that they depend on the choice of local coordinates. However, any two coordinate systems will result in equivalent

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

norms,[4] and so the definition of $H^s(\mathscr{X})$ is independent of local coordinates. Additionally, all of our upper bounds on the estimation and testing error of PCR-LE hold up to constant factors, regardless of what choice of local coordinates is made. For an alternative, more intrinsic definition of a Sobolev space on a manifold, see [34].

2.2 *Principal components regression with Laplacian Eigenmaps*

We now formally define the estimator and test statistic we study. Both are derived from eigenvectors of a graph Laplacian. For a function $\eta : [0,\infty) \to [0,\infty)$, and a radius parameter $\varepsilon > 0$, let $G = (\{X_1,\ldots,X_n\}, W)$ be the neighbourhood graph formed over the design points $\{X_1,\ldots,X_n\}$, with a weighted edge $W_{ij} = \eta(\|X_i - X_j\|/\varepsilon)$ between vertices i and j. We refer to η as a similarity kernel, or just kernel for short.

Then the *neighbourhood graph Laplacian* $L_{n,\varepsilon} : \mathbb{R}^n \to \mathbb{R}$ is defined by its action on vectors $u \in \mathbb{R}^n$ as

$$\left(L_{n,\varepsilon}u\right)_i := \frac{1}{n\varepsilon^{2+\dim(\mathscr{X})}} \sum_{j=1}^{n} \left(u_i - u_j\right) \eta\left(\frac{\|X_i - X_j\|}{\varepsilon}\right). \tag{2.1}$$

(Here $\dim(\mathscr{X})$ stands for the dimension of $\mathscr{X}$. It is equal to d in the flat Euclidean setting and equal to m in the manifold setting. The pre-factor $(n\varepsilon^{2+\dim(\mathscr{X})})^{-1}$ ensures non-degenerate stable limits as $n \to \infty, \varepsilon \to 0$). Note that $(n\varepsilon^{\dim(\mathscr{X})+2}) \cdot L_{n,\varepsilon} = D - W$, 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 $k \in \{1,\ldots,n\}$ the eigenvalue-eigenvector pair (λ_k, v_k) satisfies

$$L_{n,\varepsilon} v_k = \lambda_k v_k, \quad \|v_k\|_2^2 = 1.$$

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 $(\lambda_1, v_1), \ldots, (\lambda_n, v_n)$ in ascending order of eigenvalue, $0 = \lambda_1 < \lambda_2 < \ldots < \lambda_n$.

The PCR-LE estimator $\widehat{f}$ defined in (1.1) simply projects the response vector $\mathbf{Y}$ onto the first K eigenvectors of $L_{n,\varepsilon}$. Since the eigenvectors of the graph Laplacian are orthonormal with respect to the Euclidean inner product on $\mathbb{R}^n$, we can more simply write this as

$$\widehat{f} = V_K V_K^\top \mathbf{Y}, \tag{2.2}$$

where $V_K \in \mathbb{R}^{n\times K}$ is the matrix with kth column $V_{K,k} = v_k$.

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

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

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

The PCR-LE test statistic is

$$\widehat{T} := \|\widehat{f}\|_n^2 = \frac{1}{n}\mathbf{Y}^\top V_K V_K^\top \mathbf{Y}, \tag{2.3}$$

and can be used to determine whether $f_0 = 0$.

2.3 *Population-level spectral series regression*

We now establish some upper bounds on the error of population-level spectral series regression when $f_0 \in H^s(\mathscr{X})$, which imply that such methods achieve optimal rates of convergence for both estimation and testing. The upper bounds we establish are 'usual' in the sense that they match the rates $n^{-2s/(2s+d)}$ (estimation) and $n^{-4s/(4s+d)}$ (testing) which are already known in many cases. However, they are unusual in that we treat both the case where $s < d/2$ and thus the Sobolev space $H^s(\mathscr{X})$ does not continuously embed into $C(\mathscr{X})$, and the case where P is not the uniform distribution over the unit cube. The upper bounds given in this section serve two purposes: first, to clarify what the rates are in these less-typically studied settings; second, to show that even in this general setting, population-level spectral series regression can always obtain the optimal rates. The latter point is important since the method we focus on for the most part, PCR-LE, is an empirical approximation to population-level spectral series regression.

Spectrally defined Sobolev spaces. Suppose we have a flat Euclidean design P supported on $\mathscr{X}$. Recalling the density-weighted Laplacian Δ_P, defined in (1.2), we consider the eigenvector equation with Neumann boundary conditions,

$$\Delta_P \psi = \rho\psi, \quad \frac{\partial}{\partial \mathbf{n}}\psi = 0 \text{ on } \partial\mathscr{X}. \tag{2.4}$$

The eigenvector equation (2.4) has enumerable solutions $(\rho_1, \psi_1), (\rho_2, \psi_2), \ldots$, sorted as usual in ascending order of eigenvalue [27]. These eigenvalues and eigenfunctions can be used to give a spectral definition of Sobolev spaces: these are the spaces

$$\mathscr{H}^s(\mathscr{X}) := \left\{\sum_{k=1}^{\infty} a_k\psi_k \in L^2(\mathscr{X}) : \sum_{k=1}^{\infty} a_k^2\rho_k^s < \infty\right\}, \tag{2.5}$$

with norm $\|f\|_{\mathscr{H}^s(\mathscr{X})}^2 = a_k^2\rho_k^s$ for $f = \sum a_k\psi_k$, and corresponding ball

$$\mathscr{H}^s(\mathscr{X}; M) := \left\{f \in \mathscr{H}^s(\mathscr{X}) : \|f\|_{\mathscr{H}^s(\mathscr{X})} \le M\right\}. \tag{2.6}$$

Under appropriate regularity conditions $\mathscr{H}^s(\mathscr{X})$ consists of functions $f \in H^s(\mathscr{X})$ which also satisfy some additional boundary conditions. For instance if $p \in C^\infty(\mathscr{X})$ and $\partial\mathscr{X} \in C^{1,1}$ then Dunlop et al. [22] show that for any $s \ge 1$,

$$\mathscr{H}^{2s}(\mathscr{X}) = \left\{f \in H^{2s}(\mathscr{X}) : \frac{\partial \Delta_P^r f}{\partial \mathbf{n}} = 0 \text{ on } \partial\mathscr{X}, \text{ for all } 0 \le r \le s-1\right\}, \tag{2.7}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

and likewise $\mathscr{H}^{2s+1}(\mathscr{X}) = \mathscr{H}^{2s}(\mathscr{X}) \cap H^{2s+1}(\mathscr{X})$ for any $s \geq 0$. Additionally, the norms $\|f\|_{\mathscr{H}^s(\mathscr{X})}$ and $\|f\|_{H^s(\mathscr{X})}$ are equivalent.

Estimation with spectral series regression. Recall the population-level spectral series estimator $\widetilde{f}$ defined in (1.4). We now give an upper bound on the mean-squared error of $\widetilde{f}$.

Proposition 1. In the flat Euclidean setting, assume additionally that $\partial\mathscr{X} \in C^{1,1}$, $p \in C^\infty(\mathscr{X})$, $f_0 \in \mathscr{H}^s(\mathscr{X};M)$ and $\|f_0\|_P^2 \leq 1$. Then there exists a constant C which does not depend on f_0, M or n such that the following statement holds: if the population-level spectral series estimator in (1.4) is computed with parameter $K = \max\{\lfloor M^2 n\rfloor^{d/(2s+d)}, 1\}$, then

$$\mathbb{E}\left[\|\widetilde{f} - f_0\|_P^2\right] \leq C \max\left\{M^2\left(M^2 n\right)^{-2s/(2s+d)}, \frac{1}{n}\right\}. \tag{2.8}$$

When the Sobolev ball radius $n^{-1/2} \lesssim M$, the upper bound in (2.8) is of the order of $M^2(M^2n)^{-2s/(2s+d)}$. This is well known to be the minimax rate of estimation over the Sobolev classes $H^s([0,1]^d;M)$ when $s > d/2$; see e.g. [33, 66, 72] and references therein, and specifically Theorem 3.2 of [33] for a matching lower bound in the context of non-parametric regression with random design. On the other hand there seems to have been much less study of minimax rates over $H^s([0,1]^d;M)$ when $s < d/2$. In this *subcritical* regime, the Sobolev space contains functions without continuous representatives, and certain questions become more subtle; see our remark after Theorem 3. However, Proposition 1 confirms that in this regime the minimax rates (with loss measured in squared-$L^2(P)$ norm) are still of the order of $M^2(M^2n)^{-2s/(2s+d)}$, since a matching lower bound follows from the known estimation rates over $C^s([0,1]^d) \subseteq H^s([0,1]^d)$ [64].

Testing with spectral series regression. In the goodness-of-fit testing problem, one asks for a test function—formally, a Borel measurable function ϕ that takes values in $\{0,1\}$—which can distinguish between the hypotheses

$$\mathbf{H}_0 : f_0 = f_0^\star, \quad \text{versus} \quad \mathbf{H}_a : f_0 \in \mathscr{H}^s(\mathscr{X};M) \setminus \{f_0^\star\}. \tag{2.9}$$

To fix ideas, here and throughout we focus on the signal detection problem, meaning the special case where $f_0^\star = 0$.[5] We are interested in how large $\|f_0\|_P^2$ needs to be in order for a level-a test to have power of at least b, for some $a, b \in (0,1)$. For more background on non-parametric goodness-of-fit testing problems, see [40].

A natural test statistic for the signal detection problem is $\widetilde{T} = \|\widetilde{f}\|_P^2$. The population-level spectral series test $\widetilde{\varphi} := \mathbf{1}\{\widetilde{T} \geq K/n + \sqrt{2K/an^2}\}$ has bounded Type I error, $\mathbb{E}_0[\widetilde{\varphi}] \leq a(1 + o(1))$ so long as $(M^2n)^{2d/(4s+d)} \leq n$. Proposition 2 gives an upper bound on the Type II error that holds over all $f_0 \in \mathscr{H}^s(\mathscr{X};M)$ for which $\|f_0\|_P^2$ is sufficiently large.

Proposition 2. Fix $a, b \in (0,1)$. In the flat Euclidean setting, suppose additionally that the density p is known, that $\partial\mathscr{X} \in C^{1,1}$, $p \in C^\infty(\mathscr{X})$, $f_0 \in \mathscr{H}^s(\mathscr{X};M)$ for some $s > d/4$, and $\|f_0\|_{L^4(\mathscr{X})}^4 \leq 1$. Then there exists a constant C which does not depend on f_0, M or n such that the following statement holds: if

[5] This is without loss of generality since all the test statistics we consider 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, with no change in the analysis.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

the population-level spectral series test $\widetilde{\varphi}$ is computed with parameter $K = \max\{\lfloor M^2 n \rfloor^{2d/(4s+d)}, 1\}$, if $(M^2 n)^{2d/(4s+d)} \leq n$, and if

$$\|f_0\|_P^2 \geq C \min\left\{M^2 (M^2 n)^{-4s/(4s+d)}, \frac{1}{n}\right\} \tag{2.10}$$

then the Type II error is upper bounded, $\mathbb{E}_{f_0}[1 - \widetilde{\varphi}] \leq b$.

Assuming again that $n^{-1/2} \lesssim M$, the right-hand side of (2.10] is $M^2(M^2 n)^{-4s/(4s+d)}$, matching the usual minimax critical radius over Sobolev space. (See [32, 39, 40]; specifically, Ingster and Sapatinas [39] show that the minimax squared critical radius is of the order of $n^{-4s/(4s+d)}$ when $M = 1$, and simple alterations of their analysis imply the rate $M^2(M^2 n)^{-4s/(4s+d)}$ for general M.) On the other hand, when $s \leq d/4$ the minimax regression testing rates over $H^s(\mathscr{X})$ are not known. If one explicitly assumes $f_0 \in L^4(\mathscr{X}; 1)$—note that $H^s(\mathscr{X})$ does not continuously embed into $L^4(\mathscr{X})$ when $s \leq d/4$—then the minimax critical radius for regression testing is of the order of $n^{-1/2}$ [32], and is achieved by a test using the naive statistic $\|\mathbf{Y}\|_n^2$. In other words, the regression testing problem over Sobolev spaces fundamentally changes when $s \leq d/4$, and hereafter when we discuss testing we will limit our consideration to $s > d/4$.

The main takeaway from Propositions 1 and 2 is that population-level spectral series methods for regression achieve optimal rates of convergence, when the regression function f_0 is Sobolev smooth and the design distribution P is known a priori and satisfies an appropriate notion of smoothness.[6] We reiterate that when the design distribution is unknown, these methods have to be treated as oracle methods, in contrast to PCR-LE. As we will see, PCR-LE achieves comparable rates of convergence when p is sufficiently smooth but potentially unknown.

Of course, it is worth pointing out that other methods besides PCR-LE are statistically optimal for non-parametric regression even when p is unknown. We comment more on some of these in Section 6, after we have derived our major results regarding PCR-LE.

3. Minimax optimality of PCR-LE

In this section we give upper bounds on the error of PCR-LE in the flat Euclidean setting. We will divide our theorem statements based on whether the regression function f_0 belongs to the first order Sobolev class $H^1(\mathscr{X})$ or a higher order Sobolev class ($H_0^s(\mathscr{X})$ for some integer $s \geq 2$), since the details of the two settings are somewhat different.

3.1 *First-order Sobolev classes*

We begin assuming $f_0 \in H^1(\mathscr{X}; M)$. We show that $\widehat{f}$ and a test based on $\widehat{T}$ are minimax optimal for all values of d for which the minimax rates are known.

Estimation with PCR-LE. PCR-LE depends on the kernel η and two tuning parameters, the graph radius ε and number of eigenvectors K. We will need to make some assumptions on each.

[6] The assumption $p \in C^\infty(\mathscr{X})$ could likely be weakened, but since this would not substantially add to the main points of Propositions 1 and 2, we do not pursue the details further.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

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

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

(P1) The number of eigenvectors is given by

$$K = \min\left\{\left\lfloor (M^2 n)^{d/(2+d)} \right\rfloor \vee 1, n\right\}. \tag{3.1}$$

If $K < n$, then additionally

$$C_0\left(\frac{\log n}{n}\right)^{1/d} \le \varepsilon \le c_0 \min\{1, K^{-1/d}\}. \tag{3.2}$$

In (3.2) the numbers C_0 and c_0 are constants that do not depend on n,f_0 or M, but may depend on P and d. For instance, C_0 is sufficiently large to ensure that the neighbourhood graph is connected with high probability.

We now have our first main theorem, regarding the estimation error of PCR-LE. The proof of this theorem, along with the proofs of all subsequent results, can be found in the Appendix.

Theorem 1. In the flat Euclidean setting, suppose additionally $f_0 \in H^1(\mathscr{X}, M)$. There are constants c, C and N (not depending on f_0, M or n), such that the following statement holds for all $n \geq N$ and any $\delta \in (0, 1)$: if the PCR-LE estimator $\widehat{f}$ is computed with a kernel η satisfying **(K1)**, and parameters ε and K satisfying **(P1)**, then

$$\|\widehat{f} - f_0\|_n^2 \le C\left(\frac{1}{\delta} M^2 (M^2 n)^{-2/(2+d)} \wedge 1\right) \vee \frac{1}{n}, \tag{3.3}$$

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

From (3.3) it follows immediately that when $n^{-1/2} \lesssim M \lesssim n^{1/d}$, then with constant probability $\|\widehat{f} - f_0\|_n^2 \lesssim M^2(M^2 n)^{-2/(2+d)}$, matching the minimax estimation rate over Sobolev classes.

Some other remarks:

- *Radius of the Sobolev ball.* When $M = o(n^{-1/2})$ then computing PCR-LE 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 PCR-LE with $K = n$ achieves the rate $\|\widehat{f} - f_0\|_n^2 \lesssim 1$, which is better than the rate in (2.8). This is because we are evaluating error in-sample rather than out-of-sample. However, in truth these are edge cases, which do not fall neatly into the framework of non-parametric regression.

- *Meaning of pointwise evaluation.* There is one subtlety introduced by the use of in-sample mean squared error. Since elements $f \in H^s(\mathscr{X})$ are equivalence classes, defined only up to a set of measure zero, one cannot really speak of the pointwise evaluation $f_0(X_i)$, as we do by defining our target of estimation to be $(f_0(X_1), \ldots, f_0(X_n))$, until one selects a representative of each equivalence class f.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Implicitly, we will always pick the *precise representative* $f_0^* \in f_0$ (as defined in [24]), and the notation '$f_0(X_i)$' should always be interpreted as $f_0^*(X_i)$. 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 so any two $g_0, h_0 \in f_0$ satisfy $g_0(X_i) = h_0(X_i)$ for all $i = 1, \dots, n$ almost surely. For this reason we can write $f_0(X_i)$ without fear of ambiguity or confusion.

- *Tuning parameters*. 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 (3.2) is of the order of the connectivity threshold, the smallest radius for which the resulting graph will still be connected with high probability. On the other hand, as we will see in Section 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 ρ_K; this is essential in order to obtain a tight upper bound on the bias of $\widehat{f}$. Finally, we set $K = \lfloor (M^2 n)^{d/(2+d)} \rfloor$ (when possible) to optimally trade-off bias and variance, as is typical.
- *High-probability guarantees*. The upper bound given in (3.3) holds with probability $1 - \delta - o(1)$. Under the stronger assumption that f_0 is M-Lipschitz, we can establish the same guarantee (3.3) with probability $1 - \delta^2/n - Cn\exp(-cn\varepsilon^d) - \exp(-K)$; in other words, we can give a high-probability guarantee (for details see [31]). In this case a routine calculation shows that $\mathbb{E}[\|\widehat{f} - f_0\|_n^2]$ will also be on the some order as (3.3). We also suspect that high-probability guarantees will hold so long as $\|\nabla f\|_{L^q(\mathscr{X})}$ is bounded for some sufficiently large $q < \infty$, but it remains an open question whether such guarantees can be obtained in the Sobolev case ($q = 2$) which is the focus of this work.

Testing with PCR-LE. 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. As we show in Theorem 2, when $d < 4$, ε and K are chosen appropriately, and the alternative f_0 is sufficiently well separated from 0, the test φ has Type II error of at most b.

(P2) The number of eigenvectors is given by

$$K = \min\left\{\left\lfloor (M^2 n)^{2d/(4+d)} \right\rfloor \vee 1, n\right\}. \tag{3.4}$$

If $K < n$ then additionally the graph radius ε satisfies (3.2).

THEOREM 2. Fix $a, b \in (0, 1)$. The PCR-LE test φ, computed with threshold t_a, is a level-a test: $\mathbb{E}_0[\varphi] \leq a$. Additionally, in the flat Euclidean setting, suppose $f_0 \in H^1(\mathscr{X}; M)$, 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 $n \geq N$: if φ is computed with a kernel η satisfying **(K1)**, and parameters ε and K satisfying **(P2)**, and if f_0 satisfies

$$\|f_0\|_P^2 \geq C\left(\left(M^2(M^2 n)^{-4/(4+d)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{bn^{2/d}}\right) \vee \frac{1}{n}, \tag{3.5}$$

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

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Although (3.5) involves taking the maximum of several different terms, the important takeaway of Theorem 2 is that if $n^{-1/2} \lesssim M \lesssim n^{(4-d)/4d}$, then φ has small worst-case risk as long as f_0 is separated from 0 by at least $M^2(M^2n)^{-4/(4+d)}$. This implies that φ is a minimax rate-optimal test over $H^1(\mathscr{X};M)$ when $d \in \{1,2,3\}$. As mentioned previously, when $d \geq 4$ the first-order Sobolev space $H^1(\mathscr{X})$ does not continuously embed into $L^4(\mathscr{X})$, and in this case the optimal rates for regression testing over Sobolev spaces are unknown.

3.2 *Higher order Sobolev classes*

We now consider the situation where the regression function displays some higher order regularity, $f_0 \in H_0^s(\mathscr{X};M)$ for some $s \in \mathbb{N}, s \geq 2$. We show that the PCR-LE estimator and test continue to be optimal for all orders of s, as long as the design density is itself also sufficiently regular, $p \in C^{s-1}(\mathscr{X})$. In estimation, this is the case for any dimension d, whereas in testing it is the case only when $d \leq 4$.

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

(P3) The number of eigenvectors is given by

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

If $K < n$ then additionally

$$C_0 \max\left\{\left(\frac{\log}{n}\right)^{1/d}, (M^2n)^{-1/(2(s-1)+d)}\right\} \leq \varepsilon \leq c_0 \min\{1, K^{-1/d}\}. \tag{3.6}$$

Crucially, when n is sufficiently large the two conditions in **(P3)** are not mutually exclusive.

THEOREM 3. In the flat Euclidean setting, suppose additionally $f_0 \in H_0^s(\mathscr{X},M)$ and $p \in C^{s-1}(\mathscr{X})$. 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 PCR-LE estimator $\widehat{f}$ is computed with a kernel η satisfying **(K1)**, and parameters ε and K satisfying **(P3)**, then

$$\|\widehat{f} - f_0\|_n^2 \leq C\left(\frac{1}{\delta}M^2(M^2n)^{-2s/(2s+d)} \wedge 1\right) \vee \frac{1}{n}, \tag{3.7}$$

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

Theorem 3, in combination with Theorem 1, implies that in the flat Euclidean setting PCR-LE is a minimax rate-optimal estimator over Sobolev classes, for all values of s and d. Some other remarks:

- *Sub-critical Sobolev spaces*. Theorems 1 and 3 do not require that the smoothness index s 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 $H^s(\mathscr{X})$ is quite irregular. It is not a Reproducing Kernel Hilbert Space (RKHS), nor does it continuously embed into $C^0(\mathscr{X})$, much less into any Hölder space. As a result, for certain versions of the non-parametric 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

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

RKHS-based methods such as thin-plate splines of degree $k \leq d/2$. Nevertheless, for random design regression with error measured in squared $L^2(P)$-norm, the population-level spectral series estimator $\widetilde{f}$ obtains the standard minimax rates $n^{-2s/(2s+d)}$ for all values of s and d. Theorems 1 and 3 show that the same is true with respect to PCR-LE, when error is measured in empirical norm.

- *Smoothness of design density*. As promised, Theorem 3 shows that PCR-LE achieves optimal rates of convergence so long as the unknown design density p is sufficiently smooth, $p \in C^{s-1}(\mathscr{X})$. The requirement $p \in C^{s-1}(\mathscr{X})$ is essential to showing that $\widehat{f}$ enjoys the faster minimax rates of convergence when $s > 1$, as we discuss in Section 3.3.

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

(P4) The number of eigenvectors is given by

$$K = \min\left\{\left\lfloor (M^2 n)^{2d/(4s+d)} \right\rfloor \vee 1, n\right\}. \tag{3.8}$$

If $K < n$ then additionally the graph radius ε satisfies (3.6).

When $d \leq 4$ and n is sufficiently large, it is possible to choose ε and K such that both (3.6) and (3.8) are satisfied, and our next theorem establishes that in this situation φ is an optimal test.

THEOREM 4. Fix $a, b \in (0,1)$. The PCR-LE test φ, computed with threshold t_a, is a level-a test: $\mathbb{E}_0[\varphi] \leq a$. In the flat Euclidean setting, suppose additionally that $f_0 \in H_0^s(\mathscr{X}, M)$, that $p \in C^{s-1}(\mathscr{X})$ 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 PCR-LE test φ is computed with a kernel η satisfying **(K1)**, and parameters ε and K satisfying **(P4)**, and if f_0 satisfies

$$\|f_0\|_P^2 \geq \frac{C}{b}\left(\left(M^2(M^2n)^{-4s/(4s+d)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{bn^{2s/d}}\right) \vee \frac{1}{n}, \tag{3.9}$$

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

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

THEOREM 5. Under the same set-up as Theorem 4, but with $4 < d < 4s$. If the PCR-LE test φ is computed with a kernel η satisfying **(K1)**, number of eigenvectors K satisfying (3.8), and $\varepsilon = (M^2n)^{-1/(2(s-1)+d)}$, and if

$$\|f_0\|_P^2 \geq \frac{C}{b}\left(\left(M^2(M^2n)^{-2s/(2(s-1)+d)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{bn^{2s/d}}\right) \vee \frac{1}{n}, \tag{3.10}$$

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

Focusing on the special case where $M \asymp 1$, Theorem 5 says that φ has small Type II error whenever $\|f_0\|_P^2 \gtrsim n^{-2s/(2(s-1)+d)}$ and $4 < d < 4s$. This is smaller than the estimation rate $n^{-2s/(2s+d)}$, but larger than the minimax squared critical radius $n^{-4s/(4s+d)}$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

At a high level, it is intuitively reasonable that PCR-LE should have more difficulty achieving the minimax rates of convergence for testing, as opposed to estimation. To obtain the faster rates of convergence for testing, PCR-LE must use many more eigenvectors than are necessary for estimation, including some eigenvectors which correspond to very large eigenvalues. It is known that the approximation properties of eigenvectors corresponding to large eigenvalues are very poor [15, 28], and when $d > 4$ this prevents us from establishing that PCR-LE is an optimal test. That being said, although we suspect φ is truly suboptimal when $d > 4$, our analysis relies on an upper bound on testing bias. Since we do not prove a matching lower bound, we cannot rule out that the test φ is optimal for all $s < d/4$. We leave the matter to future work.

3.3 *Analysis of PCR-LE*

We now outline the high-level strategy we follow when proving each of Theorems 1-5. We analyse the estimation error of $\widehat{f}$, and the testing error of φ, by first conditioning on the design points $\{X_1, \ldots, X_n\}$ and deriving *design-dependent* bias and variance terms. For estimation, we show that with probability at least $1 - \exp(-K)$,

$$\|\widehat{f} - f_0\|_n^2 \le \underbrace{\frac{\langle L_{n,\varepsilon}^s f_0, f_0\rangle_n}{\lambda_{K+1}^s}}_{\text{bias}} + \underbrace{\frac{5K}{n}}_{\text{variance}} . \tag{3.11}$$

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

$$\|f_0\|_n^2 \ge \underbrace{\frac{\langle L_{n,\varepsilon}^s f_0, f_0\rangle_n}{\lambda_{K+1}^s}}_{\text{bias}} + \underbrace{32\frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]}_{\text{variance}}. \tag{3.12}$$

These design-dependent bias-variance decompositions are reminiscent of the more classical bias-variance decompositions typical in the analysis of population-level spectral series methods (for instance (B6) and (B7)), but different in certain key respects. Comparing (3.11) and (3.12) with (B6) and (B7), we see that two continuum objects in the latter pair of bounds, the Sobolev norm $\|f_0\|_{\mathcal{H}^s(\mathcal{X})}^2$ and the eigenvalue ρ_{K+1}, have been replaced by graph-based analogues: the graph Sobolev seminorm $\langle L_{n,\varepsilon}^s f_0, f_0\rangle_n$ and the graph Laplacian eigenvalue λ_{K+1}. These latter quantities, along with the empirical squared norm $\|f_0\|_n^2$, are random variables that depend on the random design points $\{X_1, \ldots, X_n\}$. We proceed to establish suitable upper and lower bounds on these quantities that hold in probability.

Graph Sobolev seminorms. In Proposition 3 we restate an upper bound on the first-order graph Sobolev semi-norm $\langle L_{n,\varepsilon} f, f\rangle_n$ from [31].

PROPOSITION 3. (Lemma 1 of Green et al. [31]). In the flat Euclidean setting, suppose 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 **(K1)** and $\varepsilon < c$, then

$$\langle L_{n,\varepsilon} f, f\rangle_n \le \frac{C}{\delta}\|f\|_{H^1(\mathcal{X})}^2, \tag{3.13}$$

with probability at least $1 - \delta$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Proposition 3 follows by upper bounding the expectation $\mathbb{E}\langle L_{n,\varepsilon}f,f\rangle_n = \langle L_{P,\varepsilon}f,f\rangle_P$—where $L_{P,\varepsilon}$ is the non-local Laplacian operator defined in (3.15)—by (a constant times) the squared Sobolev norm $\|f\|^2_{H^1(\mathcal{X})}$, and then applying Markov's inequality.

In this work, we establish that under certain conditions analogous bounds hold for the higher order graph Sobolev seminorm $\langle L^s_{n,\varepsilon}f,f\rangle_n$, when $s \in \mathbb{N}, s \geq 2$.

PROPOSITION 4. In the flat Euclidean setting, suppose additionally that $f \in H^s_0(\mathcal{X})$ and $p \in C^{s-1}(\mathcal{X})$. Then there exist constants c and C that do not depend on f 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

$$\langle L^s_{n,\varepsilon}f,f\rangle_n \leq \frac{C}{\delta}\|f\|^2_{H^s(\mathcal{X})}, \tag{3.14}$$

with probability at least $1-\delta$.

We now summarize the techniques used to prove Proposition 4, emphasizing the reasons for the conditions imposed on f, p and ε. The following discussion is non-rigorous—for the complete and formal proof see Section D.

To upper bound $\langle L^s_{n,\varepsilon}f,f\rangle_n$ in terms of $\|f\|^2_{H^s(\mathcal{X})}$, we introduce an intermediate quantity: the *non-local Sobolev seminorm* $\langle L^s_{P,\varepsilon}f,f\rangle_P$. This seminorm is defined with respect to the iterated non-local Laplacian $L^s_{P,\varepsilon} = L_{P,\varepsilon} \circ \cdots \circ L_{P,\varepsilon}$, where $L_{P,\varepsilon}$ is a non-local approximation to Δ_P,

$$L_{P,\varepsilon}f(x) := \frac{1}{\varepsilon^{d+2}}\int_{\mathcal{X}} (f(z) - f(x))\,\eta\left(\frac{\|z-x\|}{\varepsilon}\right)dP(z). \tag{3.15}$$

Proposition 4 is proved by showing the pair of inequalities,

$$\langle L^s_{n,\varepsilon}f,f\rangle_n \leq \frac{1}{\delta}\langle L^s_{P,\varepsilon}f,f\rangle_P \leq \frac{C}{\delta}\|f\|^2_{H^s(\mathcal{X})}, \tag{3.16}$$

where the first inequality above is probabilistic and holds with probability $1-\delta$, and in the second inequality C may depend on parameters such as $s, d, \mathcal{X}$ but does not depend on n or the specific $f \in H^s_0(\mathcal{X})$.

The first inequality in (3.16) is shown by bounding the expectation of the graph Sobolev semi-norm in terms of the non-local Sobolev seminorm and applying Markov's inequality. A complication is that unlike in the first-order case, when $s \geq 2$ then $\langle L^s_{n,\varepsilon}f,f\rangle_n$ is itself a biased estimate of the non-local seminorm $\langle L^s_{P,\varepsilon}f,f\rangle_P$. This is because $\langle L^s_{n,\varepsilon}f,f\rangle_n$ 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 ε is sufficiently large. This is where the lower bound $Cn^{-1/(2(s-1)+d)} < \varepsilon$ in the statement of Proposition 4 comes from.

The derivation of the second inequality in (3.16) differs in the technical details based on whether s is even or odd; we focus our discussion on the case where s is odd although the general ideas are the same in either case. When s is odd, in a nutshell we establish the desired upper bound by arguing that the iterated non-local Laplacian $L^{(s-1)/2}_{P,\varepsilon}f$ satisfies the approximate equality

$$L^{(s-1)/2}_{P,\varepsilon}f \approx \sigma_\eta^{(s-1)/2}\Delta_P^{(s-1)/2}f,$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

and then applying the known upper bound $\langle L_{P,\varepsilon} g, g\rangle_P \leq C\|g\|^2_{H^1(\mathscr{X})}$ with $g = \Delta_P^{(s-1)/2} f$. Here $\sigma_\eta := \frac{1}{d}\int_{\mathbb{R}^d} \|x\|^2 \eta(\|x\|)\, dx$ is a constant that is finite under the assumptions of **(K1)**. The approximate equality above relies on Taylor expansions of both f and p, which is the reason we require that $p \in C^{s-1}(\mathscr{X})$. The approximate equality also breaks down near the boundary of $\mathscr{X}$, and instead we show $L_{P,\varepsilon}^{(s-1)/2} f$ is close to 0 using the zero-trace property of f.

An important aspect of the ultimate result is that, since we are ultimately interested only in an upper bound on the rate of convergence of PCR-LE, it is enough to have an upper bound on the graph-Sobolev seminorm that has the right dependence on $\|f\|^2_{H^s(\mathscr{X})}$ and does not depend on n. This means we do not have to show that $\langle L^s_{n,\varepsilon} f, f\rangle_n \approx \|f\|^2_{H^s(\mathscr{X})}$ which would be substantially more challenging.

Neighbourhood graph eigenvalues. On the other hand, several recent works [15, 16, 27] have analysed the convergence of graph eigenvalues λ_k towards ρ_k, defined in (2.4). They provide explicit bounds on the relative error $|\lambda_k - \rho_k|/\rho_k$, which show that the relative error is small for sufficiently large n and small ε. Crucially, these guarantees hold simultaneously for all $1 \leq k \leq K$ as long as $\rho_K = O(\varepsilon^{-2})$. These results are actually stronger than are necessary to establish Theorems 1–4—in order to get rate-optimality, we need only show that for the relevant values of K, $\lambda_K/\rho_K = \Omega_P(1)$—but the guarantees hold only when $\mathscr{X}$ is a manifold without boundary.

In the case where $\mathscr{X}$ is assumed to have a boundary, the graph Laplacian $L_{n,\varepsilon}$ is a reasonable approximation of the operator Δ_P only at points $x \in \mathscr{X}$ for which $B(x,\varepsilon) \subseteq \mathscr{X}$. In contrast, at points x near the boundary of $\mathscr{X}$, the graph Laplacian is known to approximate a different operator altogether [11].[7] This renders analysis of λ_k substantially more challenging, since its continuum limit is not ρ_k. Rather than analysing the convergence of λ_k, we will instead use Lemma 2 of [31], whose assumptions match our own, and who give a weaker bound on the ratio λ_k/ρ_k that will nevertheless suffice for our purposes.

Proposition 5. (Lemma 2 of [31]). In the flat Euclidean setting, 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

$$\lambda_k \geq c \cdot \min\left\{\rho_k, \frac{1}{\varepsilon^2}\right\} \quad \text{for all } 1 \leq k \leq n, \tag{3.17}$$

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

By our assumptions on P, $\rho_0 = \lambda_0 = 0$. Furthermore, Weyl's Law (B4) tells us that under the assumptions of the flat Euclidean setting $k^{2/d} \asymp \rho_k$ for all $k \in \mathbb{N}, k > 1$. Combining these statements with (3.17), we conclude that with high probability $\lambda_K \gtrsim K^{2/d}$ so long as $K \lesssim \varepsilon^{-d}$.

Empirical norm. Finally, in Proposition 6 we establish that a one-sided bound of the form $\|f_0\|_n^2 \gtrsim \|f_0\|_P^2$ holds whenever $\|f_0\|_P^2$ is itself sufficiently large.

Proposition 6. In the flat Euclidean setting, suppose additionally that $f \in H^s(\mathscr{X}, M)$ 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

[7] This is directly related to the boundary bias of kernel smoothing, since the graph Laplacian can be viewed as a kernel-based estimator of Δ_P.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

any $\delta > 0$: if

$$\|f\|_P \geq CM\left(\frac{1}{\delta n}\right)^{s/d} \tag{3.18}$$

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

$$\|f\|_n^2 \geq \frac{1}{2}\|f_0\|_P^2. \tag{3.19}$$

To prove Proposition 6, we use a Gagliardo–Nirenberg interpolation inequality (see e.g. Theorem 12.83 of Leoni [47]) to control the fourth moment of $f \in H^s(\mathscr{X})$ in terms of $\|f\|_P$ and $|f|_{H^s(\mathscr{X})}$, then invoke a one-sided Bernstein's inequality as in [70, section 14.2]. Note carefully that the statement (3.19) is *not* a uniform guarantee over all $f \in H^s(\mathscr{X}; M)$. Indeed, such a statement cannot hold in the sub-critical regime ($2s \leq d$).[8] Fortunately, a pointwise bound—meaning a bound that holds with high probability for a single $f \in H^s(\mathscr{X})$—is sufficient for our purposes.

Finally, invoking the bounds of Propositions 3–6 inside the bias-variance tradeoffs (3.11) and (3.12) and then choosing K to balance bias and variance (when possible) leads to the conclusions of Theorems 1–5.

As pointed out by a reviewer, the proof techniques outlined above rely heavily on special properties of the Euclidean norm. To analyse the statistical error of PCR-LE for more general types of losses, very different approaches might be needed. We think the question of whether PCR-LE would attain the optimal rates for, say, ℓ^p losses with $p > 2$, would be an interesting direction for future work.

3.4 *Computational considerations*

Our focus in this paper is primarily on the statistical efficiency of PCR-LE. In this section we briefly discuss some computational aspects of the method.

Sparsification. First, we note that one very nice aspect of neighbourhood graphs is that the graph Laplacian $L_{n,\varepsilon}$ is typically quite sparse. For instance, choosing $\varepsilon \asymp (\log n/n)^{1/d}$ will result in a Laplacian with $O(n \log n)$ non-zero entries. Our theory shows that this choice of ε results in optimal estimators and tests when $s = 1$.

However, in the higher order case ($s \geq 2$), our optimality results hold only under meaningfully larger choices of ε (see **(P3)** and **(P4)**), and the resulting neighbourhood graph G will be much denser: the average degree will grow polynomially in the sample size n as $n \to \infty$, so there will be more non-zero entries in the graph Laplacian, which increases the computational burden of PCR-LE. To address this issue, in Appendix J we review some approaches to *spectral sparsification*, in which one efficiently computes a sparse graph $\check{G}$ that approximates G in a spectral sense. The hope is that the PCR-LE estimator $\check{f}$, computed with respect to the sparsified graph $\check{G}$, has similar statistical properties as $\widehat{f}$ while being much faster to compute. To that end, we provide upper bounds on $\|\check{f} - f_0\|_n^2$, which show that under

[8] This is because in the sub-critical regime, for any set of points $\{x_1, \ldots, x_n\}$ there exists a sequence of functions $\{f_k : k \in \mathbb{N}\} \subset H^s(\mathscr{X}; 1)$ satisfying $f_k(x_i) = 1$ for each $i = 1, \ldots, n$—and therefore $\|f_k\|_n^2 = 1$—but for which $\|f_k\|_P^2 \to 0$ as $k \to \infty$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

mild conditions on $\check{G}$—provably achieved by many spectral sparsification algorithms—the estimator $\check{f}$ achieves the same rates of convergence as $\widehat{f}$.

Beyond eigendecomposition. As currently defined PCR-LE requires taking a full eigendecomposition of L. Naively this requires $O(n^3)$ time, and when n is large this may be prohibitive. However Frostig et al. [25] show that in a general setting—outside of the context of graphs or graph Laplacians—it is possible to compute an approximate solution to PCR by solving a 'few' ridge regression problems, without ever needing to find the spectral decomposition. In the case of PCR-LE each ridge regression is equivalent to Laplacian regularization, which can be approximately computed in time nearly-linear in the number of edges in G [68]; this means that when G is sparse or has been sparsified, Laplacian regularization can be solved in $\widetilde{O}(n)$ time. [Here $\widetilde{O}(\cdot)$ hides $\mathrm{poly}(\log n)$ factors.]

Unfortunately, in Frostig et al. [25] the number of ridge regressions used to approximately compute PCR depends inversely on the spectral gap $\lambda_{K+1}/(\lambda_{K+1}-\lambda_K)$ which for neighbourhood graphs is usually quite small. The subsequent work of Allen-Zhu and Li [2, Jin and Sidford 41] sharpen the dependence on the spectral gap, and the topic remains an area of active research.

Another interesting idea is to consider whether graph-based estimators besides PCR-LE could achieve the optimal rates of convergence. An obvious candidate would be penalized regression involving the Laplacian raised to a certain power, i.e. the solution to

$$\underset{f \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{n} (Y_i - f_i)^2 + \lambda f^\top L^s f.$$

Obviously this can also be solved without needing a full eigendecomposition of the Laplacian. However we are unable to show that it is statistically optimal except in a very few special cases; namely when $s = 1$ and $d = 1, 2, 3, 4$.

3.5 *Out-of-sample error*

LE is defined only at the observed samples and so our statements on the estimation error of PCR-LE are based on mean-squared error at the data. Although it is common to measure error in this way, in the context of random design regression it is arguably a bit unnatural; for instance, in-sample mean-squared error has no formal relationship to prediction risk. Also, the lower bounds on estimation error for random design regression—discussed in Section 2.3—are with respect to integrated L^2 loss. So as a formal matter our upper bounds cannot be compared with known lower bounds, at least in the sub-critical regime where $s \leq d/2$ and there is no coupling between $\|\cdot\|_n$ and $\|\cdot\|_P$.

In Green [30], we discussed a post-processing scheme based on kernel smoothing that takes an estimator $\widehat{f}$ defined only at the design points, and gives the function

$$T_{h,n}\widehat{f}(x) = \begin{cases} \frac{1}{\sum_{i=1}^{n} \psi(\|X_i - x\|/h)} \sum_{i=1}^{n} \psi\left(\frac{\|X_i - x\|}{h}\right) \widehat{f}_i, & \text{if } \sum_{i=1}^{n} \psi\left(\frac{\|X_i - x\|}{h}\right) > 0 \\ 0, \quad \text{otherwise.} \end{cases}$$

This is well defined for any $x \in \mathbb{R}^d$.

We showed that under appropriate conditions the resulting $T_{h,n}\widehat{f}$ has comparable $L^2(P)$ error to the in-sample mean-squared error of $\widehat{f}$. These conditions are exactly the ones of Theorem 1 (for $f_0 \in H^1(\mathscr{X})$) or Theorem 3 (for $f_0 \in H_0^s(\mathscr{X}), s \geq 2$), plus the following.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

(K2) 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

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

If $s \geq 2$ then ψ additionally satisfies the higher order kernel conditions,

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

(P5) For constants c_0 and C_0, the bandwidth parameter h satisfies

$$C_0\left(\frac{\log(1/h)}{n}\right)^{1/d} \leq h \leq c_0 n^{-1/(2s+d)}.$$

Combined with our results on the in-sample error of PCR-LE, Theorem 19 of Green [30] yields the following upper bound on estimation error.

THEOREM 6. In the flat Euclidean setting, 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 $1 - \delta - Cn\exp\{-cn\varepsilon^d\} - C\exp\{-cnh^d\}$, for all $n \geq N$, and for any $\delta \in (0, 1)$.

- If $f_0 \in H^1(\mathscr{X}; M)$ for $Cn^{-1/2} \leq M \leq cn^{1/d}$, the LE 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 $h = n^{-1/(2+d)}$ and kernel ψ that satisfies **(K2)**, then
$$\|T_{h,n}\widehat{f} - f_0\|_P^2 \leq \frac{C}{\delta} M^2 (M^2 n)^{-2/(2+d)}.$$
- If $f_0 \in H_0^s(\mathscr{X}; M)$ for $Cn^{-1/2} \leq M \leq cn^{s/d}$ and $p \in C^{s-1}(\mathscr{X})$ for some $s \in \mathbb{N}, s > 1$, and the LE 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 **(K2)**, then
$$\|T_{h,n}\widehat{f} - f_0\|_P^2 \leq \frac{C}{\delta} M^2 (M^2 n)^{-2s/(2s+d)}.$$

The implication is that $T_{h,n}\widehat{f}$ is a rate-optimal estimator with loss measured in $L^2(P)$ norm.

Another advantage of extending $\widehat{f}$ to a function defined out-of-sample is that in principle it allows for tuning the hyperparameters of PCR-LE using cross-validation or other sample-splitting techniques.

4. Manifold setting

In this section we consider the manifold setting, where it is known that the minimax rates depend only on the intrinsic dimension m; more specifically, [3, 12] show that for functions with Hölder smoothness

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

s, the minimax estimation rate is $n^{-2s/(2s+m)}$ and the testing rate is $n^{-4s/(4s+m)}$.[9] On the other hand, a theory has been developed [5–7, 10, 49, 50] establishing that the neighbourhood graph G can 'learn' the manifold $\mathscr{X}$ in various senses, so long as $\mathscr{X}$ is locally linear. We build on this work by showing that when P is supported on a manifold $\mathscr{X}$ and $f_0 \in H^s(\mathscr{X})$, PCR-LE achieves the faster minimax estimation and testing rates.

4.1 *Upper bounds*

This section will proceed similarly to Section 3.2, with two differences. First when $\mathscr{X}$ is without boundary, as is assumed in the manifold setting, it is easy to deal with the first-order ($s = 1$) and higher order ($s > 1$) cases all at once. Second and more importantly, we will establish PCR-LE is optimal only when the regression function $f_0 \in H^s(\mathscr{X}; M)$ for $s \in \{1, 2, 3\}$.

Estimation with PCR-LE. 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.

(K4) The kernel function η is a non-increasing function supported on a subset of $[0, 1]$. Its restriction to $[0, 1]$ is Lipschitz, and $\eta(1) > 0$. Additionally, it is normalized so that

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

(P5) The number of eigenvectors is given by

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

If $K < n$ then additionally

$$C_0 \max\left\{\left(\frac{\log}{n}\right)^{1/m}, n^{-1/(2(s-1)+m)}\right\} \le \varepsilon \le c_0 \min\{1, K^{-1/m}\}. \tag{4.1}$$

THEOREM 7. In the manifold setting, suppose additionally $f_0 \in H^s(\mathscr{X}, M)$ and $p \in C^{s-1}(\mathscr{X})$ for $s \le 3$. There exist constants c, C and N that do not depend on f_0, such that the following statement holds all for all $n \ge N$ and for any $\delta \in (0, 1)$: if the PCR-LE estimator $\widehat{f}$ is computed with a kernel η satisfying **(K4)**, and parameters ε and K satisfying **(P5)**, then

$$\|\widehat{f} - f_0\|_n^2 \le C\left(\frac{1}{\delta} M^2 (M^2 n)^{-2s/(2s+m)} \wedge 1\right) \vee \frac{1}{n}, \tag{4.2}$$

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

[9] Although [3] considers density testing, usual arguments regarding equivalence of experiments [14] imply that the same rates apply to regression testing.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Testing with PCR-LE. 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 number of eigenvectors is given by

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

If $K < n$ then additionally the graph radius ε satisfies (4.1).

THEOREM 8. Fix $a, b \in (0,1)$. The PCR-LE test φ, computed with threshold t_a, is a level-a test: $\mathbb{E}_0[\varphi] \leq a$. In the manifold setting, suppose additionally $f_0 \in H^s(\mathscr{X}, M)$, that $p \in C^{s-1}(\mathscr{X})$ 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 $n \geq N$: if the PCR-LE test φ is computed with a kernel η satisfying **(K4)**, and parameters ε and K satisfying **(P6)**, and if f_0 satisfies

$$\|f_0\|_P^2 \geq \frac{C}{b}\left(\left(M^2(M^2n)^{-4s/(4s+m)} \wedge n^{-1/2}\right)\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right] \vee \frac{M^2}{bn^{2s/m}}\right) \vee \frac{1}{n}, \tag{4.3}$$

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

Focusing on the case $M \asymp 1$,[10] the upper bounds in Theorems 7 and 8 imply that PCR-LE attain the optimal rates of convergence over Sobolev balls $H^s(\mathscr{X})$ for $s \in \{1, 2, 3\}$.

Unlike in the full-dimensional case, in the manifold setting our upper bounds on the estimation and testing error of PCR-LE do not match the minimax rate when $s \geq 4$. In this case, the containment $H^s(\mathscr{X}; 1) \subset H^3(\mathscr{X}; 1)$ implies that the PCR-LE estimator $\widehat{f}$ has in-sample mean-squared error of at most of the order of $n^{-6/(6+m)}$, and that the PCR-LE test has small Type II error whenever $\|f_0\|_P^2 \gtrsim n^{-12/(12+m)}$; however, these are slower than the minimax rates.

We now explain this difference between the flat Euclidean and manifold settings. At a high level, thinking of the graph G as an estimate of the manifold $\mathscr{X}$, we incur some error 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 $x, z \in \mathscr{X}$ that are sufficiently close to each other and far from the boundary of $\mathscr{X}$. This extra error incurred in the manifold setting 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 $\mathscr{X}$—and in particular its geodesics—are unknown. Likewise, a population-level spectral series 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 $\mathscr{X}$ is known. It is not clear whether this gap between population-level spectral series regression and the PCR-LE estimator is real, or a product of loose upper bounds.

Finally, as in the full-dimensional case, when the intrinsic dimension $m > 4$ we cannot choose the graph radius ε and number of eigenvectors K to optimally balance testing bias and variance. Instead, reasoning as in the proof of Theorem 5 shows that when $1 \leq s \leq 3$, the PCR-LE test has power against

[10] To the best of our knowledge, the minimax rates for general M in the manifold setting have not been worked out.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

alternatives with $L^2(P)$ norm satisfying the inequality in (3.10), except with the ambient dimension d replaced by m.

4.2 *Analysis*

The high-level strategy used to prove Theorems 7 and 8 is the same as in the flat-Euclidean setting. More specifically, we will use precisely the same bias-variance decompositions (3.11) (for estimation) and (3.12) (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 7–9.

Proposition 7. In the manifold setting, suppose additionally that $f \in H^s(\mathscr{X}; M)$ and $p \in C^{s-1}(\mathscr{X})$ for $s = 1, 2$ or 3. Then there exist constants c and C that do not depend on f, n or M such that the following statement holds for any $\delta \in (0, 1)$: if η satisfies **(K4)** and $Cn^{-1/(2(s-1)+m)} < \varepsilon < c$, then

$$\langle L_{n,\varepsilon}^s f, f\rangle_n \leq \frac{C}{\delta}\|f\|_{H^s(\mathscr{X})}^2, \tag{4.4}$$

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

In the manifold setting, appropriate bounds on the graph eigenvalues λ_k have already been derived in [15, 28, 29]. The precise result we use is a direct consequence of Theorem 2.4 of [16].

Proposition 8. In the manifold setting, there exist constants c and C such that the following statement holds: if η satisfies **(K4)** and $C(\log n/n)^{1/m} < \varepsilon < c$, then

$$\lambda_k \geq c \cdot \min\left\{k^{2/m}, \frac{1}{\varepsilon^2}\right\} \quad \text{for all } 1 \leq k \leq n, \tag{4.5}$$

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

(For the specific computation used to deduce Proposition 8 from Theorem 2.4 of [16], see [31].)

Finally, we have the following lower bound on the empirical norm $\|f\|_n$.

Proposition 9. In the manifold setting, suppose additionally that $f_0 \in H^s(\mathscr{X}, M)$ 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

$$\|f_0\|_P \geq CM\left(\frac{1}{\delta n}\right)^{s/m}, \tag{4.6}$$

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

$$\|f_0\|_n^2 \geq \frac{1}{2}\|f_0\|_P^2. \tag{4.7}$$

We prove Proposition 9 in a parallel manner to its flat Euclidean counterpart (Proposition 6), by first using a Gagliardo–Nirenberg inequality to upper bound the $L^4(\mathscr{X})$ norm of a Sobolev function defined

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

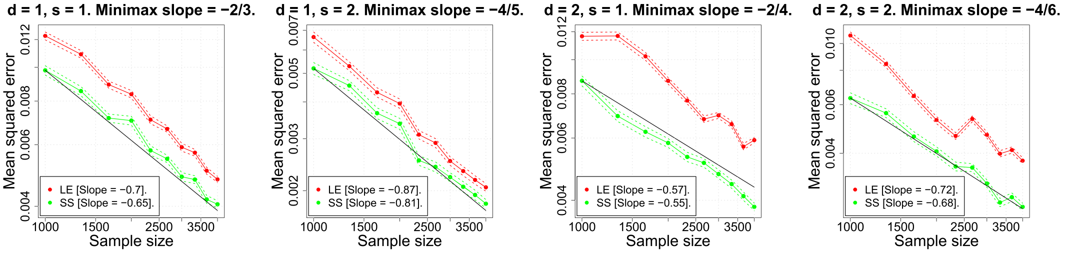

FIG. 1. In-sample mean squared error (mse) of PCR-LE (LE) vs. population-level spectral series (SS) estimator, as a function of sample size n. Each plot is on the log–log scale, and the results are averaged over 400 repetitions. All estimators are tuned for optimal average mse, separately at each value of n. The black line shows the minimax rate (in slope only; the intercept is chosen to match the observed error).

on a compact Riemannian manifold, and then applying a one-sided Bernstein's inequality. Finally, combining Propositions 7–9 with the conditional-on-design bias-variance decompositions (3.11) and (3.12) leads to the conclusions of Theorems 7 and 8.

5. Experiments

In this section we empirically demonstrate that the PCR-LE estimator and test are reasonably good alternatives to population-level spectral series methods, even at moderate sample sizes n. In order to compare the two approaches, in our experiments we stick to the simple case where the design distribution P is the uniform distribution over $\mathcal{X} = [-1, 1]^d$, and we have simple closed-form expressions for the eigenfunctions of Δ_P. In general, it is not easy to analytically compute these eigenfunctions, which is part of the appeal of LE and PCR-LE.

Estimation. In our first experiment, we compare the mean-squared error of the PCR-LE estimator $\widehat{f}$ to that of its population-level counterpart $\widetilde{f}$. 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 (1.3) with regression function $f_0 = M/\rho_K^{s/2} \cdot \psi_K$ for $K \asymp n^{d/(2s+d)}$ (the pre-factor $M/\rho_K^{s/2}$ is chosen so that $|f_0|^2_{H^s(\mathcal{X})} = M^2$). In Fig. 1 we show the in-sample mean-squared error of the two estimators as a function of n, for different dimensions d and order of smoothness s. We see that both estimators have mean-squared error converging to zero at roughly the minimax rate. While unsurprisingly the population-level spectral series estimator has the smaller error, generally speaking the error of PCR-LE approaches that of the population-level spectral series method as n gets larger.

Testing. In our second experiment, we compare the PCR-LE test φ against the population-level spectral series test $\widetilde{\varphi}$.[11] The set-up is generally the same as that of our first experiment, but the details are necessarily somewhat more complicated. First we take a collection $\mathcal{F} = \{M/\rho_k^{s/2}\psi_k\}_{k=1}^n$ of functions $H^1(\mathcal{X}; M)$. Then, for each $f_0 \in \mathcal{F}$, we run a given test ϕ (either the PCR-LE test $\phi = \varphi$, or the population-level spectral series test $\phi = \widetilde{\varphi}$) and record whether it was a false negative or true positive. We repeat this process over 100 replications, giving a Monte Carlo estimate of the type II error $E_{f_0}[1-\phi]$ for each $f_0 \in \mathcal{F}$. Finally, we measure quality of the test by reporting the smallest value of $\|f_0\|_P^2$ such that $E_{f_0}[1 - \phi] \leq b$.

[11] In this experiment, we calibrate the tests by simulation rather than using theoretically-motivated cut-offs.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

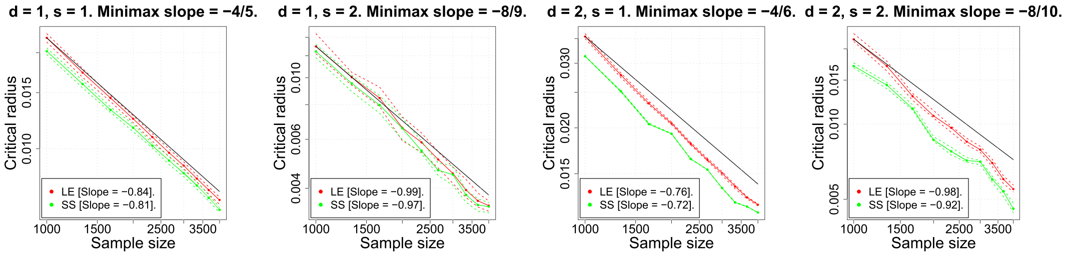

FIG. 2. Worst-case testing risk for PCR-LE (LE) and spectral series (SP) tests, as a function of sample size n. Plots are on the same scale as Fig. 1, and the black line shows the minimax rate. All tests are set to have .05 Type I error, and are calibrated by simulation under the null.

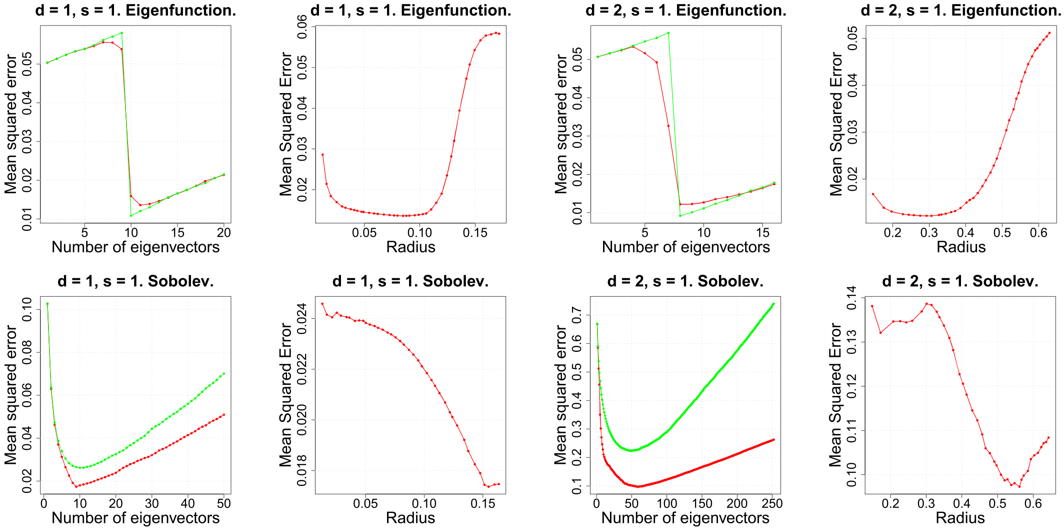

FIG. 3. Mean squared error of PCR-LE (red), and population-level spectral series (green) estimators as a function of tuning parameters. Top row: the same regression function f_0 as used in Fig. 1. Bottom row: the regression function $f_0 \propto \sum_k 1/\rho_k^{1/2}\psi_k$. For all experiments, the 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.

In Fig. 2, we see that both the PCR-LE and population-level spectral series tests perform similarly, and converge at roughly the minimax rate.

Tuning parameters. Our first two experiments demonstrate that PCR-LE methods have comparable statistical performance to population-level spectral series methods. PCR-LE depends on two tuning parameters, and in our final experiment we investigate the importance of both, focusing now on estimation. In Fig. 3, we see how the mean-squared error of PCR-LE 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.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

6. Discussion

In this work, we have derived upper bounds on the rates of convergence for regression with PCR-LE, which imply that in various settings the PCR-LE estimator and test are minimax rate-optimal over Sobolev classes. Importantly, these upper bounds hold under non-parametric conditions on the design density p, and allow for p to be unknown and, potentially, supported on a low-dimensional manifold. Our results help explain the practical success of methods which leverage graph Laplacian eigenvectors for regression. They also distinguish such methods from more traditional spectral series procedures, which rely on a density-dependent basis and thus require the density be known a priori.

Of course, there do exist other methods for non-parametric regression which achieve optimal rates of convergence under similar (or indeed weaker) conditions on p. These include other graph-based approaches—e.g. Laplacian smoothing—methods besides spectral series methods—e.g. kernel smoothing, local polynomial regression, thin-plate splines—and continuum spectral projection methods which use the eigenfunctions of an operator defined independently of p. To be clear, we do not advocate PCR-LE over these alternatives. Rather, we view our results as theoretically justifying a place for regression using LE in the non-parametric regression toolbox.

That being said, PCR-LE does have certain advantages over each of the aforementioned approaches. We now conclude by outlining some of these advantages (limiting our discussion to estimation):

- *Optimality when $d \geq 5$.* As mentioned in the introduction, Laplacian smoothing (defined via (1.5)) provably achieves minimax optimal rates over $H^1(\mathscr{X})$ only when $d \in \{1,2,3,4\}$ [31, 53]. In contrast, PCR-LE is optimal over $H^1(\mathscr{X})$ for all dimensions d, and also over the higher order Sobolev spaces $H^s(\mathscr{X})$.

- *Dependence on intrinsic dimension.* When the design distribution is non-uniform, an oft-recommended alternative to population-level spectral series regression is to run OLS using eigenfunctions of a density-independent differential operator. As a concrete example, let Δ be the unweighted Laplacian operator on $\mathbb{R}^d$, $\Delta = \sum_{i=1}^{d} \partial^2 f/\partial x_i^2$. Denoting the eigenfunctions of Δ (under Neumann boundary conditions) by $\phi_1, \phi_2, \ldots$, and letting $\Phi \in \mathbb{R}^{n \times K}$ be the matrix with entries $\Phi_{ik} = \phi_k(X_i)$ and columns $\Phi_1, \ldots, \Phi_K$, one could compute an estimator by solving the following OLS problem:

$$\underset{f \in \mathrm{span}\{\Phi_1,\ldots,\Phi_K\}}{\mathrm{minimize}} \ \|\mathbf{Y} - f\|_n^2.$$

  Unlike with spectral series regression, this approach can produce reasonable estimates even when the sampled eigenfunctions $(\phi_k(X_1), \ldots, \phi_k(X_n)) \in \mathbb{R}^n$ are not approximately orthogonal. Indeed in the flat Euclidean setting such a method will in fact be minimax rate-optimal, though the upper bounds may come with undesirably large constants if p is very non-uniform. However in the manifold setting we know of no guarantees for the method, and suspect it may converge at suboptimal rates or even be inconsistent. The justification for this claim is that the eigenfunctions ϕ_k have no underlying relationship to the Sobolev space $H^s(\mathscr{X})$ except when $\mathscr{X}$ is a full-dimension set in $\mathbb{R}^d$. In contrast, PCR-LE uses features which are empirical approximations to eigenfunctions ψ_k of the density-weighted Laplace–Beltrami operator Δ_P. The eigenfunctions of Δ_P are appropriately adapted to the geometry of the manifold $\mathscr{X}$, and as a result PCR-LE is consistent and in certain cases minimax optimal, as we have shown.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

- *Dependence on design density*. In Appendix H, we give a simple univariate example of a sequence of densities and regression functions $\{(p^{(n)}, f_0^{(n)} : n \in \mathbb{N}\}$ such that the expected in-sample mean squared error of PCR-LE is smaller than that of either kernel smoothing or least squares using eigenfunctions of Δ. This is possible because PCR-LE induces a completely different bias than these latter two methods. In particular, when f_0 and p satisfy the so-called *cluster assumption*—meaning f_0 is piecewise constant in high-density regions (clusters) of p—then the bias of PCR-LE can be much smaller (for equivalent levels of variance) than that of kernel smoothing or least-squares with eigenfunctions of Δ.
- We emphasize that this does not contradict the well-known optimality properties of, for example, kernel smoothing over Hölder balls. Rather, in the standard non-parametric regression set-up—which we adopt in the main part of this paper, and in which P is assumed to be equivalent to Lebesgue measure—the biases of PCR-LE and kernel smoothing happen to be equivalent. But when P is sufficiently non-uniform, this is no longer the case.

Grounding each of these three points on a firmer and more complete theoretical basis would be, in our view, a valuable direction for future work.

Acknowledgements

The authors would like to thank the reviewers of this manuscript for their constructive suggestions. AG would also like to thank Krishnakumar Balakrishnan, Zhaoyang Shi and Addison Hu for catching several errors in an earlier draft.

Funding

This work was partially supported by funding from the NSF grants DMS-2113684 and DMS-2310632, as well as an Amazon AI and a Google Research Scholar Award to SB. AG and RJT were supported by ONR grant N00014- 20-1-2787.

Data Availability Statement

Code used to generate the synthetic data and plots in Section 5 is available at this\ignorespaceslink.

References

1. Aamari, E., Kim, J., Chazal, F., Michel, B., Rinaldo, A. & Wasserman, L. (2019) Estimating the reach of a manifold. *Electron. J. Stat.*, **13**, 1359–1399.
2. Allen-Zhu, Z. & Li, Y. (2017) Faster principal component regression and stable matrix chebyshev approximation. *International Conference on Machine Learning*. Proceedings of Machine Learning Research, pp. 107–115.
3. Arias-Castro, E., Pelletier, B. & Saligrama, V. (2018) Remember the curse of dimensionality: the case of goodness-of-fit testing in arbitrary dimension. *J. Nonparametr. Stat.*, **30**, 448–471.
4. Aubin, T. (2012) *Nonlinear analysis on manifolds. Monge-Ampere equations*, vol. **252**. Springer Science & Business Media.
5. Balakrishnan, S., Rinaldo, A., Sheehy, D., Singh, A. & Wasserman, L. Minimax rates for homology inference. In *International Conference on Artificial Intelligence and Statistics*, volume **22**, 2012.
6. Balakrishnan, S., Narayanan, S., Rinaldo, A., Singh, A. & Wasserman, L. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems, volume* **26**, 2013.
7. Belkin, M. (2003) *Problems of Learning on Manifolds*, PhD thesis. University of Chicago.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

8. Belkin, M. & Niyogi, P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, **15**, 1373–1396.
9. Belkin, M. & Niyogi, P. (2007) Convergence of Laplacian eigenmaps. *In Advances in Neural Information Processing Systems, volume*, **20**.
10. Belkin, M. & Niyogi, P. (2008) Towards a theoretical foundation for Laplacian-based manifold methods. *J. Comput. Syst. Sci.*, **74**, 1289–1308.
11. Belkin, M., Que, Q., Wang, Y. & Zhou, X. Toward understanding complex spaces: Graph laplacians on manifolds with singularities and boundaries. In Mannor, S., Srebro, N. & Williamson, R. C., 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.
12. Bickel, P. J. & Li, B. (2007) Local polynomial regression on unknown manifolds. *Complex datasets and inverse problems*, vol. **54**. Institute of Mathematical Statistics, pp. 177–186.
13. Bousquet, O., Chapelle, O. & Hein, M. Measure based regularization. In *Advances in Neural Information Processing Systems*, volume **16**, 2004.
14. Brown, L. D. & Low, M. G. (1996) Asymptotic equivalence of nonparametric regression and white noise. *Ann. Stat.*, **24**, 2384, 12–2398.
15. Burago, D., Ivanov, S. & Kurylev, Y. (2014) A graph discretization of the Laplace-Beltrami operator. *J. Spectr. Theory*, **4**, 675–714.
16. Calder, J. & García Trillos, N. (2019) *Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs* arXiv preprint arXiv:1910.13476.
17. Calder, J. & Slepčev, D. (2019) Properly-weighted graph laplacian for semi-supervised learning. *Appl. Math. Optim.*, 1–49.
18. Calder, J., Slepčev, D. & Thorpe, M. (2020) *Rates of convergence for laplacian semi-supervised learning with low labeling rates* arXiv preprint arXiv:2006.02765.
19. Cheng, X. & Wu, N. (2021) *Eigen-convergence of gaussian kernelized graph laplacian by manifold heat interpolation* arXiv preprint arXiv:2101.09875.
20. Dhillon, P. S., Foster, D. P., Kakade, S. M. & Ungar, L. H. (2013) A risk comparison of ordinary least squares vs ridge regression. *J. Mach. Learn. Res.*, **14**, 1505–1511.
21. Dicker, L. H., Foster, D. P. & Hsu, D. (2017) Kernel ridge vs. principal component regression: minimax bounds and the qualification of regularization operators. *Electron. J. Stat.*, **11**, 1022–1047.
22. Dunlop, M. M., Slepčev, D., Stuart, A. M. & Thorpe, M. (2020) Large data and zero noise limits of graph-based semi-supervised learning algorithms. *Appl. Comput. Harmon. Anal.*, **49**, 655–697.
23. Dunson, D. B., Wu, H.-T. & Wu, N. (2021) Spectral convergence of graph laplacian and heat kernel reconstruction in l-infinity from random samples. *Appl. Comput. Harmon. Anal.*.
24. Evans, L. C. & Gariepy, R. F. (2015) *Measure theory and fine properties of functions*. Chapman and Hall/CRC.
25. Frostig, R., Musco, C., Musco, C. & Sidford, A. (2016) Principal component projection without principal component analysis. *International Conference on Machine Learning*. PMLR, pp. 2349–2357.
26. García Trillos, N. & Murray, R. W. (2020) A maximum principle argument for the uniform convergence of graph Laplacian regressors. *SIAM journal on mathematics of data. Science*, **2**, 705–739.
27. García Trillos, N. & Slepčev, D. (2018) A variational approach to the consistency of spectral clustering. *Appl. Comput. Harmon. Anal.*, **45**, 239–281.
28. García Trillos, N., Gerlach, M., Hein, M. & Slepcev, D. (2019a) Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Found. Comput. Math.*, **20**, 1–61.
29. Trillos, N. G., Hoffmann, F. & Hosseini, B. (2019b) *Geometric structure of graph laplacian embeddings* arXiv preprint arXiv:1901.10651.
30. Green, A. (2021) *Statistical Guarantees for Spectral Methods on Neighborhood Graphs* PhD thesis,. Carnegie Mellon University.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

31. Green, A., Balakrishnan, S. & Tibshirani, R. Minimax optimal regression over sobolev spaces via laplacian regularization on neighborhood graphs. In *A. Banerjee and K. Fukumizu, editors, Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume **130** of *Proceedings of Machine Learning Research*, pages 2602–2610. PMLR, 13–15 Apr 2021.

32. Guerre, E. & Lavergne, P. (2002) Optimal minimax rates for nonparametric specification testing in regression models. *Econom. Theory*, **18**, 1139–1171.

33. Györfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2006) *A Distribution-Free Theory of Nonparametric Regression*. Springer.

34. Hebey, E. (1996) *Sobolev spaces on Riemannian manifolds*, vol. **1635**. Springer Science & Business Media.

35. Hoffmann, F., Hosseini, B., Oberai, A. A. & Stuart, A. M. (2019) *Spectral analysis of weighted laplacians arising in data clustering* arXiv preprint arXiv:1909.06389.

36. Hörmander, L. (2007) *The analysis of linear partial differential operators III: Pseudo-differential operators*. Springer Science & Business Media.

37. Hsu, D., Kakade, S. M. & Zhang, T. Random design analysis of ridge regression. In *Conference on learning theory*, pages **9–1**, 2012.

38. Hütter, J.-C. and Rigollet, P. Optimal rates for total variation denoising. In *Conference on Learning Theory, volume* **29**, 2016.

39. Ingster, Y. I. & Sapatinas, T. (2009) Minimax goodness-of-fit testing in multivariate nonparametric regression. *Math. Methods Stat.*, **18**, 241–269.

40. Ingster, Y. I. & Suslina, I. A. (2012) *Nonparametric goodness-of-fit testing under Gaussian models*. Springer Science & Business Media.

41. Jin, Y. and Sidford, A. Principal component projection and regression in nearly linear time through asymmetric svrg. In *Advances in Neural Information Processing Systems, volume* **32**, 2019.

42. Kirichenko, A. & Zanten, H. van (2017) Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *Electron. J. Stat.*, **11**, 891–915.

43. Kirichenko, A. & Zanten, H. van (2018) Minimax lower bounds for function estimation on graphs. *Electron. J. Stat.*, **12**, 651–666.

44. Koltchinskii, V. & Gine, E. (2000) Random matrix approximation of spectra of integral operators. *Bernoulli*, **6**, 113–167 02.

45. Laurent, B. & Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.*, 1302–1338.

46. Lee, A. B. & Izbicki, R. (2016) A spectral series approach to high-dimensional nonparametric regression. *Electron. J. Stat.*, **10**, 423–463.

47. Leoni, G. (2017) *A first Course in Sobolev Spaces*. American Mathematical Society.

48. Nadler, B., Srebro, N. & Zhou, X. Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data. In *Neural Information Processing Systems*, volume **19**, 2009.

49. Niyogi, P. (2013) Manifold regularization and semi-supervised learning: some theoretical analyses. *J. Mach. Learn. Res.*, **14**, 1229–1250.

50. Niyogi, P., Smale, S. & Weinberger, S. (2008) Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, **39**, 419–441.

51. Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Stat., pages*, **1215–1230**.

52. Sadhanala, V., Wang, Y.-X. & Tibshirani, R. Graph sparsification approaches for laplacian smoothing. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume **51**, pages 1250–1259, 2016a.

53. Sadhanala, V., Wang, Y.-X. & Tibshirani, R. J. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, volume **29**, 2016b.

54. Sadhanala, V., Wang, Y.-X., Sharpnack, J. L. & Tibshirani, R. J. Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems*, volume **30**, 2017.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

55. Sharpnack, J. and Singh, A. Identifying graph-structured activation patterns in networks. In *Advances in Neural Information Processing Systems*, volume **23**, 2010.
56. Sharpnack, J., Krishnamurthy, A. & Singh, A. Near-optimal anomaly detection in graphs using Lovasz extended scan statistic. In *Advances in Neural Information Processing Systems*, volume **26**, 2013a.
57. Sharpnack, J., Singh, A. & Krishnamurthy, A. Detecting activations over graphs using spanning tree wavelet bases. In *International Conference on Artificial Intelligence and Statistics*, volume **16**, 2013b.
58. Sharpnack, J., Rinaldo, A. & Singh, A. (2015) Detecting anomalous activity on networks with the graph Fourier scan statistic. *IEEE Trans. Signal Process.*, **64**, 364–379.
59. Shi, Z. (2015) *Convergence of laplacian spectra from random samples* arXiv preprint arXiv:1507.00151.
60. Singer, A. & Wu, H.-T. (2017) Spectral convergence of the connection laplacian from random samples. *Inf. Inference*, **6**, 58–123.
61. Spielman, D. A. & Teng, S.-H. (2011) Spectral sparsification of graphs. *SIAM J. Comput.*, **40**, 981–1025.
62. Spielman, D. A. & Teng, S.-H. (2013) A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM J. Comput.*, **42**, 1–26.
63. Spielman, D. A. & Teng, S.-H. (2014) Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.*, **35**, 835–885.
64. Stone, C. J. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Stat.*, 1348–1360.
65. Stone, C. J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Stat.*, 1040–1053.
66. Tsybakov, A. B. (2008) *Introduction to Nonparametric Estimation*. Springer.
67. Čencov, N. N. (1962) Estimation of an unknown distribution density from observations. *Soviet Math.*, **3**, 1559–1566.
68. Vishnoi, N. K. (2012) Laplacian solvers and their algorithmic applications. *Found. Trends Theor. Comput. Sci.*, **8**, 1–141.
69. Luxburg, U. von, Belkin, M. & Bousquet, O. (2008) Consistency of spectral clustering. *Ann. Stat.*, **36**, 555–586.
70. Wainwright, M. J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Biewpoint*. Cambridge University Press.
71. Wang, Y.-X., Sharpnack, J., Smola, A. J. & Tibshirani, R. J. (2016) Trend filtering on graphs. *J. Mach. Learn. Res.*, **17**, 3651–3691.
72. Wasserman, L. (2006) *All of Nonparametric Statistics*. Springer.
73. Yang, Y. & Dunson, D. B. (2016) Bayesian manifold regression. *Ann. Stat.*, **44**, 876–905.
74. Zhou, X. and Srebro, N. 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.
75. Zhu, X., Ghahramani, Z. & Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning*, volume **20**, 2003.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

APPENDIX

A. Notation Table

TABLE A3 *Notation.*

Symbol	Definition
$\mathscr{X}$	domain, either an open set in $\mathbb{R}^d$ or a compact manifold embedded in $\mathbb{R}^d$.
ν	Lebesgue measure
μ	volume form induced by the embedding of $\mathscr{X}$ into $\mathbb{R}^d$
P	probability measure associated with the design points
p	density of the probability measure, either with respect to ν or μ.
$L^2(\mathscr{X})$	set of square-integrable functions, either with respect to ν or μ $\int_{\mathscr{X}} f^2\,d\nu < \infty$ or $\int_{\mathscr{X}} f^2\,d\mu < \infty$
$C^k(\mathscr{X})$	functions which are k-times continuously differentiable in $\mathscr{X}$
$C_c^\infty(\mathscr{X})$	functions in $C^\infty(\mathscr{X})$ which are compactly supported in $\mathscr{X}$
$H^s(\mathscr{X})$	order-s Sobolev space (Definition 1 in the flat Euclidean setting, Definition 3 in the manifold setting.)
$H_0^s(\mathscr{X})$	order-s zero-trace Sobolev space (Definition 2)
$\|\cdot\|_2$	Euclidean distance
$d_{\mathscr{X}}(\cdot,\cdot)$	geodesic distance
$B(x,\delta)$	Ball in Euclidean distance, centred at x with radius δ
$B_{\mathscr{X}}(x,\delta)$	Ball in geodesic distance

B. Upper bounds on population-level spectral series regression

In this section we first give the proof of Proposition 1, then of Proposition 2. In both cases the structure of the analysis, which is fairly classical and straightforward, can be usefully compared with our analysis of PCR-LE (see Section 3.3).

Proof of Proposition 1. We decompose risk into squared bias and variance,

$$\mathbb{E}\|\widetilde{f} - f_0\|_P^2 = \mathbb{E}\|\mathbb{E}[\widetilde{f}] - f_0\|_P^2 + \mathbb{E}\|\widetilde{f} - \mathbb{E}[\widetilde{f}]\|_P^2. \tag{B.1}$$

Since the eigenfunctions $\{\psi_k\}$ form an orthonormal basis of $L^2(\mathscr{X})$ (with respect to the inner-product $\langle\cdot,\cdot\rangle_P$) and $f_0 \in \mathscr{H}^s(\mathscr{X}) \subseteq L^2(\mathscr{X})$, we can write the squared bias in terms of squared Fourier coefficients of f_0, leading to the following upper bound,

$$\|f_0 - \mathbb{E}[\widetilde{f}]\|_P^2 = \sum_{k=K+1}^{\infty} \langle f_0, \psi_k\rangle_P^2 \leq \frac{1}{\rho_{K+1}^s} \sum_{k=K+1}^{\infty} \rho_{k+1}^s \langle f_0, \psi_k\rangle_P^2 \leq \frac{\|f_0\|_{\mathscr{H}^s(\mathscr{X})}^2}{\rho_{K+1}^s}.$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

On the other hand, the variance term can be written as the sum of the variance of each empirical Fourier coefficient, and subsequently by the law of total variance we derive that

$$\begin{aligned}\mathbb{E}\|\widetilde{f}-\mathbb{E}[\widetilde{f}]\|_P^2 = \sum_{k=1}^{K}\mathrm{Var}\left[\langle \mathbf{Y},\psi_k\rangle_n\right] &= \sum_{k=1}^{K}\mathrm{Var}\left[\mathbb{E}[\langle Y,\psi_k\rangle_n|X_1,\ldots,X_n]\right]+\mathbb{E}\left[\mathrm{Var}[\langle Y,\psi_k\rangle_n|X_1,\ldots,X_n]\right]\\ &= \sum_{k=1}^{K}\mathrm{Var}\left[\langle f_0,\psi_k\rangle_n\right]+\frac{1}{n}\mathbb{E}\left[\|\psi_k\|_n^2\right]\\ &\le \frac{1}{n}\sum_{k=1}^{K}\mathbb{E}\left[\left(f_0(X)\psi_k(X)\right)^2\right]+\frac{K}{n}. \qquad \text{(B.2)}\end{aligned}$$

Consequently,

$$\mathbb{E}\|\widetilde{f}-f_0\|_P^2 \le \frac{\|f_0\|_{\mathscr{H}^s(\mathscr{X})}^2}{\left[\rho_{K+1}\right]^s}+\frac{K}{n}+\frac{1}{n}\mathbb{E}\left[(f_0(X))^2\cdot\sum_{k=1}^{K}(\psi_k(X))^2\right]. \qquad \text{(B.3)}$$

The claim of the proposition then follows from variants of two classical results in spectral geometry. The first is a Weyl's Law asymptotic scaling of the eigenvalues of Δ_P due to [22]; formally, there exist constants c and C (which will depend on P and d) such that

$$ck^{2/d}\le\rho_k\le Ck^{2/d}\quad\text{for all } k\in\mathbb{N}, k\ge 2. \qquad \text{(B.4)}$$

The second is a local analogue to Weyl's Law, which says that there exists a constant C (again depending on P and d) such that

$$\sup_{x\in\mathscr{X}}\left\{\sum_{k=1}^{K}\left(\psi_k(x)\right)^2\right\}\le CK\quad\text{for all } K\in\mathbb{N}. \qquad \text{(B.5)}$$

Equation (B.5) is a direct implication of (B.4) along with Theorem 17.5.3 of [36]. Plugging the upper bounds (B.4) and (B.5) back into (B.3), and recalling that $\mathbb{E}[(f_0(X))^2]=\|f_0\|_P^2\le 1$, we conclude that

$$\mathbb{E}\|\widetilde{f}-f_0\|_P^2\le C\left(\frac{\|f_0\|_{\mathscr{H}^s(\mathscr{X})}^2}{(K+1)^{2s/d}}+\frac{K}{n}\right). \qquad \text{(B.6)}$$

If $n^{-1/2}\ge M$, then taking $K=1$ implies $\mathbb{E}\|\widetilde{f}-f_0\|_P^2\le C(M^2+1/n)$. Otherwise, setting $K=\lfloor M^2 n\rfloor^{d/(2s+d)}$ balances squared bias and variance, and yields the claim. □

Proof of Proposition 2. We briefly lay out the main ideas needed to prove Proposition 2, following the lead of [39] who prove a similar result in the special case where $M=1$ and P is the uniform distribution over $\mathscr{X}=[0,1]^d$, and referring to that work for more details.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

We begin by computing the first two moments of the test statistic $\widetilde{T}$. The expectation is

$$\mathbb{E}[\widetilde{T}] = \frac{(n-1)}{n}\sum_{k=1}^{K}\langle f_0, \psi_k\rangle_P^2 + \frac{K}{n} + \mathbb{E}\Big[(f_0(X))^2 \cdot \sum_{k=1}^{K}(\psi_k(X))^2\Big].$$

To compute the variance, we decompose $\widetilde{T} = \widetilde{T}_{1,1} + \widetilde{T}_{1,2} + \widetilde{T}_{1,3} + \widetilde{T}_2$ into the sum of 3 U-statistics and the remaining diagonal terms, defined in terms of the equivalent kernel $\kappa(x,x') = \sum_{k=1}^{K}\psi_k(x)\psi_k(x')$ as

$$T_{1,1} := \frac{1}{n^2}\sum_{1\le i\ne j\le n} w_i w_j \kappa(X_i,X_j), \qquad T_{1,2} := \frac{1}{n^2}\sum_{1\le i\ne j\le n}\Big(w_i f_0(X_j) + w_j f_0(X_i)\Big)\kappa(X_i,X_j)$$

$$T_{1,3} := \frac{1}{n^2}\sum_{1\le i\ne j\le n} f_0(X_i) f_0(X_j)\kappa(X_i,X_j), \qquad T_2 := \frac{1}{n^2}\sum_{i=1}^{n} Y_i^2 \kappa(X_i,X_i).$$

The variances of each statistic can be found by routine computation (see [39]), and in particular satisfy the upper bounds

$$\mathrm{Var}(T_{1,1}) \le \frac{2K}{n^2}, \qquad \mathrm{Var}(T_{1,2}) \overset{\text{(i)}}{\le} \frac{C}{n}\|f_0\|_P^2$$

$$\mathrm{Var}(T_{1,3}) \overset{\text{(ii)}}{\le} C\bigg(\frac{K}{n}\|f_0\|_P^4 + \frac{K}{n^2}\|f_0\|_{L^4(\mathscr{X})}^4\bigg), \qquad \mathrm{Var}(T_2) \overset{\text{(iii)}}{\le} \frac{CK^2}{n^3}\bigg(1 + \|f_0\|_{L^4(\mathscr{X})}^4\bigg)$$

where (i)–(iii) hold due to local Weyl's law, i.e. (B.5). Upper bounds on Type I and Type II error,

$$\mathbb{E}_0[\widetilde{\varphi}] \le \bigg(1 + CK/n^2\bigg)a, \quad \mathbb{E}_{f_0}[1-\widetilde{\varphi}] \le \frac{C(K/n^2 + \|f_0\|_P^2 + K/n\|f_0\|_P^4 + K/n^2\|f_0\|_{L^4(\mathscr{X})}^4)}{(\sum_{k=1}^{K}\langle f_0,\psi_k\rangle_P^2 - \sqrt{2K/an})^2},$$

follow from Chebyshev's inequality. From (B.4) (Weyl's Law), we have that

$$\sum_{k=1}^{K}\langle f_0,\psi_k\rangle_P^2 \ge \|f_0\|_P^2 - \frac{\|f_0\|_{\mathscr{H}^s(\mathscr{X})}^2}{\rho_{K+1}^s} \ge \|f_0\|_P^2 - C\frac{\|f_0\|_{\mathscr{H}^s(\mathscr{X})}^2}{(K+1)^{2s/d}},$$

and it can be verified that so long as

$$\|f_0\|_P^2 \ge C\bigg(\frac{\|f_0\|_{\mathscr{H}^s(\mathscr{X})}^2}{(K+1)^{2s/d}} + \frac{\sqrt{K}}{n}\bigg(\sqrt{\frac{1}{a}} + \sqrt{\frac{1}{b}}\bigg)\bigg) \tag{B.7}$$

for a sufficiently large constant C, then $\mathbb{E}_{f_0}[1-\widetilde{\varphi}] \le b$. The two summands in (B.7) are bias and standard deviation terms, respectively. When $M^2 \le n^{-1}$, setting $K = 1$ gives the desired result. Otherwise, choosing $K = \lfloor M^2 n\rfloor^{2d/(4s+d)}$ balances these two terms, and leads to (2.10). □

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

C. Graph-dependent error bounds

In this section, we adopt the fixed design perspective; or equivalently, condition on $X_i = x_i$ for $i = 1,\dots,n$. We take $G = \big(\{x_1,\dots,x_n\}, W\big)$ to be a fixed graph on $\{x_1,\dots,x_n\}$ with Laplacian matrix $L = \sum_{k=1}^n \lambda_k v_k v_k^\top$. 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 $f_0 = (f_0(x_1),\dots,f_0(x_n)) \in \mathbb{R}^n$.

C.1 *Upper bound on Estimation Error of PCR-LE*

Lemma C.1. Suppose we observe $(Y_1,x_1),\dots,(Y_n,x_n)$ according to (C.1). Then for any integer $s \geq 1$, and any integer $1 \leq K \leq n$, the PCR-LE estimator $\widehat{f}$ of (2.2) satisfies

$$\|\widehat{f} - f_0\|_n^2 \leq \frac{\langle L^s f_0, f_0\rangle_n}{\lambda_{K+1}^s} + \frac{5K}{n}, \tag{C.2}$$

with probability at least $1 - \exp(-K)$ if $1 \leq K \leq n$.

Proof of Lemma C.1. By the triangle inequality,

$$\|\widehat{f} - f_0\|_n^2 \leq 2\left(\|\mathbb{E}\widehat{f} - f_0\|_n^2 + \|\widehat{f} - \mathbb{E}\widehat{f}\|_n^2\right). \tag{C.3}$$

The first term in (C.3) (approximation error) is non-random, since the design is fixed. The expectation $\mathbb{E}\widehat{f} = \sum_{k=1}^K \langle v_k, f_0\rangle v_k$, so that

$$\|\mathbb{E}\widehat{f} - f_0\|_n^2 = \left\| \sum_{k=K+1}^n \langle v_k, f_0\rangle v_k \right\|_n^2 = \frac{1}{n}\sum_{k=K+1}^n \langle v_k, f_0\rangle^2.$$

In the above, the last equality relies on the fact that v_k are orthonormal with respect to the usual Euclidean inner product $\langle\cdot,\cdot\rangle$. Using the fact that the eigenvalues are in increasing order, we obtain

$$\frac{1}{n}\sum_{k=K+1}^n \langle v_k, f_0\rangle^2 \leq \frac{1}{n\lambda_{K+1}^s}\sum_{k=K+1}^n \lambda_k^s \langle v_k, f_0\rangle^2 \leq \frac{\langle L^s f_0, f_0\rangle_n}{\lambda_{K+1}^s}.$$

Observe that $\langle v_k, \varepsilon\rangle \overset{d}{=} Z_k$, where $(Z_1,\dots,Z_n) \sim N(0, I_{n\times n})$. Again using the orthonormality of the eigenvectors v_k, we have

$$\|\widehat{f} - \mathbb{E}\widehat{f}\|_n^2 = \frac{1}{n}\sum_{k=1}^K \langle v_k, \varepsilon\rangle^2 \overset{d}{=} \frac{1}{n}\sum_{k=1}^K Z_k^2.$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

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 [45] 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.2 *Upper bound on Testing Error of PCR-LE*

In the following Lemma, we upper bound the Type I and Type II error of the test $\varphi = \mathbf{1}\{\widehat{T} \geq t_a\}$.

LEMMA C.2. Suppose we observe $(Y_1, x_1), \ldots, (Y_n, x_n)$ according to (C.1). Fix $(a, b) \in (0, 1)$. Then $\mathbb{E}_0[\varphi] \leq a$, and if additionally $f_0 \neq 0$ 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 $s \in \mathbb{N}, s \geq 1$, then $\mathbb{E}_{f_0}[1 - \phi] \leq b$.

Proof of Lemma C.2. 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 $\widehat{T} = \frac{1}{n}\sum_{k=1}^{K}\langle \mathbf{Y}, v_k\rangle^2$. Expanding the square gives

$$\mathbb{E}[\widehat{T}] = \frac{1}{n}\sum_{k=1}^{K}\mathbb{E}\big[\langle \mathbf{Y}, v_k\rangle^2\big] = \frac{K}{n} + \sum_{k=1}^{K}\langle f_0, v_k\rangle^2.$$

Thus $\mathbb{E}[\widehat{T}] - t_a = \frac{1}{n}\sum_{k=1}^{K}\langle f_0, v_k\rangle^2 - \sqrt{2K}/n \cdot \sqrt{1/a}$. Furthermore, it is a consequence of (C.4) that

$$\frac{1}{n}\sum_{k=1}^{K}\langle f_0, v_k\rangle^2 - \frac{\sqrt{2K}}{n}\sqrt{1/a} \geq \|f_0\|_n^2 - \frac{\langle L^s f_0, f_0\rangle_n}{\lambda_{K+1}^s} - \frac{\sqrt{2K}}{n}\sqrt{1/a} \geq \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \sqrt{\frac{2}{b}} + \frac{32}{bn}\right]. \tag{C.5}$$

Variance. Recall from the proof of Lemma C.1 that $\langle \varepsilon, v_k\rangle \overset{d}{=} Z_k$ for $(Z_1, \ldots, Z_n) \sim N(0, I_{n\times n})$. Expanding the square, and recalling that $\mathrm{Cov}[Z, Z^2] = 0$ for Gaussian random variables, we have that

$$\mathrm{Var}\left[\langle \mathbf{Y}, v_k\rangle^2\right] = \mathrm{Var}\left[2\langle f_0, v_k\rangle Z_k + 2Z_k^2\right] = 4\langle f_0, v_k\rangle^2 + 2.$$

Moreover, since $\mathrm{Cov}[Z_k^2, Z_\ell^2] = 0$ for each $k = 1, \ldots, K$, we see that

$$\mathrm{Var}\left[\widehat{T}\right] = \frac{1}{n^2}\sum_{k=1}^{K}\mathrm{Var}\left[\langle \mathbf{Y}, v_k\rangle^2\right] = \frac{2K}{n^2} + \sum_{k=1}^{K}\frac{4\langle f_0, v_k\rangle^2}{n^2}.$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

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.5) implies $\mathbb{E}_{f_0}[\widehat{T}] \geq t_a$, and apply Chebyshev's inequality to deduce

$$\mathbb{P}_{f_0}\left(\widehat{T} < t_a\right) \leq \mathbb{P}_{f_0}\left(|\widehat{T} - \mathbb{E}_{f_0}[\widehat{T}]|^2 > |\mathbb{E}_{f_0}[\widehat{T}] - t_a|^2\right) \leq \frac{\mathrm{Var}\left[\widehat{T}\right]}{\left[\mathbb{E}_{f_0}[\widehat{T}] - t_a\right]^2} = \frac{2K/n^2 + 4/n^2 \sum_{k=1}^{K} \langle f_0, v_k\rangle^2}{\left[\mathbb{E}_{f_0}[\widehat{T}] - t_a\right]^2}.$$

Thus we have upper bounded the Type II error by the sum of two terms, each of which are no more than $b/2$, as we now show. For the first term, after noting that (C.5) implies $\mathbb{E}_{f_0}[\widehat{T}] - t_a \geq \sqrt{2K}/n \cdot \sqrt{2/b}$, the upper bound follows:

$$\frac{2K/n^2}{\left[\mathbb{E}_{f_0}[\widehat{T}] - t_a\right]^2} \leq \frac{b}{2}.$$

On the other hand, for the second term we use (C.5) in two ways: first to conclude that $\mathbb{E}_{f_0}[\widehat{T}] - t_a \geq \frac{1}{2n} \cdot \sum_{k=1}^{K} \langle f_0, v_k\rangle^2$, and second to obtain

$$\frac{4/n^2 \sum_{k=1}^{K} \langle f_0, v_k\rangle^2}{\left[\mathbb{E}_{f_0}[\widehat{T}] - t_a\right]^2} \leq \frac{4/n^2 \sum_{k=1}^{K} \langle f_0, v_k\rangle^2}{\left(\frac{1}{n} \sum_{k=1}^{K} \langle f_0, v_k\rangle^2/2\right)^2} \leq \frac{16}{\sum_{k=1}^{K} \langle f_0, v_k\rangle^2} \leq \frac{b}{2}.$$

D. Graph Sobolev semi-norm, flat Euclidean domain

In this section we prove Proposition 4. The proposition will follow from several intermediate results.

1. In Section D.1, we show that if $f \in H_0^s(\mathcal{X}; M)$, then

   $$\langle L_{n,\varepsilon}^s f, f\rangle_n \leq \frac{1}{\delta} \langle L_{P,\varepsilon}^s f, f\rangle_P + \frac{C\varepsilon^2}{n\varepsilon^{2s+d}} M^2, \tag{D.1}$$

   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 with the non-local Sobolev semi-norm as long as $\varepsilon \ll n^{-1/(2(s-1)+d)}$.

2. In Section D.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 D.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 $s > 2k$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

4. In Section D.4, we use the results of the preceding two sections to show that if $f \in H_0^s(\mathscr{X};M)$ and $p \in C^{s-1}(\mathscr{X})$, there exists a constant C which does not depend on f such that

$$\langle L_{P,\varepsilon}^s f, f\rangle_P \leq CM^2. \tag{D.2}$$

Finally, in Section D.5 we provide some assorted estimates used in Sections D.1.

Proof of Proposition 4. Proposition 4 follows immediately from (D.1) and (D.2). □

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 $H^\ell(V)$.

D.1 *Decomposition of graph Sobolev semi-norm*

In Lemma D.3, 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 D.3. For any $f \in L^2(\mathscr{X})$, the graph Sobolev semi-norm satisfies

$$\langle L_{n,\varepsilon}^s f, f\rangle_n = U_{n,\varepsilon}^{(s)}(f) + B_{n,\varepsilon}^{(s)}(f), \tag{D.3}$$

such that $\mathbb{E}[U_{n,\varepsilon}^{(s)}(f)] = \frac{\binom{n}{s+1}}{n^{s+1}} \cdot \langle L_{P,\varepsilon}^s f, f\rangle_P$. If additionally $f \in H^1(\mathscr{X};M)$ and $\varepsilon \geq n^{-1/d}$, then the bias term $B_{n,\varepsilon}^{(s)}(f)$ satisfies

$$\mathbb{E}\left[|B_{n,\varepsilon}^{(s)}(f)|\right] \leq \frac{C\varepsilon^2}{\delta n\varepsilon^{2+d}} M^2. \tag{D.4}$$

Notice that $\|f\|_{H^1(\mathscr{X})}^2 \leq \|f\|_{H^s(\mathscr{X})}^2$ and $\frac{\binom{n}{s+1}}{n^{s+1}} \leq 1$. Then (D.1) follows immediately from Lemma D.3, by Markov's inequality.

Proof of Lemma D.3. We begin by introducing some notation. We will use bold notation $\mathbf{j} = (j_1, \ldots, j_s)$ 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,

$$D_i f(x) := \big(f(x) - f(X_i)\big)\, \eta\left(\frac{\|X_i - x\|}{\varepsilon}\right),$$

and we let $D_{\mathbf{j}} f(x) := \left(D_{j_1} \circ \cdots \circ D_{j_s} f\right)(x)$.

In this notation,

$$L_{n,\varepsilon} f(x) = \frac{1}{n\varepsilon^{d+2}} \sum_{i=1}^n D_i f(x),$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

and it is easy $\langle L_{n,\varepsilon}^s f, f\rangle_n$ as the sum of a U-statistic and a bias term,

$$\begin{aligned}\langle L_{n,\varepsilon}^s f, f\rangle_n &= \frac{1}{n}\sum_{i=1}^{n} L_{n,\varepsilon}^s f(X_i)\cdot f(X_i)\\ &= \underbrace{\frac{1}{n^{s+1}\varepsilon^{s(d+2)}}\sum_{i\mathbf{j}\in(n)^{s+1}} D_{\mathbf{j}}f(X_i)\cdot f(X_i)}_{=:U_{n,\varepsilon}^{(s)}(f)} + \underbrace{\frac{1}{n^{s+1}\varepsilon^{s(d+2)}}\sum_{\substack{i\mathbf{j}\in\\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}}f(X_i)\cdot f(X_i)}_{=:B_{n,\varepsilon}^{(s)}(f)}\end{aligned}$$

When the indices of $i\mathbf{j}$ are all distinct, it follows straightforwardly from the law of iterated expectation that

$$\mathbb{E}[D_{\mathbf{j}}f(X_i)\cdot f(X_i)] = \varepsilon^{s(d+2)}\mathbb{E}[L_{P,\varepsilon}^s f(X_i)\cdot f(X_i)] = \varepsilon^{s(d+2)}\langle L_{P,\varepsilon}^s f, f\rangle_P,$$

which in turn implies $\mathbb{E}[U_{n,\varepsilon}^{(s)}(f)] = \frac{\binom{n}{s+1}}{n^{s+1}}\cdot\langle L_{P,\varepsilon}^s f, f\rangle_P$.

It remains to show (D.4). Notice that for any $i,j\in[n]$ it is the case that $D_j f(X_i) = -D_i f(X_j)$. Thus, by adding and subtracting $f(X_{\mathbf{j}_s})$, we obtain by symmetry that

$$\sum_{\substack{i\mathbf{j}\in\\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}}f(X_i)\cdot f(X_i) = \frac{1}{2}\cdot\sum_{\substack{i\mathbf{j}\in\\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}}f(X_i)\cdot\Big(f(X_i) - f(X_{\mathbf{j}_s})\Big),$$

and consequently

$$\mathbb{E}\left[\sum_{\substack{i\mathbf{j}\in\\ [n]^{s+1}\setminus(n)^{s+1}}} D_{\mathbf{j}}f(X_i)\cdot f(X_i)\right] \le \frac{1}{2}\cdot\sum_{\substack{i\mathbf{j}\in\\ [n]^{s+1}\setminus(n)^{s+1}}} \mathbb{E}\Big[\Big|D_{\mathbf{j}}f(X_i)\Big|\cdot\Big|f(X_i) - f(X_{\mathbf{j}_s})\Big|\Big].$$

It follows from Lemma D.8—given later in Section D.5—$i\mathbf{j}\in[n]^{s+1}$ which contains a total of $k+1$ distinct indices,

$$\mathbb{E}\Big[\Big|D_{\mathbf{j}}f(X_i)\Big|\cdot\Big|f(X_i) - f(X_{\mathbf{j}_s})\Big|\Big] \le C_1\varepsilon^{2+kd}M^2.$$

This shows us that the expectation of $|B_{n,\varepsilon}^s(f)|$ can bounded from above by the sum over several different terms, grouped according to the number of distinct indices $|i\mathbf{j}|$, as follows:

$$\begin{aligned}\mathbb{E}\left[|B_{n,\varepsilon}^s(f)|\right] &\le C_1\frac{\varepsilon^2}{n\varepsilon^{2s}}M^2\sum_{\substack{i\mathbf{j}\in\\ [n]^{s+1}\setminus(n)^{s+1}}}\frac{1}{(n\varepsilon^d)^s}\varepsilon^{(|i\mathbf{j}|-1)d}\\ &\le C_1\frac{\varepsilon^2}{n\varepsilon^{2s}}M^2\sum_{k=1}^{s-1}\frac{(n\varepsilon^d)^k}{(n\varepsilon^d)^s}n.\end{aligned}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Finally, we note that by assumption $n\varepsilon^d \geq 1$, so that in the above sum the factor of $(n\varepsilon^d)^k$ is largest when $k = s - 1$. We conclude that

$$\mathbb{E}\left[|B^s_{n,\varepsilon}(f)|\right] \leq C_1(s-1)\frac{\varepsilon^2}{n\varepsilon^{2s+d}}M^2,$$

which is the desired result.

D.2 *Approximation error of non-local Laplacian*

In this section, we establish the convergence $L^k_{P,\varepsilon}f \to \sigma^k_\eta \Delta^k_P f$ as $\varepsilon \to 0$. More precisely, we give an upper bound on the squared difference between $L^k_{P,\varepsilon}f$ and $\sigma^k_\eta \Delta^k_P f$ as a function of ε. The bound holds for all $x \in \mathscr{X}_{k\varepsilon}$, and $f \in H^s(\mathscr{X})$, as long as $s \geq 2k + 1$.

Lemma D.4. Let $s \in \mathbb{N}, s \geq 3$. In the flat Euclidean setting, suppose additionally that $f \in H^s(\mathscr{X}; M)$, and $p \in C^{s-1}(\mathscr{X})$. Let $L_{P,\varepsilon}$ be defined 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

$$\|L^k_{P,\varepsilon}f - \sigma^k_\eta \Delta^k_P f\|_{L^2(\mathscr{X}_{k\varepsilon})} \leq C_1 M\varepsilon \tag{D.5}$$

- If s is even and $k = (s-2)/2$, then

$$\|L^k_{P,\varepsilon}f - \sigma^k_\eta \Delta^k_P f\|_{L^2(\mathscr{X}_{k\varepsilon})} \leq C_2 M\varepsilon^2. \tag{D.6}$$

We remark that when $k = 1$ and $f \in C^3(\mathscr{X})$ or $C^4(\mathscr{X})$, statements of this kind are well known, and indeed stronger results—with $L^\infty(\mathscr{X})$ norm replacing $L^2(\mathscr{X})$ 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 in result is similar in spirit.

Proof of Lemma D.4. Throughout this proof, we shall assume that f and p are smooth functions, meaning they belong to $C^\infty(\mathscr{X})$. This is without loss of generality, since $C^\infty(\mathscr{X})$ is dense in both $H^s(\mathscr{X})$ and $C^{s-1}(\mathscr{X})$, and since both sides of the inequalities (D.5) and (D.6) are continuous with respect to $\|\cdot\|_{H^s(\mathscr{X})}$ and $\|\cdot\|_{C^{s-1}(\mathscr{X})}$ norms.

We will actually prove a more general set of statements than contained in Lemma D.4, 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 $x \in \mathscr{X}_{k\varepsilon}$,

$$L^k_{P,\varepsilon}f(x) = g_s(x)\varepsilon^{s-2k} \tag{D.7}$$

  for a function g_s that satisfies

$$\|g_s\|_{L^2(\mathscr{X}_{k\varepsilon})} \leq C\|p\|^k_{C^q(\mathscr{X})}M, \tag{D.8}$$

  where $q = 1$ if $s = 0$ or $s = 1$, and otherwise $q = s - 1$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

- If $k < s/2$, then for every $x \in \mathscr{X}_{k\varepsilon}$,

$$L^k_{P,\varepsilon} f(x) = \sigma^k_\eta \cdot \Delta^k_P f(x) + \sum_{j=1}^{\lfloor (s-1)/2 \rfloor - k} g_{2(j+k)}(x)\varepsilon^{2j} + g_s(x)\varepsilon^{s-2k}. \tag{D.9}$$

for functions g_j that satisfy

$$\|g_j\|_{H^{s-j}(\mathscr{X}_{k\varepsilon})} \leq C\|p\|^k_{C^{s-1}(\mathscr{X})} M. \tag{D.10}$$

In the statement above, recall that $H^0(\mathscr{X}_{k\varepsilon}) = L^2(\mathscr{X}_{k\varepsilon})$. Additionally, note that we may speak of the pointwise behaviour of derivatives of f because we have assumed that f is a smooth function. Observe that (D.5) follows upon taking $k = \lfloor (s-1)/2 \rfloor$ in (D.9), whence we have

$$\left(L^k_{P,\varepsilon} f(x) - \sigma^k_\eta \Delta^k_P f(x) \right)^2 = \varepsilon^2 \left(g_s(x) \right)^2$$

for some $g_s \in L^2(\mathscr{X}_{k\varepsilon}, C \cdot M \cdot \|p\|_{C^{s-1}(\mathscr{X})})$, and integrating over $\mathscr{X}_{k\varepsilon}$ gives the desired result. (D.6) follows from (D.9) in an identical fashion.

It thus remains to establish (D.9), and (D.7) which is an important part of proving (D.9). 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 (D.10).

Proof of (D.7) and (D.9), base case. We begin with the base case, where $k = 1$. Again, we point out that although the 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 (D.7) is implied by Lemma D.6. When $s \geq 1$, we proceed using Taylor expansion. For any $x \in \mathscr{X}_\varepsilon$, we have that $B(x, \varepsilon) \subseteq \mathscr{X}$. 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 I.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,

$$\begin{aligned} L_{P,\varepsilon} f(x) &= \frac{1}{\varepsilon^{d+2}} \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{q-1} \frac{1}{j_1! j_2!} \int_{\mathscr{X}} \left(d^{j_1}_x f \right)(x' - x) \left(d^{j_2}_x p \right)(x' - x) \eta\left(\frac{\|x' - x\|}{\varepsilon} \right) dx' \\ &+ \frac{1}{\varepsilon^{d+2}} \sum_{j=1}^{s-1} \frac{1}{j!} \int_{\mathscr{X}} \left(d^j_x f \right)(x' - x) r^q_{x'}(x; p) \eta\left(\frac{\|x' - x\|}{\varepsilon} \right) dx' \\ &+ \frac{1}{\varepsilon^{d+2}} \int_{\mathscr{X}} r^j_{x'}(x; f) \eta\left(\frac{\|x' - x\|}{\varepsilon} \right) dP(x'). \end{aligned}$$

Here we have adopted the convention that $\sum_{j=1}^{0} = 0$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Changing variables to $z = (x' - x)/\varepsilon$, we can rewrite the above expression as

$$\begin{aligned} L_{P,\varepsilon}f(x) &= \frac{1}{\varepsilon^2}\sum_{j_1=1}^{s-1}\sum_{j_2=0}^{q-1}\frac{\varepsilon^{j_1+j_2}}{j_1!j_2!}\int d_x^{j_1}f(z)d_x^{j_2}p(z)\eta\left(\|z\|\right)\,dz \\ &+ \frac{1}{\varepsilon^2}\sum_{j=1}^{s-1}\frac{\varepsilon^j}{j!}\int d_x^j f(z) r^q_{zh+x}(x;p)\eta\left(\|z\|\right)\,dz \\ &+ \frac{1}{\varepsilon^2}\int r^j_{zh+x}(x;f)\eta\left(\|z\|\right)p(zh+x)\,dz \\ &:= G_1(x) + G_2(x) + G_3(x). \end{aligned}$$

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 $s \geq 2$

$$\|G_2\|_{L^2(\mathscr{X}_\varepsilon)} \leq C\varepsilon^{s-2}M\|p\|_{C^{s-1}(\mathscr{X})}; \tag{D.11}$$

this same estimate will hold for G_3 for all $s \geq 1$. Together these will imply (D.7) and (D.9).

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

$$G_1(x) = \sum_{j_1=1}^{s-1}\sum_{j_2=0}^{q-1}\frac{\varepsilon^{j_1+j_2-2}}{j_1!j_2!}\underbrace{\int_{B(0,1)} d_x^{j_1}f(z)d_x^{j_2}p(z)\eta(\|z\|)\,dz}_{:=g_{j_1,j_2}(x)} \tag{D.12}$$

The nature of $g_{j_1,j_2}(x)$ depends on the sum $j_1 + j_2$. Since $d_x^{j_1}fd_x^{j_2}$ is an order $j_1 + j_2$ (multivariate) monomial, we have (see Section I.2) that whenever $j_1 + j_2$ is odd,

$$g_{j_1,j_2}(x) = \int_{\mathscr{X}} d_x^{j_1}f(z)d_x^{j_2}p(z)\eta(\|z\|)\,dz = 0.$$

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 (D.12) are those where $j_1 + j_2 = 2$, so that either

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

$j_1 = 1$ and $j_2 = 1$, or $j_1 = 2$ and $j_2 = 0$. We have that

$$\begin{aligned}
g_{1,1}(x) + \frac{1}{2}g_{2,0}(x) &= \int_{\mathscr{X}} d_x^1 f(z) d_x^1 p(z)\eta(\|z\|)\,dz + \frac{p(x)}{2}\int_{\mathscr{X}} d_x^2 f(z)\eta(\|z\|)\,dz \\
&= \sum_{i_1=1}^{d}\sum_{i_2=1}^{d} D^{e_{i_1}}f(x)D^{e_{i_2}}p(x)\int_{\mathscr{X}} z^{e_{i_1}+e_{i_2}}\eta(\|z\|)\,dz \\
&\quad + \frac{p(x)}{2}\sum_{i_1=1}^{d}\sum_{i_2=1}^{d} D^{e_{i_2}+e_{i_2}}f(x)\int_{\mathscr{X}} z^{e_{i_1}+e_{i_2}}\eta(\|z\|)\,dz \\
&= \sum_{i=1}^{d} D^{e_i}f(x)D^{e_i}p(x)\int_{\mathscr{X}} z^2\eta(\|z\|)\,dz + \frac{p(x)}{2}\sum_{i=1}^{d} D^{2e_i}f(x)\int_{\mathscr{X}} z^2\eta(\|z\|)\,dz \\
&= \sigma_\eta \Delta_P f(x),
\end{aligned}$$

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 (D.10) 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

$$\begin{aligned}
\|D^{\alpha_1}fD^{\alpha_2}p\|_{H^{s-(j+2)}(\mathscr{X})} &\leq \|D^{\alpha_1}f\|_{H^{s-(j+2)}(\mathscr{X})}\|D^{\alpha_2}p\|_{C^{s-(j+2)}(\mathscr{X})} \\
&\leq \|D^{\alpha_1}f\|_{H^{s-j_1}(\mathscr{X})}\|D^{\alpha_2}p\|_{C^{s-(j_2+1)}(\mathscr{X})} \\
&\leq M \cdot \|p\|_{C^{s-1}(\mathscr{X})},
\end{aligned}$$

and summing over all $|\alpha_1| = j_1$ and $|\alpha_2| = j_2$ establishes that g_{j_1,j_2} satisfies (D.10).

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 $|r_{x+z\varepsilon}^{s-1}(x;p)| \leq C\varepsilon^{s-1}\|p\|_{C^{s-1}(\mathscr{X})}$ for any $z \in B(0,1)$, and that $d_x^j f(\cdot)$ is a j-homogeneous function, we have that

$$\begin{aligned}
|G_2(x)| &\leq \sum_{j=1}^{s-1}\frac{\varepsilon^{j-2}}{j!}\int_{B(0,1)}\left|\left(d_x^j f\right)(z)\right| \cdot |r_{x+z\varepsilon}^{s-1}(x;p)| \cdot \eta(\|z\|)\,dz \\
&\leq C\varepsilon^{s-2}\|p\|_{C^{s-1}(\mathscr{X})}\sum_{j=1}^{s-1}\frac{1}{j!}\int_{B(0,1)}\left|\left(d_x^j f\right)(z)\right| \cdot \eta(\|z\|)\,dz. \qquad \text{(D.13)}
\end{aligned}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Furthermore, for each $j = 1, \dots, s-1$ convolution of $d_x^j f$ with η only decreases the $L^2(\mathscr{X}_\varepsilon)$ norm, meaning

$$\begin{aligned}\int_{\mathscr{X}_\varepsilon} \left(\int_{B(0,1)} \left| \left(d_x^j f \right)(z) \right| \cdot \eta(\|z\|)\, dz \right)^2 dx &\le \int_{\mathscr{X}_\varepsilon} \left(\int_{B(0,1)} \left| \left(d_x^j f \right)(z) \right|^2 \eta(\|z\|)\, dz \right) \cdot \left(\int_{B(0,1)} \eta(\|z\|)\, dz \right) dx \\ &\le \int_{B(0,1)} \int_{\mathscr{X}_\varepsilon} \left[\left(d^j f \right)(x) \right]^2 \eta(\|z\|)\, dx\, dz \\ &\le \|d^j f\|^2_{L^2(\mathscr{X}_\varepsilon)}.\end{aligned} \tag{D.14}$$

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 (D.13), we conclude that

$$\begin{aligned}\int_{\mathscr{X}_\varepsilon} |G_2(x)|^2\, dx &\le C \left(\varepsilon^{s-2} \|p\|_{C^{s-1}(\mathscr{X})} \right)^2 \sum_{j=1}^{s-1} \int_{\mathscr{X}_\varepsilon} \left(\frac{1}{j!} \int_{B(0,1)} \left| \left(d_x^j f \right)(z) \right| \cdot |\eta(\|z\|)|\, dz \right)^2 dx \\ &\le C \left(\varepsilon^{s-2} \|p\|_{C^{s-1}(\mathscr{X})} \right)^2 \sum_{j=1}^{s-1} \|d^j u\|^2_{L^2(\mathscr{X}_\varepsilon)},\end{aligned}$$

establishing the desired estimate.

Estimate on $G_3(x)$. Applying the Cauchy–Schwarz inequality, we deduce a pointwise upper bound on $|G_3(x)|^2$,

$$\begin{aligned}|G_3(x)|^2 &\le \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \cdot \left(\int_{B(0,1)} \left| r^s_{x+\varepsilon z}(x; u) \right|^2 \eta(\|z\|)\, dz \right) \cdot \left(\int_{B(0,1)} \eta(\|z\|)\, dz \right) \\ &\le \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \int_{B(0,1)} \left| r^s_{x+\varepsilon z}(x; u) \right|^2 \eta(\|z\|)\, dz.\end{aligned}$$

Applying this pointwise over all $x \in \mathscr{X}_\varepsilon$ and integrating, we obtain

$$\begin{aligned}\int_{\mathscr{X}_\varepsilon} |G_3(x)|^2\, dx &\le \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \int_{\mathscr{X}_\varepsilon} \int_{B(0,1)} \left| r^s_{x+\varepsilon z}(x; f) \right|^2 \eta(\|z\|)\, dz\, dx \\ &= \left(\frac{p_{\max}}{\varepsilon^2} \right)^2 \int_{B(0,1)} \int_{\mathscr{X}_\varepsilon} \left| r^s_{x+\varepsilon z}(x; f) \right|^2 \eta(\|z\|)\, dx\, dz \\ &\le \left(\frac{p_{\max} \varepsilon^s}{\varepsilon^2} \right)^2 \|d^s f\|^2_{L^2(\mathscr{X}_\varepsilon)},\end{aligned}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

with the last inequality following from (I.2). Noting that $p_{\max} = \|p\|_{C^0(\mathscr{X})} \le \|p\|_{C^{s-1}(\mathscr{X})}$, we see that this is a sufficient bound on $\|G_3\|_{L^2(\mathscr{X}_\varepsilon)}$.

Proof of (D.7) and (D.9), induction step. We now assume that (D.7) and (D.9) 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

$$\|g_j^\ell\|_{H^{\ell-j}(\mathscr{X}_{(k+1)\varepsilon})} \le C\|p\|_{C^q(\mathscr{X})}^{k+1} M. \tag{D.15}$$

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 (D.15). We immediately illustrate the purpose of this notation. Suppose $g \in H^\ell(\mathscr{X}_{k\varepsilon}; C\|p\|^k_{C^q(\mathscr{X})}M)$ for some $\ell \le s$. If $\ell \le 2$, then by the inductive hypothesis, it follows that for any $x \in \mathscr{X}_{(k+1)\varepsilon}$

$$L_{P,\varepsilon} g(x) = g_\ell^\ell(x)\varepsilon^{\ell-2}. \tag{D.16}$$

On the other hand if $2 < \ell \le s$, then by the inductive hypothesis, it follows that for any $x \in \mathscr{X}_{(k+1)\varepsilon}$,

$$L_{P,\varepsilon} g(x) = \sigma_\eta \Delta_P g(x) + \sum_{j=1}^{\lfloor(\ell-1)/2\rfloor-1} g^\ell_{2j+2}(x)\varepsilon^{2j} + g_\ell^\ell(x)\varepsilon^{\ell-2}. \tag{D.17}$$

Proof of (D.7). If $s \le 2(k+1)$, then by the inductive hypothesis it follows that for all $x \in \mathscr{X}_{k\varepsilon}$, we have $L^k_{P,\varepsilon} f(x) = g_s(x)\cdot\varepsilon^{s-2k}$, for some $g_s \in L^2(\mathscr{X}_{k\varepsilon}, C\|p\|^k_{C^{s-1}(\mathscr{X})}M)$. Note that we may know more about $L_P^k f(x)$ 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 (D.16) we deduce that for any $x \in \mathscr{X}_{(k+1)\varepsilon}$,

$$L^{k+1}_{P,\varepsilon} f(x) = (L_{P,\varepsilon}\circ L^k_{P,\varepsilon} f)(x) = L_{P,\varepsilon} g_s(x)\varepsilon^{s-2k} = g_s^s(x)\varepsilon^{s-2(k+1)}, \tag{D.18}$$

establishing (D.7).

Proof of (D.9). If $s > 2(k+1)$, then by the inductive hypothesis we have that for all $x \in \mathscr{X}_{k\varepsilon}$,

$$L^k_{P,\varepsilon} f(x) = \sigma^k_\eta \Delta^k_P f(x) + \sum_{j=1}^{\lfloor(s-1)/2\rfloor-k} g_{2(j+k)}(x)\varepsilon^{2j} + g_s(x)\varepsilon^{s-2k}.$$

Thus for any $x \in \mathscr{X}_{(k+1)\varepsilon}$,

$$L^{k+1}_{P,\varepsilon} f(x) = \left(L_{P,\varepsilon}\circ L^k_{P,\varepsilon} f\right)(x) = \sigma^k_\eta L_{P,\varepsilon}\Delta^k_P f(x) + \sum_{j=1}^{\lfloor(s-1)/2\rfloor-k} L_{P,\varepsilon} g_{2(j+k)}(x)\varepsilon^{2j} + L_{P,\varepsilon} g_s(x)\varepsilon^{s-2k}$$

There are three terms on the right-hand side of this equality, and we now analyse each separately.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

1. Noting that $\Delta_P^k f \in H^{s-2k}(\mathscr{X}; C\|p\|^k_{C^{s-1}(\mathscr{X})}M)$, we use (D.17) to derive that

$$\begin{aligned} L_{P,\varepsilon}\Delta_P^k f(x) &= \sigma_\eta \Delta_P^{k+1} f(x) + \sum_{j=1}^{(s-2k-1)/2-} g^{s-2k}_{2j+2}(x)\varepsilon^{2j} + g^{s-2k}_{s-2k}(x)\varepsilon^{s-2k-2} \\ &= \sigma_\eta \Delta_P^{k+1} f(x) + \sum_{j=1}^{(s-1)/2-(k+1)} g_{2(k+1+j)}(x)\varepsilon^{2j} + g_s(x)\varepsilon^{s-2(k+1)}, \end{aligned} \tag{D.19}$$

   where in the second equality we have simply used the fact $g_j^\ell(x) = g_{(s-\ell)+j}(x)$ to rewrite the equation.

2. Suppose $j < \lfloor (s-1)/2 \rfloor - k$. Then we use (D.17) 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 $g_j^\ell(x) = g_{(s-\ell)+j}(x)$, and also written $\sigma_\eta \Delta_P f = g_2^{s-2(j+k)} = g_{2(j+k+1)}$, since the particular dependence on the Laplacian Δ_P will not matter. From here, multiplying by ε^{2j}, we conclude that

$$\begin{aligned} \varepsilon^{2j} L_{P,\varepsilon} g_{2(j+k)}(x) &= g_{2(j+k+1)}(x)\varepsilon^{2j} + \sum_{i=1}^{\lfloor (s-1)/2\rfloor-(j+k+1)} g_{2(i+j+k+1)}(x)\varepsilon^{2(i+j)} + g_s(x)\varepsilon^{s-2(k+1)} \\ &= g_{2(j+k+1)}(x)\varepsilon^{2j} + \sum_{m=1}^{\lfloor (s-1)/2\rfloor-(k+1)} g_{2(m+k+1)}(x)\varepsilon^{2m} + g_s(x)\varepsilon^{s-2(k+1)}, \end{aligned} \tag{D.20}$$

   with the second equality following upon changing variables to $m = i + j$.

3. On the other hand if $j = \lfloor (s-1)/2 \rfloor - k$, then the calculation is much simpler,

$$\varepsilon^{2j} L_{P,\varepsilon} g_{2(j+k)}(x) = g^{s-2(j+k)}_{s-2(j+k)}(x)\varepsilon^{2j}\varepsilon^{s-2(j+k)-2} = g_s(x)\varepsilon^{s-2(k+1)}. \tag{D.21}$$

4. Finally, it follows immediately from (D.17) that

$$L_{P,\varepsilon} g_s(x)\varepsilon^{s-2k} = g_s(x)\varepsilon^{s-2(k+1)}. \tag{D.22}$$

Plugging (D.19)–(D.22) back into (D.18) proves the claim.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

D.3 *Boundary behaviour of non-local Laplacian*

In Lemma D.5, we establish that if f is Sobolev smooth of order $s > 2k$ and zero-trace, then near the boundary of $\mathscr{X}$ the non-local Laplacian $L_{P,\varepsilon}^k f$ is close to 0 in the L^2-sense.

LEMMA D.5. Let $s, k \in \mathbb{N}$. In the flat Euclidean setting, suppose additionally that $f \in H_0^s(\mathscr{X}; M)$. Then there exist numbers $c, C > 0$ that do not depend on M, such that for all $\varepsilon < c$,

$$\|L_{P,\varepsilon}^k f\|_{L^2(\partial_{k\varepsilon}\mathscr{X})}^2 \leq C\varepsilon^{2(s-2k)} M^2.$$

Proof of Lemma D.5. Applying Lemma D.6, we have that

$$\|L_{P,\varepsilon}^k f\|_{L^2(\partial_{k\varepsilon}(\mathscr{X}))}^2 \leq \frac{(Cp_{\max})^2}{\varepsilon^4} \|L_{P,\varepsilon}^{k-1} f\|_{L^2(\partial_{k\varepsilon}(\mathscr{X}))}^2 \leq \cdots \leq \frac{(Cp_{\max})^2}{\varepsilon^{4k}} \|f\|_{L^2(\partial_{k\varepsilon}(\mathscr{X}))}^2$$

Thus it remains to show that for all $\varepsilon < c$,

$$\|f\|_{L^2(\partial_{k\varepsilon}(\mathscr{X}))}^2 = \int_{\partial_{k\varepsilon}(\mathscr{X})} (f(x))^2 \, dx \leq C_1 \varepsilon^{2s} \|f\|_{H^s(\mathscr{X})}^2. \tag{D.23}$$

We will build to (D.23) by a series of intermediate steps, following the same rough structure as the proof of Theorem 18.1 in Leoni [47]. 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

$$U_\phi(c_0) = \left\{y \in Q(0, c_0) : \phi(y_{-d}) \leq y_d\right\},$$

and here $Q(0, c_0)$ is the d-dimensional cube of side length c_0, centred at 0. We will show that for all $0 < \varepsilon < c_0$, and for the tubular neighbourhood $V_\phi(\varepsilon) = \{y \in Q(0, c_0) : \phi(y_{-d}) \leq y_d \leq \phi(y_{-d}) + \varepsilon\}$, 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

$$f(y) = f(y_0) + \sum_{j=1}^{s-1} \frac{1}{j!} D^{je_d} f(y_0) \left(y_d - \phi(y')\right)^j + \frac{1}{(s-1)!} \int_{\phi(y')}^{y_d} (1-t)^{s-1} D^{se_d} f(y', z) \left(y_d - z\right)^{s-1} \, dz \Longrightarrow$$

$$|f(y)| \leq C\varepsilon^{s-1} \int_{\phi(y')}^{y_d} \left|D^{se_d} f(y', z)\right| \, dz.$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Consequently, by squaring both sides and applying Cauchy–Schwarz, we have that

$$|f(y)|^2 \leq C\varepsilon^{2(s-1)}\left(\int_{\phi(y')}^{y_d} \left|D^{se_d}f(y',z)\right| \, dz\right)^2 \leq C\varepsilon^{2s-1}\int_{\phi(y')}^{y_d} \left|D^{se_d}f(y',z)\right|^2 \, dz.$$

Applying this bound for each $y \in V_\phi(\varepsilon)$, and then integrating, we obtain

$$\begin{aligned}\int_{V_\phi(\varepsilon)} |f(y)|^2 \, dy &\leq \int_{Q_{d-1}(c_0)}\int_{\phi(y')}^{\phi(y')+\varepsilon} |f(y',y_d)|^2 \, dy_d \, dy' \\ &\leq C\varepsilon^{2s-1}\int_{Q_{d-1}(c_0)}\int_{\phi(y')}^{\phi(y')+\varepsilon}\int_{\phi(y')}^{y_d} \left|D^{se_d}f(y',z)\right|^2 \, dz \, dy_d \, dy', \qquad \text{(D.24)}\end{aligned}$$

where we have written $Q_{d-1}(0, c_0)$ for the $d-1$-dimensional cube of side length c_0, centred at 0. Exchanging the order of the inner two integrals then gives

$$\begin{aligned}\int_{\phi(y')}^{\phi(y')+\varepsilon}\int_{\phi(y')}^{y_d} \left|D^{se_d}f(y',z)\right|^2 \, dz \, dy_d &= \int_{\phi(y')}^{\phi(y')+\varepsilon}\int_{z}^{\varepsilon} \left|D^{se_d}f(y',z)\right|^2 \, dy_d \, dz \\ &\leq C\varepsilon\int_{\phi(y')}^{\phi(y')+\varepsilon} \left|D^{se_d}f(y',z)\right|^2 \, dz \\ &\leq C\varepsilon\int_{\phi(y')}^{c_0} \left|D^{se_d}f(y',z)\right|^2 \, dz.\end{aligned}$$

Finally, plugging back into (D.24), we conclude that

$$\int_{V_\phi(\varepsilon)} |f(y)|^2 \, dy \leq C\varepsilon^{2s}\int_{Q_{d-1}(0,c_0)}\int_{\phi(y')}^{c_0} \left|D^{se_d}f(y',z)\right|^2 \, dz \, dy' \leq C\varepsilon^{2s}|u|^2_{H^s(U_\phi(c_0))}.$$

Step 2: Rigid motion of local patch. Now, suppose that at a point $x_0 \in \partial\mathscr{X}$, 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 $\varepsilon \cdot C_0 \leq c_0$,

$$T\left(Q_T(x_0, c_0) \cap \partial_\varepsilon\mathscr{X}\right) \subseteq V_\phi\left(C_0\varepsilon\right) \quad \text{and} \quad T\left(Q_T(x_0, c_0) \cap \mathscr{X}\right) = U_\phi(c_0).$$

Here $Q_T(x_0, c_0))$ is a (not necessarily coordinate-axis-aligned) cube of side length c_0), centred at x_0. Define $v(y) := f(T^{-1}(y))$ for $y \in U_\phi(c_0)$. If $u \in C_c^\infty(\mathscr{X})$, then $v \in C_c^\infty(U_\phi(c_0))$, and moreover $\|v\|^2_{H^s(U_\phi(c_0))} = \|f\|^2_{H^s(Q_T(x_0,c_0)\cap\mathscr{X})}$. Therefore, using the upper bound that we derived in Step 1,

$$\int_{V_\phi(C_0\cdot\varepsilon)} |v(y)|^2 \, dy \leq C\varepsilon^{2s}\|v\|^2_{H^s(U_\phi(c_0))},$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

we conclude that

$$\begin{aligned}\int_{Q_T(x_0,c_0)\cap\partial_\varepsilon\mathscr{X}} |f(x)|^2\,dx &= \int_{T(Q_T(x_0,c_0))\cap\partial_\varepsilon\mathscr{X})} |v(y)|^2\,dy\\ &\leq \int_{V_\phi(C_0\cdot\varepsilon)} |v(y)|^2\,dy\\ &\leq C\varepsilon^{2s}\|v\|^2_{H^s(U_\phi(c_0))} = C\varepsilon^{2s}\|f\|^2_{H^s(Q_T(x_0,c_0))\cap\mathscr{X})} \leq C\varepsilon^{2s}\|f\|^2_{H^s(\mathscr{X})}.\end{aligned}$$

Step 3: Lipschitz domain. Finally, we deal with the case where $\mathscr{X}$ is assumed to be an open, bounded subset of $\mathbb{R}^d$, with Lipschitz boundary. In this case, at every $x_0 \in \partial\mathscr{X}$, 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 $\varepsilon \cdot C_0(x_0) \leq c_0(x_0)$,

$$T\left(Q_T(x_0, c_0(x_0)) \cap \partial_\varepsilon\mathscr{X}\right) \subseteq V_\phi\left(C_0(x_0)\cdot\varepsilon\right) \quad \text{and} \quad T\left(Q_T(x_0, c_0(x_0)) \cap \mathscr{X}\right) = U_\phi(c_0(x_0)).$$

Therefore for every $x_0 \in \partial\mathscr{X}$, it follows from the previous step that

$$\int_{Q_{T_{x_0}}(x_0,c_0(x_0))\cap\partial_\varepsilon\mathscr{X}} |f(x)|^2\,dx \leq C(x_0)\varepsilon^{2s}\|f\|^2_{H^s(\mathscr{X})},$$

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 $\partial_\varepsilon\mathscr{X}$ for all ϵ sufficiently small. First, we note that by a compactness argument there exists a finite subset of the collection of cubes $\{Q_{T_{x_0}}(x_0, c_0(x_0)/2) : x_0 \in \partial\mathscr{X}\}$ which covers $\partial\mathscr{X}$, 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

$$\partial_\varepsilon\mathscr{X} \subseteq \bigcup_{i=1}^{N} Q_{T_{x_i}}(x_i, c_0(x_i)).$$

As a result,

$$\int_{\partial_\varepsilon\mathscr{X}} |f(x)|^2 \leq \sum_{i=1}^{N}\int_{Q_{T_{x_i}}(x_i,c_0(x_i))\cap\partial_\varepsilon(\mathscr{X})} |f(x)|^2 \leq \varepsilon^{2s}\|f\|^2_{H^s(\mathscr{X})}\sum_{i=1}^{N} C_0(x_i),$$

which proves the claim of (D.23).

D.4 *Estimate of non-local Sobolev seminorm*

Now, we use the results of the preceding two sections to prove (D.2). 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 $\langle L_{P,\varepsilon}f, g\rangle_P =$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

$\langle f, L_{P,\varepsilon} g\rangle_P$ for all $f, g \in L^2(\mathscr{X})$. We also recall the definition of the Dirichlet energy $E_{P,\varepsilon}(f; \mathscr{X})$,

$$\langle L_{P,\varepsilon} f, f\rangle_P = \frac{1}{\varepsilon^{d+2}} \int_{\mathscr{X}} \int_{\mathscr{X}} \big(f(x) - f(x')\big)^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x')\, dP(x) =: E_{P,\varepsilon}(f; \mathscr{X}). \tag{D.25}$$

Finally, we recall a result of [31]: there exist constants c_0 and C_0 which do not depend on M, such that for all $\varepsilon < c_0$ and for any $f \in H^1(\mathscr{X}; M)$,

$$E_{P,\varepsilon}(f; \mathscr{X}) \le C_0 M^2. \tag{D.26}$$

Case 1: s odd. Suppose s is odd, so that $s \ge 3$. 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,

$$\langle L^s_{P,\varepsilon} f, f\rangle_P = \langle L^{k+1}_{P,\varepsilon} f, L^k_{P,\varepsilon} f\rangle_P = E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \mathscr{X}).$$

We now separate this energy into integrals over $\mathscr{X}_{k\varepsilon}$ and $\partial_{k\varepsilon}(\mathscr{X})$,

$$\begin{aligned} E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \mathscr{X}) = \frac{1}{\varepsilon^{d+2}} \Bigg\{ &\int_{\mathscr{X}_{k\varepsilon}} \int_{\mathscr{X}_{k\varepsilon}} \Big(L^k_{P,\varepsilon} f(x) - L^k_{P,\varepsilon} f(x')\Big)^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x')\, dP(x) \\ &+ \int_{\partial_{k\varepsilon} \mathscr{X}} \int_{\partial_{k\varepsilon} \mathscr{X}} \Big(L^k_{P,\varepsilon} f(x) - L^k_{P,\varepsilon} f(x')\Big)^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x')\, dP(x) \Bigg\} \\ := E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \mathscr{X}_{k\varepsilon}) &+ E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \partial_{k\varepsilon} \mathscr{X}) \end{aligned} \tag{D.27}$$

and upper bound each energy separately. For the first term, we add and substract $\sigma^k_\eta \Delta^k_P f(x)$ and $\sigma^k_\eta \Delta^k_P f(x')$ within the integrand, then use the triangle inequality and the symmetry between x and x' to deduce that

$$\begin{aligned} E_{P,\varepsilon}(L^k_{P,\varepsilon} f; \mathscr{X}_{k\varepsilon}) \le & 3\sigma^{2k}_\eta E_{P,\varepsilon}(\Delta^k_P f; \mathscr{X}_{k\varepsilon}) + \frac{2}{\varepsilon^{d+2}} \int_{\mathscr{X}_{k\varepsilon}} \int_{\mathscr{X}_{k\varepsilon}} (L^k_{P,\varepsilon} f(x) \\ & - \sigma^k_\eta \Delta^k_P f(x))^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x')\, dP(x). \end{aligned} \tag{D.28}$$

Noticing that $\Delta^k_P f \in H^1(\mathscr{X}; \|p\|^k_{C^{s-1}(\mathscr{X})} M)$, we use (D.26) to conclude that $E_{P,\varepsilon}(\Delta^k_P f; \mathscr{X}_{k\varepsilon}) \le C_0 M^2$. On the other hand, it follows from Assumption **(K1)** and (D.5) that

$$\begin{aligned} &\frac{2}{\varepsilon^{d+2}} \int_{\mathscr{X}_{k\varepsilon}} \int_{\mathscr{X}_{k\varepsilon}} \Big(L^k_{P,\varepsilon} f(x) - \sigma^k_\eta \Delta^k_P f(x)\Big)^2 \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) dP(x')\, dP(x) \\ &\le \frac{2p_{\max}}{\varepsilon^2} \int_{\mathscr{X}_{k\varepsilon}} \Big(L^k_{P,\varepsilon} f(x) - \sigma^k_\eta \Delta^k_P f(x)\Big)^2 dP(x) \\ &\le C_1 M^2. \end{aligned}$$

Plugging these two bounds into (D.28) gives the desired upper bound on $E_{P,\varepsilon}(L^k_{P,\varepsilon}; \mathscr{X}_{k\varepsilon})$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

For the second term in (D.27), we apply Lemmas D.7 and D.5 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 $\partial_{k\varepsilon}\mathcal{X}$,

$$\|L_{P,\varepsilon}^{k+1} f\|_P^2 \leq p_{\max}\|L_{P,\varepsilon}^{k+1} f\|_{L^2(\mathcal{X})}^2 \leq p_{\max}\left\{\int_{\mathcal{X}_{k\varepsilon}} \left(L_{P,\varepsilon}^{k+1} f(x)\right)^2 \, dP(x) + \int_{\partial_{k\varepsilon}\mathcal{X}} \left(L_{P,\varepsilon}^{k+1} f(x)\right)^2 \, dP(x)\right\}, \tag{D.29}$$

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 (D.7) since $\Delta_P^k f \in H^2(\mathcal{X}; M\|p\|_{C^{s-1}(\mathcal{X})}^l)$, (ii) following from Lemma D.6, and (iii) following from (D.6).

Then Lemma D.5 shows that the second term in (D.29) satisfies

$$\int_{\partial_{k\varepsilon}\mathcal{X}} \left(L_{P,\varepsilon}^{k+1} f(x)\right)^2 \, dP(x) \leq CM^2.$$

D.5 *Assorted integrals*

Lemma D.6. In the flat Euclidean setting, suppose additionally that $f \in L^2(U; M)$ 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

$$\|L_{P,\varepsilon} f\|_{L^2(U)} \leq \frac{2p_{\max}}{\varepsilon^2}\|f\|_{L^2(U)} \tag{D.30}$$

Lemma D.7. In the flat Euclidean setting, suppose additionally that $f \in L^2(U; M)$ 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

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

which does not depend on f or M such that

$$E_{P,\varepsilon}(f;U) \leq \frac{4p_{\max}^2}{\varepsilon^2}\|f\|_{L^2(U)}^2 \tag{D.31}$$

LEMMA D.8. In the flat Euclidean setting, suppose additionally that $f \in H^1(\mathscr{X};M)$, 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 $i,j \in [n]$ and $\mathbf{j} \in [n]^s$,

$$\mathbb{E}\left[|D_{\mathbf{j}}f(X_i)| \cdot |f(X_i) - f(X_j)|\right] \leq C\varepsilon^{2+dk}M^2,$$

where $k+1$ is the number of distinct indices in $ij\mathbf{j}$.

Proof of Lemma D.6. 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:

$$\begin{aligned}|L_{P,\varepsilon}f(x)|^2 &\leq \left(\frac{p_{\max}}{\varepsilon^{2+d}}\right)^2\left(\int_U \left(|f(x')| + |f(x)|\right)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'\right)^2 \\ &\leq \left(\frac{p_{\max}}{\varepsilon^{2+d}}\right)^2\left(\int_U \left(|f(x')| + |f(x)|\right)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx' \cdot \int \eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'\right) \\ &= \frac{p_{\max}^2}{\varepsilon^{4+d}}\int_U \left(|f(x')| + |f(x)|\right)^2 \eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'.\end{aligned}$$

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

$$\begin{aligned}\|L_{P,\varepsilon}\|_{L^2(U)}^2 &\leq \frac{2p_{\max}^2}{\varepsilon^{4+d}}\int_U\int_U \left(|f(x')|^2 + |f(x)|^2\right)\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'\,dx \\ &\leq \frac{2p_{\max}^2}{\varepsilon^{4+d}}\int_U\int_U \left(|f(x')|^2 + |f(x)|^2\right)\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'\,dx.\end{aligned} \tag{D.32}$$

Finally, using Fubini's Theorem we determine that

$$\begin{aligned}\int_U\int_U \left(|f(x')|^2 + |f(x)|^2\right)\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'\,dx &= 2\int_U\int_U |f(x)|^2\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx \\ &\leq 2\varepsilon^d\int_U |f(x)|^2\,dx = 2\varepsilon^d\|f\|_{L^2(U)}^2,\end{aligned} \tag{D.33}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

and by combining (D.32) and (D.33) we conclude that

$$\|L_{P,\varepsilon}\|_{L^2(U)}^2 \le \frac{4p_{\max}^2}{\varepsilon^4}\|f\|_{L^2(U)}^2.$$

Proof of Lemma D.7. We have

$$\begin{aligned}E_{P,\varepsilon}(f) &= \frac{1}{\varepsilon^{2+d}}\int_U\int_U \big(f(x)-f(x')\big)^2\,\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dP(x')\,dP(x)\\ &\le \frac{2p_{\max}^2}{\varepsilon^{2+d}}\int_U\int_U \left(|f(x)|^2+|f(x')|^2\right)\eta\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'\,dx,\end{aligned}$$

and the claim follows from (D.33).

Proof of Lemma D.8. Let $G_{n,\varepsilon}[X_{i\mathbf{j}}]$ be the subgraph induced by vertices $X_i, X_{\mathbf{j}_1},\ldots,X_{\mathbf{j}_s}$. An inductive argument shows that

$$|D_{\mathbf{j}}f(X_i)| \le C|f(X_{\mathbf{j}_s})-f(X_i)|\cdot\mathbf{1}\left\{G_{n,\varepsilon}[X_{i\mathbf{j}}]\ \text{ is connected.}\right\},$$

from which it follows that

$$|D_{\mathbf{j}}f(X_i)|\cdot|f(X_i)-f(X_j)| \le C|f(X_{\mathbf{j}_s})-f(X_i)|\cdot|f(X_j)-f(X_i)|\cdot\mathbf{1}\left\{G_{n,\varepsilon}[X_{ij\mathbf{j}}]\ \text{ is connected.}\right\}$$

Taking expectation and applying Cauchy–Schwarz (if $j\ne\mathbf{j}_s$) gives

$$\mathbb{E}\Big[|D_{\mathbf{j}}f(X_i)|\cdot|f(X_i)-f(X_j)|\Big] \le C\cdot\mathbb{E}\Big[|f(X_{\mathbf{j}_s})-f(X_i)|^2\cdot\mathbf{1}\left\{G_{n,\varepsilon}[X_{ij\mathbf{j}}]\ \text{ is connected.}\right\}\Big]$$

(If $j=\mathbf{j}_s$ the inequality is of course immediate.) Marginalizing out the contribution of all indices in $\mathbf{j}$ not equal to i or j gives

$$\mathbb{E}\left[|f(X_j)-f(X_i)|^2\cdot\mathbf{1}\left\{G_{n,\varepsilon}[X_{ij\mathbf{j}}]\ \text{ is connected.}\right\}\right] \tag{D.34}$$

$$\le \left((s+1)p_{\max}\nu_d\varepsilon^d\right)^{|\mathbf{j}\setminus\{j\cup i\}|}\cdot\mathbb{E}\left[|f(X_j)-f(X_i)|^2\mathbf{1}\{\|X_i-X_j\|\le\varepsilon\}\right]$$

$$\le \left((s+1)p_{\max}\nu_d\varepsilon^d\right)^{|\mathbf{j}\setminus\{j\cup i\}|}\cdot p_{\max}^2\nu_d\varepsilon^{2+d}M^2 \tag{D.35}$$

with the second inequality following from the proof of Lemma 1 in [31]. Finally, we notice that $|\mathbf{j}\setminus\{i\cup j\}|=k-1$, so that (D.35) gives the desired result.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

E. Graph Sobolev semi-norm, manifold domain

In this section we prove Proposition 7. Note that when $s = 1$, the upper bound (4.4) follows immediately from Lemma E.10 and Markov's inequality.

On the other hand when $s = 2$ or $s = 3$, we prove Proposition 7 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 analyse 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 using Euclidean rather than geodesic distance. We upper bound this error by comparing the non-local operator

$$L_{P,\varepsilon}f(x) := \frac{1}{\varepsilon^{2+m}} \int_{\mathcal{X}} \big(f(x') - f(x)\big)\, \eta\left(\frac{\|x' - x\|}{\varepsilon}\right) p(x')\, d\mu(x').$$

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

$$\widetilde{L}_{P,\varepsilon}f(x) := \frac{1}{\varepsilon^{2+m}} \int_{\mathcal{X}} \big(f(x') - f(x)\big)\, \eta\left(\frac{d_{\mathcal{X}}(x', x)}{\varepsilon}\right) p(x')\, d\mu(x').$$

We show the following results, each of which hold under the same assumptions as Proposition 7.

- In Section E.1 we show that the graph Sobolev seminorm $\langle L_{n,\varepsilon}^{s} f, f\rangle_n$ is upper bounded by the sum of a non-local seminorm and a pure bias term: specifically, with probability at least $1 - 2\delta$,

$$\langle L_{n,\varepsilon}^{s} f, f\rangle_n \leq \frac{\langle L_{P,\varepsilon}^{s} f, f\rangle_P}{\delta} + C_1 \frac{\varepsilon^2}{n\varepsilon^{2s+m}} M^2. \tag{E.1}$$

  This upper bound is essentially the same as (D.1), 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 E.2, we show that the error incurred using the 'wrong' metric is negligible. Precisely, we find that

$$\|L_{P,\varepsilon}f - \widetilde{L}_{P,\varepsilon}f\|_{L^2(\mathcal{X})}^2 \leq C_2 \varepsilon^2 |f|_{H^1(\mathcal{X})}^2. \tag{E.2}$$

- In Section E.3, we analyse the approximation error of $\widetilde{L}_{P,\varepsilon}$. We show that when $f \in H^2(\mathcal{X})$ and $p \in C^1(\mathcal{X})$,

$$\|\widetilde{L}_{P,\varepsilon}f\|_{L^2(\mathcal{X})}^2 \leq C_3 \|f\|_{H^2(\mathcal{X})}^2. \tag{E.3}$$

  On the other hand, if $f \in H^3(\mathcal{X})$ and $p \in C^2(\mathcal{X})$, then

$$\|\widetilde{L}_{P,\varepsilon}f - \sigma_\eta \Delta_P f\|_{L^2(\mathcal{X})}^2 \leq C_3 \varepsilon^2 \|f\|_{H^3(\mathcal{X})}^2. \tag{E.4}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Here Δ_P is the density-weighted Laplace–Beltrami operator. It is defined precisely as in (1.2), but with div and ∇ interpreted as the divergence and gradient operators on the manifold $\mathcal{X}$.

- In Section E.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

$$\langle L^s_{P,\varepsilon} f, f\rangle_P \leq C_4 \|f\|^2_{H^s(\mathcal{X})}. \tag{E.5}$$

- In Section E.5 we state some technical results used in the previous sections.

We point out that when f is Hölder smooth, results analogous to (E.4) have been established in Calder and García Trillos [16]. 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.

In the proof of (E.1)–(E.5), we recall the following estimates from differential geometry: (i) letting K_0 be an upper bound on the absolute value of the sectional curvatures of $\mathcal{X}$, $K_0 \leq 2R$, and letting (ii) i_0 be a lower bound on the injectivity radius of $\mathcal{X}$, $i_0 \geq \pi R$; see Proposition 1 of [1]. Additionally, recall that for all $\delta < i_0$, 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 $x \in \mathcal{X}$. We shall therefore always assume $\varepsilon < i_0$.

Proof of Proposition 7. Follows immediately from (E.1) and (E.5). □

E.1 *Decomposition of graph Sobolev seminorm*

The proof of (E.1) is identical to the proof of (D.1), except substituting the intrinsic dimension m for ambient dimension d, and using Lemma E.12 rather than Lemma D.8.

E.2 *Error due to Euclidean Distance*

In this section, we prove (E.2). By applying Cauchy–Schwarz we obtain an upper bound on $|L_{P,\varepsilon}f(x) - \widetilde{L}_{P,\varepsilon}f(x)|^2$:

$$\begin{aligned}\left[L_{P,\varepsilon}f(x) - \widetilde{L}_{P,\varepsilon}f(x)\right]^2 &\leq \frac{p^2_{\max}}{\varepsilon^{2(2+m)}} \int_{\mathcal{X}} \left[f(x') - f(x)\right]^2 \left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right) - \eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right)\right| d\mu(x') \\ &\quad \cdot \int_{\mathcal{X}} \left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right) - \eta\left(\frac{d_{\mathcal{X}}(x',x)}{\varepsilon}\right)\right| d\mu(x') \\ &= \frac{1}{\varepsilon^{2(2+m)}} A_1(x)\cdot A_2(x)\end{aligned} \tag{E.6}$$

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 estimate, from proposition 4 of [28]: for all $\|x'-x\| \leq R/2$,

$$\|x'-x\| \leq d_{\mathcal{X}}(x',x) \leq \|x'-x\| + \frac{8}{R^2}\|x'-x\|^3. \tag{E.7}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

From here forward we will assume $\varepsilon < R/2$.

Upper bound on $A_1(x)$. Consequently $\eta(\|x'-x\|/\varepsilon) \geq \eta(d_{\mathscr{X}}(x',x)/\varepsilon)$. For a corresponding upper bound, let L_η denote the Lipschitz constant of η on $[0,1]$ and set $\widetilde{\varepsilon} := (1+27\varepsilon^2/R^2)\varepsilon$. Then

$$\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right) - \eta\left(\frac{d_{\mathscr{X}}(x',x)}{\varepsilon}\right)\right| \leq \frac{L_\eta 8\varepsilon^2}{R^2} \cdot \mathbf{1}\left\{d_{\mathscr{X}}(x',x) \leq \varepsilon\right\} + \|\eta\|_\infty \cdot 1\{\varepsilon < d_{\mathscr{X}}(x',x) \leq \widetilde{\varepsilon}\}.$$

Thus,

$$\begin{aligned} A_1(x) \leq & \frac{8L_\eta\varepsilon^2}{R^2}\int_{\mathscr{X}} \left[f(x') - f(x)\right]^2 \mathbf{1}\{\|x'-x\| \leq \varepsilon\}\, d\mu(x') \\ & + \|\eta\|_\infty \int_{\mathscr{X}} \left[f(x') - f(x)\right]^2 \mathbf{1}\left\{\varepsilon < d_{\mathscr{X}}(x',x) \leq \widetilde{\varepsilon}\right\}\, d\mu(x') \end{aligned}$$

Integrating over $\mathscr{X}$, we conclude from Lemma 3.3 of [15] and Lemma E.11 that

$$\begin{aligned} \int_{\mathscr{X}} A_1(x)\, d\mu(x) \leq & \frac{8L_\eta \nu_m \varepsilon^2}{R^2(m+2)}\left(1 + CmK_0R^2\right)\varepsilon^{m+2}|f|^2_{H^1(\mathscr{X})} \\ & + C\|\eta\|_\infty \varepsilon^{m+4}|f|^2_{H^1(\mathscr{X})} =: C_5\varepsilon^{m+4}|f|^2_{H^1(\mathscr{X})}. \end{aligned}$$

Upper bound on $A_2(x)$. Integrating over $x' \in \mathscr{X}$, we see that

$$\begin{aligned} & \int_{\mathscr{X}}\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right) - \eta\left(\frac{d_{\mathscr{X}}(x',x)}{\varepsilon}\right)\right| d\mu(x') \\ & \leq \frac{8L_\eta\varepsilon^2}{R^2}\int_{\mathscr{X}} \mathbf{1}\left\{d_{\mathscr{X}}(x',x) \leq \varepsilon\right\}\, d\mu(x') + p_{\max}\|\eta\|_\infty \int_{\mathscr{X}} \mathbf{1}\left\{\varepsilon < d_{\mathscr{X}}(x',x) \leq \widetilde{\varepsilon}\right\}\, d\mu(x') \\ & = \frac{8L_\eta\varepsilon^2}{R^2} \cdot \mu\left(B_{\mathscr{X}}(x,\varepsilon)\right) + p_{\max}\|\eta\|_\infty\left[\mu\left(B_{\mathscr{X}}(x,\widetilde{\varepsilon})\right) - \mu\left(B_{\mathscr{X}}(x,\varepsilon)\right)\right]. \end{aligned} \tag{E.8}$$

Equation (1.36) in [28] states that

$$\left|\mu(B_{\mathscr{X}}(x,\varepsilon)) - \omega_m\varepsilon^m\right| \leq CmK_0\varepsilon^{m+2},$$

where we recall K_0 is an upper bound on the sectional curvature of $\mathscr{X}$. Plugging this back into (E.8), we conclude that

$$\begin{aligned} & \int_{\mathscr{X}}\left|\eta\left(\frac{\|x'-x\|}{\varepsilon}\right) - \eta\left(\frac{d_{\mathscr{X}}(x',x)}{\varepsilon}\right)\right| d\mu(x') \\ & \leq \frac{8L_\eta\varepsilon^2}{R^2}\left[\omega_m\varepsilon^m + CmK_0\varepsilon^{m+2}\right] + \|\eta\|_\infty\left[\omega_m(\widetilde{\varepsilon}^m - \varepsilon^m) + 2CmK_0\varepsilon^{m+2}\right] \\ & \leq \frac{8L_\eta\varepsilon^2}{R^2}\left[\omega_m\varepsilon^m + R^2CmK_0\varepsilon^m\right] + \|\eta\|_\infty\varepsilon^{m+2}\left[\frac{27\omega_m}{R^2} + 2CmK_0\right] =: C_6\varepsilon^{m+2}. \end{aligned}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Putting together the pieces. Plugging our upper bounds on $A_1(x)$ and $A_2(x)$ back into (E.6), we deduce that

$$\begin{aligned}\|\widetilde{L}_{P,\varepsilon}f - L_{P,\varepsilon}f\|^2_{L^2(\mathscr{X})} &\leq \frac{1}{\varepsilon^{2(2+m)}}\int_{\mathscr{X}} A_1(x)\cdot A_2(x)\,d\mu(x)\\ &\leq \frac{C_6}{\varepsilon^{(2+m)}}\int_{\mathscr{X}} A_1(x)\,d\mu(x)\\ &\leq C_5C_6\varepsilon^2|f|^2_{H^1(\mathscr{X})},\end{aligned}$$

thus proving the claimed result.

E.3 *Approximation error of non-local Laplacian*

Fix $x \in \mathscr{X}$. 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 [16]. Let $J_x(\cdot)$ be the Jacobian of the exponential map $\exp_x$, we have

$$\begin{aligned}\widetilde{L}_{P,\varepsilon}f(x) &= \frac{1}{\varepsilon^{m+2}}\int_{\mathscr{X}}\big(f(x') - f(x)\big)\,\eta\left(\frac{d_{\mathscr{X}}(x',x)}{\varepsilon}\right)dP(x')\\ &= \frac{1}{\varepsilon^{m+2}}\int_{B(0,\varepsilon)\subset T_x(\mathscr{X})}(w(v) - w(0))\,\eta\left(\frac{\|v\|}{\varepsilon}\right)J_x(v)q(v)\,dv\\ &= \frac{1}{\varepsilon^2}\left\{\int_{B(0,1)}(w(\varepsilon v) - w(0))\,\eta(\|v\|)q(\varepsilon v)\,dv + \int_{B(0,1)}(w(\varepsilon v) - w(0))\,\eta(\|v\|)q(\varepsilon v)\,(J_x(\varepsilon v) - 1)\,dv\right\}\\ &= A_1(x) + A_2(x)\end{aligned}$$

[[DmEquation201]]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 $f \in H^2(\mathscr{X})$ and $p \in C^1(\mathscr{X})$, then

$$\|A_1\|^2_{L^2(\mathscr{X})} \leq C\|f\|^2_{H^2(\mathscr{X})},$$

whereas if $f \in H^3(\mathscr{X})$ and $p \in C^2(\mathscr{X})$ then

$$\|A_1 - \sigma_\eta\Delta_P f\|^2_{L^2(\mathscr{X})} \leq C\|f\|^2_{H^3(\mathscr{X})}\varepsilon^2.$$

Therefore it remains only to upper bound A_2 in $L^2(\mathscr{X})$ norm. To do so, we recall (1.34) of [28]: for any $\varepsilon < i_0$ and all $x \in \mathscr{X}$, the Jacobian $J_x(v)$ satisfies the upper bound

$$|J_x(v) - 1| \leq CmK_0\varepsilon^2, \quad \text{for all } v \in B(0,\varepsilon) \subseteq T_x(\mathscr{X}).$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Combining this estimate with the Cauchy–Schwarz inequality, we conclude that

$$\begin{aligned}\|A_2\|^2_{L^2(\mathscr{X})} &\le Cm^2K_0^2\left[\int_{B(0,1)} (w(\varepsilon v) - w(0))^2\,\eta(\|v\|)q(\varepsilon v)\,dv\right]\cdot\left[\int_{B(0,1)} \eta(\|v\|)q(\varepsilon v)\,dv\right]\\ &\le Cm^2K_0^2\sigma_\eta(1+L_q\varepsilon)\int_{B(0,1)} (w(\varepsilon v) - w(0))^2\,\eta(\|v\|)q(\varepsilon v)\,dv\\ &\le Cm^2K_0^2\sigma_\eta^2(1+L_q\varepsilon)p_{\max}\varepsilon^2|f|^2_{H^1(\mathscr{X})},\end{aligned}$$

with the final inequality following from (3.2) of [15]. Combining our estimates on A_1 and A_2 yields the claim.

E.4 *Estimate of non-local Sobolev seminorm*

In this subsection we establish that the upper bound (E.5) holds when $f \in H^s(\mathscr{X})$ and $p \in C^{s-1}(\mathscr{X})$. We first consider $s = 2$, and then $s = 3$.

Case 1: $s = 2$. When $s = 2$, the triangle inequality implies that

$$\langle L^s_{P,\varepsilon} f, f\rangle_P \le 2p_{\max}\left(\|L_{P,\varepsilon} f - \widetilde{L}_{P,\varepsilon}\|^2_{L^2(\mathscr{X})} + \|\widetilde{L}_{P,\varepsilon} f\|^2_{L^2(\mathscr{X})}\right)$$

The first term on the right-hand side is upper bounded in (E.2), and the second term is upper bounded in (E.3). Together these estimates imply the claim.

Case 2: $s = 3$. When $s = 3$, the triangle inequality implies that

$$\begin{aligned}\langle L^s_{P,\varepsilon} f, f\rangle_P = E_{P,\varepsilon}(L_{P,\varepsilon} f;\mathscr{X}) &\le 3(E_{P,\varepsilon}(L_{P,\varepsilon} f - \widetilde{L}_{P,\varepsilon} f;\mathscr{X}) + E_{P,\varepsilon}(\widetilde{L}_{P,\varepsilon} f - \sigma_\eta \Delta_P f;\mathscr{X})\\ &\quad + \sigma_\eta^2 E_{P,\varepsilon}(\Delta_P f;\mathscr{X}))\end{aligned}$$

We now upper bound each of the three terms on the right-hand side of the above inequality. First, we note that by Lemma E.9 and (E.2),

$$E_{P,\varepsilon}(L_{P,\varepsilon} f - \widetilde{L}_{P,\varepsilon} f;\mathscr{X}) \le \frac{C}{\varepsilon^2}\|L_{P,\varepsilon} f - \widetilde{L}_{P,\varepsilon} f\|^2_{L^2(\mathscr{X})} \le C|f|^2_{H^1(\mathscr{X})}.$$

An equivalent upper bound on $E_{P,\varepsilon}(\widetilde{L}_{P,\varepsilon} f - \sigma_\eta \Delta_P f;\mathscr{X})$ follows from Lemma E.9 and (E.4). Finally, we notice that $f \in H^3(\mathscr{X})$ and $p \in C^2(\mathscr{X})$ implies $\Delta_P f \in H^1(\mathscr{X})$, and furthermore $|\Delta_P f|_{H^1(\mathscr{X})} \le \|p\|_{C^2(\mathscr{X})} \cdot \|f\|_{H^3(\mathscr{X})}$. We conclude from Lemma E.10 that

$$E_{P,\varepsilon}(\Delta_P f;\mathscr{X}) \le C|\Delta_P f|^2_{H^1(\mathscr{X})} \le C\|f\|^2_{H^3(\mathscr{X})},$$

where in the final inequality we have absorbed $\|p\|_{C^2(\mathscr{X})}$ into the constant C. Together, these upper bounds prove the claim.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

E.5 *Integrals*

Recall the Dirichlet energy $E_{P,\varepsilon}(f;\mathscr{X}) = \langle L_{P,\varepsilon}f,f\rangle_P$, defined in (D.25). Now we establish some estimates on $E_{P,\varepsilon}(f;\mathscr{X})$ that hold in the manifold setting, and under various assumptions regarding the regularity of f.

Lemma E.9. In the manifold setting, suppose additionally that $f \in L^2(\mathscr{X})$. Then there exists a constant C such that

$$E_{P,\varepsilon}(f;\mathscr{X}) \le \frac{C}{\varepsilon^2}\|f\|^2_{L^2(\mathscr{X})}. \tag{E.9}$$

Lemma E.10. In the manifold setting, suppose additionally that $f \in H^1(\mathscr{X})$. Then there exist constants c and C which do not depend on f such that for any $0 < \varepsilon < c$,

$$E_{P,\varepsilon}(f;\mathscr{X}) \le C|f|^2_{H^1(\mathscr{X})}. \tag{E.10}$$

We use Lemma E.11 to help upper bound the error incurred using $\|\cdot\|$ rather than $d_{\mathscr{X}}(\cdot,\cdot)$. Recall the notation $\widetilde{\varepsilon} = (1 + 27\varepsilon^2/R^2)\varepsilon$, where R is the reach of $\mathscr{X}$.

Lemma E.11. In the manifold setting, suppose additionally that $f \in H^1(\mathscr{X})$. There exist constants c and C such that for any $\varepsilon < c$,

$$\int_{\mathscr{X}}\int_{\mathscr{X}} \big(f(x') - f(x)\big)^2\,\mathbf{1}\{\varepsilon < d_{\mathscr{X}}(x',x) \le \widetilde{\varepsilon}\}\,d\mu(x')\,d\mu(x) \le C\varepsilon^{4+m}\|f\|^2_{H^1(\mathscr{X})} \tag{E.11}$$

Finally, we use Lemma E.12 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 D.8.

Lemma E.12. In the manifold setting, suppose additionally that $f \in H^1(\mathscr{X})$, and let $D_j f$ be defined with respect to a kernel η that satisfies **(K4)**. Then there exists a constant C which does not depend on f or n, such that for any $i,j \in [n]$ and $\mathbf{j} \in [n]^s$,

$$\mathbb{E}\Big[|D_{\mathbf{j}}f(X_i)|\cdot|f(X_i) - f(X_j)|\Big] \le C\varepsilon^{2+mk}\cdot\|f\|^2_{H^1(\mathscr{X})},$$

where $k+1$ is the number of distinct indices in $ij\mathbf{j}$.

Proof of Lemmas E.9 and E.10. Define the non-local energy $\widetilde{E}_{P,\varepsilon}$ with respect to geodesic distance,

$$\widetilde{E}_{P,\varepsilon}(f;\mathscr{X}) := \langle \widetilde{L}_{P,\varepsilon}f,f\rangle_P = \int_{\mathscr{X}}\int_{\mathscr{X}} \big(f(x') - f(x)\big)^2\,\eta\left(\frac{d_{\mathscr{X}}(x',x)}{\varepsilon}\right)dP(x')\,dP(x).$$

From the lower bound in (E.7), it follows that $E_{P,\varepsilon}(f;X) \le \widetilde{E}_{P,\varepsilon}(f;\mathscr{X})$, and from the upper bounds $p(x) \le p_{\max}$ and $\eta(|x|) \le \|\eta\|_\infty\cdot\mathbf{1}\{x \in [-1,1]\}$ we further have

$$\widetilde{E}_{P,\varepsilon}(f;\mathscr{X}) \le p^2_{\max}\|\eta\|_\infty\cdot\int_{\mathscr{X}}\int_{B_{\mathscr{X}}(\varepsilon)} \big(f(x') - f(x)\big)^2\,d\mu(x')\,d\mu(x).$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

The estimates (E.9) and (E.10) then respectively follow from (3.1) and Lemma 3.3 of [15].

Proof (of Lemma E.11). Following exactly the steps of the proof of Lemma 3.3 of Burago et al. [15], 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

$$\int_{\mathscr{X}}\int_{\mathscr{X}} \big(f(x') - f(x)\big)^2\, \mathbf{1}\{\varepsilon < d_{\mathscr{X}}(x', x) \le \widetilde{\varepsilon}\}\, d\mu(x')\, d\mu(x) \le (1 + CmK_0\varepsilon^2) \cdot \int_{\mathscr{X}}\int_{\Delta(\varepsilon,\widetilde{\varepsilon})} |d_x^1 f(v)|^2\, dv\, d\mu(x).$$

Here $\Delta_m(\varepsilon,\widetilde{\varepsilon}) = \{v : \varepsilon \le \|v\| \le \widetilde{\varepsilon}\}$, and $d_x^1 f(v)$ is the directional derivative of f at the point x in the direction v. From (2.7) of Burago et al. [15], we further have

$$\begin{aligned}\int_{\mathscr{X}}\int_{\Delta_m(\varepsilon,\widetilde{\varepsilon})} |d_x^1 f(v)|^2\, dv\, d\mu(x) &= \frac{\nu_m}{2+m}(\widetilde{\varepsilon}^{2+m} - \varepsilon^{2+m})\int_{\mathscr{X}} |d_x^1 f|^2\, d\mu(x) \\ &= 27\frac{\nu_m}{(2+m)R^2}\varepsilon^{4+m}\|d^1 f\|_{L^2(\mathscr{X})}^2.\end{aligned}$$

Noting that $\|d^1 f\|_{L^2(\mathscr{X})}^2 \le \|f\|_{H^1(\mathscr{X})}^2$, we see that this implies the claim of Lemma E.11.

Proof of Lemma E.12. The proof of Lemma E.12 is identical to the proof of Lemma D.8, upon substituting the ambient dimension m for the intrinsic dimension d, and using Lemma E.10 rather than Lemma D.7 to establish (D.35).

F. Lower bound on empirical norm

In this section we prove Propositions 6 (in Section F.1) and 9 (in Section F.2).

F.1 *Proof of Proposition 6*

In this section we establish Proposition 6. As mentioned, the proof of this proposition follows from the Gagliardo–Nirenberg interpolation inequality, and a one-sided Bernstein's inequality (Lemma I.17).

LEMMA F.13. (Gagliardo–Nirenberg interpolation inequality). In the flat Euclidean setting, suppose additionally that $f \in H^s(\mathscr{X})$ for some $s \ge d/4$. Then there exist constants C_1 and C_2 that do not depend on f, such that

$$\|f\|_{L^4(\mathscr{X})} \le C_1 |f|_{H^s(\mathscr{X})}^{d/4s} \|f\|_{L^2(\mathscr{X})}^{1-d/(4s)} + C_2\|f\|_{L^2(\mathscr{X})} \tag{F.1}$$

Proof of Proposition 6. Rearranging (F.1) and raising both sides to the fourth power, we see that

$$\frac{\mathbb{E}[f^4(X)]}{\|f\|_P^4} \le C\left(\frac{\|f\|_{L^4(\mathscr{X})}}{\|f\|_{L^2(\mathscr{X})}}\right)^4 \le C_1\left(\frac{|f|_{H^s(\mathscr{X})}}{\|f\|_{L^2(\mathscr{X})}}\right)^{d/s} + C_2;$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

here the constants C_1, C_2 are not the same as in (F.1). Taking the constant C in assumption (3.18) to be sufficiently large relative to C_1 and C_2, such that

$$C_1\left(\frac{|f|_{H^s(\mathscr{X})}}{\|f\|_{L^2(\mathscr{X})}}\right)^{d/s} \leq \frac{\delta n}{64},$$

we conclude

$$\frac{\mathbb{E}[f^4(X)]}{\|f\|_P^4} \leq \frac{\delta n}{8} + 8C_2^3.$$

The claim then follows from Lemma I.17, upon taking $c = 1/(64C_2^3)$ in the statement of Proposition 6.

F.2 *Proof of Proposition 9*

The proof of Proposition 9 follows exactly the same steps as the proof of Proposition 6, upon replacing Lemma F.13 by Lemma F.14.

Lemma F.14. (c.f Theorem 3.70 of Aubin [4]). In the manifold setting, suppose additionally that $f \in H^s(\mathscr{X})$ for some $s \geq m/4$. Then there exist constants C_1 and C_2 that do not depend on f, such that

$$\|f\|_{L^4(\mathscr{X})} \leq C_1 |f|_{H^s(\mathscr{X})}^{m/4s} \|f\|_{L^2(\mathscr{X})}^{1-m/(4s)} + C_2\|f\|_{L^2(\mathscr{X})}. \tag{F.2}$$

G. Proofs of main results

G.1 *Estimation Results*

Proof of Theorem 1. We condition on the event that the design points $X_1, \ldots, X_n$ satisfy

$$\langle L_{n,\varepsilon} f_0, f_0\rangle_n \leq \frac{C}{\delta} M^2 \quad \text{and} \quad \lambda_k \geq c \cdot \min\left\{\rho_k, \frac{1}{\varepsilon^2}\right\} \text{ for all } 1 \leq k \leq n. \tag{G.1}$$

Note that by Propositions 3 and 5, these statements are both satisfied with probability at least $1 - \delta - Cn\exp\{-cn\varepsilon^d\}$.

Conditional on (G.1), we have from Lemma C.1 that for any $1 \leq K \leq n$,

$$\|\widehat{f} - f_0\|_n^2 \leq C\left\{\frac{M^2}{\delta(\rho_{K+1} \wedge \varepsilon^{-2})} + \frac{K}{n}\right\},$$

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 $\rho_{K+1} \geq c(K+1)^{2/d}$ ((B.4), Weyl's Law) we can simplify the above expression to the following:

$$\|\widehat{f} - f_0\|_n^2 \leq C\left\{\frac{M^2}{\delta}(K+1)^{-2/d} + \frac{K}{n}\right\}. \tag{G.2}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

We now upper bound the right-hand side of (G.2), 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 (G.2) 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 (G.2) 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 1.

Proof of Theorem 3. Follows identically to the proof of Theorem 1, except substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 4 rather than Proposition 3 and Assumption **(P3)** rather than Assumption **(P1)**.

Proof of Theorem 7. Follows identically to the proof of Theorem 1, substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 7 rather than Proposition 3, Proposition 8 rather than Proposition 5 and Assumption **(P5)** rather than Assumption **(P2)**.

G.2 *Testing Results*

Proof of Theorem 2. We have already upper bounded the Type I error of φ in Lemma C.2, 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

$$\langle L_{n,\varepsilon} f_0, f_0 \rangle_n \leq \frac{C}{\delta} M^2, \quad \text{and} \quad \lambda_k \geq c \cdot \min\{\rho_k, \varepsilon^{-2}\} \text{ for all } 2 \leq k \leq n, \tag{G.3}$$

as well as that

$$\|f_0\|_n^2 \geq \frac{1}{2} \|f_0\|_P^2. \tag{G.4}$$

Note that by Propositions 3 and 5, both statements in (G.3) are satisfied with probability at least $1 - \delta - Cn \exp\{-cn\varepsilon^d\}$. Additionally, by Proposition 6 and the assumption in (3.5) that $\|f_0\|_P^2 \geq CM^2/(bn^{2/d})$, the one-sided inequality (G.4) 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 (G.3) and (G.4) are together satisfied with probability at least $1 - b/2$.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Now, to complete the proof of Theorem 2, we would like to invoke Lemma C.2, and conclude that conditional on $X_1, \ldots, X_n$ satisfying (G.3) and (G.4), our test φ will equal 1 with probability at least $1 - b/2$. To use Lemma C.2, we will need to establish that (C.4) is satisfied.

On the one hand, we have that the right-hand side of (C.4) is upper bounded,

$$\frac{\langle L_{n,\varepsilon} f_0, f_0 \rangle_n}{\lambda_{K+1}} + \frac{\sqrt{2K}}{n}\left[2\sqrt{\frac{1}{a}} + \sqrt{\frac{2}{b}} + \frac{32}{bn}\right] \leq C\left(\frac{M^2}{b \min\{\rho_{K+1}, \varepsilon^{-2}\}} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right)$$
$$\leq C\left(\frac{M^2}{b}(K+1)^{-2/d} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right),$$

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 2, it remains only to verify that

$$\|f_0\|_P^2 \geq C\left(\frac{M^2}{b}(K+1)^{-2/d} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right). \tag{G.5}$$

As in the estimation case, we can further upper bound the right-hand side of (G.5), depending on the value of K chosen in **(P2)**. If $K = \lfloor M^2 n \rfloor^{d/(2+d)}$ then (G.5) is satisfied as long as

$$\|f_0\|_P^2 \geq CM^2(M^2n)^{-4/(4+d)}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right].$$

If $M^2 < n^{-1}$, then we take $K = 1$, and (G.5) 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 (G.5) 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 (3.5) implies each of these three inequalities, and thus implies (G.5).

Proof of Theorem 4. Follows identically to the proof of Theorem 1, except substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 4 rather than Proposition 3 and Assumption **(P4)** rather than Assumption **(P2)**.

Proof of Theorem 8. Follows identically to the proof of Theorem 1, except substituting $L_{n,\varepsilon}^s$ for $L_{n,\varepsilon}$, λ_k^s for λ_k, and using Proposition 7 rather than Proposition 3, Proposition 8 rather than Proposition 5, Proposition 9 rather than Proposition 6 and Assumption **(P6)** rather than Assumption **(P2)**.

Proof of Theorem 5. Note that our choices of K and ε ensure that (G.3) (with $L_{n,\varepsilon}^s$ replacing $L_{n,\varepsilon}$) and (G.4) are satisfied with probability at least $1 - b/2$. Proceeding as in the proof of Theorem 2, we upper

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

bound the right-hand side of (C.4),

$$\frac{\langle L_{n,\varepsilon}f_0, f_0\rangle_n}{\lambda_{K+1}} + \frac{\sqrt{2K}}{n}\left[2\sqrt{\frac{1}{a}} + \sqrt{\frac{2}{b}} + \frac{32}{bn}\right] \leq C\left(\frac{M^2}{b\min\{\rho_{K+1}, \varepsilon^{-2}\}} + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right)$$
$$\leq C\left(\frac{M^2}{b}\varepsilon^2 + \frac{\sqrt{2K}}{n}\left[\sqrt{\frac{1}{a}} + \frac{1}{b}\right]\right).$$

Unlike in the proof of Theorem 2, we note that in this case $\varepsilon^2 \leq C\rho_K$ rather than vice versa. From here, proceeding as in the proof of Theorem 2 gives the claimed result.

H. Graph Laplacian methods and the cluster assumption

A main conclusion of our paper is that PCR-LE is minimax optimal for non-parametric regression over certain Sobolev classes. It is not the only optimal method. 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 PCR-LE is better than these two alternatives, in the sense of having (much) smaller risk. This is possible because PCR-LE performs 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 PCR-LE *density adaptivity*.

H.1 *Set-up*

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

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

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 now define kernel smoothing and least squares using eigenfunction of an unweighted Laplace operator. For a kernel function ψ and bandwidth parameter h, the kernel smoothing estimator $\widetilde{f}_{\mathrm{KS}}$ is

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

defined at a point $x \in \mathscr{X}$ as

$$\widetilde{f}_{\mathrm{KS}}(x) := \begin{cases} 0, & \text{if } d_{n,h}(x) = 0, \\ \frac{1}{d_{n,h}(x)} \sum_{i=1}^{n} Y_i \psi\left(\frac{\|X_i - x\|}{h}\right), & \text{otherwise.} \end{cases} \tag{H.2}$$

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

$$\Delta \phi_k = \lambda_k \phi_k, \;\; \|\phi_k\|_{L^2([0,1])} = 1, \;\; \frac{d}{dx}\phi_k(0) = \frac{d}{dx}\phi_k(1) = 0. \tag{H.3}$$

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

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

The least squares estimator using $\phi_1, \ldots, \phi_K$ $(1 \leq K \leq n)$ eigenfunctions as features is simply[12]

$$\widetilde{f}_K := \operatorname*{argmin}_{f \in \mathrm{span}\{\phi_1, \ldots, \phi_K\}} \|Y - f\|_n^2 = \Phi(\Phi^\top \Phi)^{-1} \Phi^\top Y. \tag{H.4}$$

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

H.2 *Upper bounds on risk of PCR-LE*

Now we are in a position to state our results. Both PCR-LE and kernel smoothing depend in part on the choice of kernel. For simplicity, in our analysis we only consider the boxcar kernel,

$$\eta(z) = \psi(z) = \mathbf{1}\{z \leq 1\}. \tag{H.5}$$

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

PROPOSITION H.10. Suppose $(X_1, Y_1), \ldots (X_n, Y_n)$ are sampled according to (H.1). Compute the PCR-LE estimator $\widehat{f}$ using a kernel η which satisfies (H.5), number of eigenvectors $K = 2$, and radius $\varepsilon = r/2$. Then,

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

Proof of Proposition H.10. 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:

[12] 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 $\|\widetilde{f}_K - f_0\|_n^2 \geq 1/2$ with high probability.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

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 I.19) 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 neighbourhood 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,

$$\mathbb{P}\left(\mathrm{span}\{v_1, v_2\} = \mathrm{span}\{u_1, u_2\}\right) \geq 1 - 4/\varepsilon \exp(-n\varepsilon/4). \tag{H.7}$$

Let us condition on the 'good' event $\mathscr{E}$ that the design points $X_1, \ldots, X_n$ satisfy (I.1), and therefore that $\mathrm{span}\{v_1, v_2\} = \mathrm{span}\{u_1, u_2\}$. Consider the empirical mean $\overline{Y}_Q := \frac{1}{\sharp\{Q \cup \mathbf{X}\}} \sum_{i:X_i \in Q} Y_i$. Since $\mathrm{span}\{v_1, v_2\} = \mathrm{span}\{u_1, u_2\}$, the estimator $\widehat{f} = \widehat{f}_{\mathrm{LE}}$ will be piecewise constant on Q_1 and Q_2, and in fact we have that

$$\widehat{f} = \overline{Y}_{Q_1} u_1 + \overline{Y}_{Q_2} u_2. \tag{H.8}$$

Therefore conditional on $\mathscr{E}$,

$$\|\widehat{f} - f_0^{(n)}\|_n^2 = P_n(Q_1) \cdot (\overline{Y}_{Q_1} - \theta)^2 + P_n(Q_2) \cdot (\overline{Y}_{Q_2} + \theta)^2$$

and consequently,

$$\mathbb{E}\left[\|\widehat{f} - f_0^{(n)}\|_n^2 \middle| \mathscr{E}\right] = \mathbb{E}\left[\mathbb{E}\left[\|\widehat{f} - f_0^{(n)}\|_n^2 | X_1, \ldots, X_n\right] \middle| \mathscr{E}\right] = \frac{1}{n}. \tag{H.9}$$

Now we derive a crude upper bound on $\|\widehat{f} - f_0^{(n)}\|_n$ that will suffice to control the error conditional on $\mathscr{E}^c$. We observe that the empirical norm of $\widehat{f}$ is bounded,

$$\|\widehat{f}\|_n^2 \leq \frac{2}{n} \sum_{i=1}^{n} \langle Y, v_1\rangle_n^2 v_{1,i}^2 + \langle \mathbf{Y}, v_2\rangle_n^2 v_{2,i}^2 \leq 2\left(\langle \mathbf{Y}, v_1\rangle_n^2 + \langle \mathbf{Y}, v_2\rangle_n^2\right) \leq 4\|\mathbf{Y}\|_n^2.$$

Noting that $\mathbb{E}[\|\mathbf{Y}\|_n^2 | X_1, \ldots, X_n] = \|f_0\|_n^2 + 1/n = \theta^2 + 1/n$, we conclude that

$$\mathbb{E}\left[\|\widehat{f} - f_0\|_n^2 \cdot \mathbf{1}\{\mathscr{E}^c\}\right] \leq \mathbb{E}\left[(2\|f_0\|_n^2 + 4(\theta^2 + 1/n) \cdot \mathbf{1}\{\mathscr{E}^c\}\right] \leq (6\theta^2 + n^{-1}) \cdot 4\varepsilon^{-1} \exp(-n\varepsilon/4).$$

Combining this with (H.9) implies (H.6).

H.3 *Lower bounds on risk of kernel smoothing and least squares*

Proposition H.11. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled according to (H.1). Suppose $(\log n)^2/n \leq r \leq c$, where c is a universal constant.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

- Compute the kernel smoothing estimator $\widetilde{f} = \widetilde{f}_{\mathrm{KS}}$ as in (H.2), using a kernel ψ which satisfies (H.5). Then there exist universal constants $c, N > 0$ such that for all $n > N$,

$$\inf_{h'>0} \mathbb{E}\left[\|\widetilde{f} - f_0^{(n)}\|_n^2\right] \geq c \min\left\{\frac{r^{-1}}{n}, \frac{\theta}{\sqrt{n}}\right\}. \tag{H.10}$$

- Compute the least squares estimator $\widetilde{f} = \widetilde{f}_{\mathrm{SP}}$ as in (H.4). Then there exist universal constants $c, N > 0$ such that for all $n > N$,

$$\inf_{1\leq K\leq n} \mathbb{E}\left[\|\widetilde{f} - f_0^{(n)}\|_n^2\right] \geq c \min\left\{\frac{r^{-1}}{n}, \frac{1}{\log(n)}, \frac{r^{-2/3}}{n}, \frac{\sqrt{\theta}}{n^{3/4}}\right\}. \tag{H.11}$$

The proof of Proposition H.11 is long, and we defer it until after some discussion of the implications of the proposition.

Together, Propositions H.10 and H.11 illustrate that the risk of PCR-LE 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, $\widehat{f}$ satisfies

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

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

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

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

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

Other choices of θ and r lead to even more dramatic gaps between the risk of PCR-LE, and the risk of kernel smoothing and least squares. The overall takeaway is that under Model H.1, 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—non-parametric 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 (H.10) is larger than that of $\widetilde{f}_{\mathrm{SP}}$ given by (H.11). 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.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

- 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 PCR-LE should have the advantage on non-density adaptive linear smoothers, such as kernel smoothing or uniform least squares. Indeed, in the case of Model H.1 we see that

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

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 analyse 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. [35] for an extensive discussion.

H.4 *Proof of Proposition H.11*

First we show (H.10), then (H.11).

H.4.1 *Proof of (H.10).* 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,

$$\mathbb{E}\Big[\big(\tilde{f}(X_i) - f_0(X_i)\big)^2 | X_i = x\Big] \geq \frac{(n-1)}{n}\mathbb{E}\Big[\big(f_0(X) - f_0(x)\big)^2 | X \in B(x,h')\Big] + \mathbb{E}\Big[\frac{1}{d_{n,h'}(x)}\Big].$$

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

$$\mathbb{E}[(f_0(X) - f_0(x))^2 | X \in B(x,h')] \geq \frac{\theta^2}{8} \quad \text{for all } x \in [(1-h'/2)_+, 1/2 - r].$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Combining these lower bounds on variance and squared bias terms and summing over $X_1,\dots,X_n$, we arrive at the following: if $h' \leq 4r$, then

$$\mathbb{E}\left[\|\widetilde{f} - f_0^{(n)}\|_n^2\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\left(\widetilde{f}(X_i) - f_0(X_i)\right)^2 |X_i\right]\right] \geq \frac{1}{16rn},$$

whereas if $h' > 4r$ then

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{f} - f_0^{(n)}\|_n^2\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\left(\widetilde{f}(X_i) - f_0(X_i)\right)^2 |X_i\right]\right] \\ &\geq \frac{1}{4nh'} + \frac{\theta^2}{8}\frac{(n-1)}{n}P^{(n)}\left([(1-h'/2)_+, 1/2 - r]\right) \\ &\geq \frac{1}{4nh'} + \frac{\theta^2 h'}{64}.\end{aligned}$$

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 (H.10).

H.4.2 *Proof of (H.11).* 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 $\mathbb{E}_n[w] = 0$,

$$\mathbb{E}\left[\|\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\right] = \mathbb{E}\left[\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\right] + \mathbb{E}\left[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\right].$$

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; also, we will show that the expected conditional squared bias is lower bounded,

$$\mathbb{E}\left[\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\right] = \frac{K}{n} \quad \text{and} \quad \mathbb{E}\left[\|\mathbb{E}_n\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\right] \geq \frac{\theta^2}{2601\pi^2K^3}, \tag{H.12}$$

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,

$$\mathbb{E}\left[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\right] \geq \min\left\{\frac{1}{16rn}, \frac{1}{8\log(8n)}, \frac{(\sqrt{160}\pi)^{2/3}}{r^{2/3}n}\right\} \tag{H.13}$$

Otherwise (H.12) implies that the in-sample risk is always at least

$$\mathbb{E}\left[\|\widetilde{f}_{\mathrm{SP}} - f_0\|_n^2\right] \geq \frac{\theta^2}{2601\pi^2K^3} + \frac{K}{n} \geq 2\frac{\theta^{1/2}}{n^{3/4}}\frac{1}{(2601\pi^2)^{1/4}}.$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

Along with (H.13), 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

$$\mathbb{E}_n\left[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\right] = \frac{1}{n}\mathrm{tr}(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top) = \frac{K}{n};$$

and finally by the law of iterated expectation and the independence of the noise $(w_1,\ldots,w_n)$ and the design points $X_1,\ldots,X_n$,

$$\mathbb{E}\Big[\mathbb{E}_n\left[\|\widetilde{f}_{\mathrm{SP}} - \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}\|_n^2\right]\Big] = K/n.$$

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 $(\phi_1,\ldots,\phi_K)$, 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 H.15. (Lipschitz regularity of $\mathbb{E}_n\widetilde{f}_{\mathrm{SP}}$). Let $\widetilde{f}_n = \mathbb{E}_n\widetilde{f}_{\mathrm{SP}}$. Then

$$\|\widetilde{f}_n\|^2_{C^1(\mathscr{X})} \le \pi^2\frac{K^3\cdot\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}}}{(1-\|I_K-\Sigma\|_F)}\cdot\|\widetilde{f}_n\|_n^2. \tag{H.14}$$

Moreover, suppose $K \le 1/(16r)$ and $r \le (1-2^{-1/2})/2$.

- (Matrix perturbation) Then

$$\|\Sigma - I_K\|_F \le \frac{1}{2}. \tag{H.15}$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

- (Matrix concentration, cf. [37]) If additionally $n \geq 8K\log(K/\delta)$ for some $\delta \in (0, 1/2)$, then with probability at least $1 - 2\delta$,

$$\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\|_{\mathrm{op}} \leq 5. \tag{H.16}$$

Therefore, if $K \leq \min\{1/(16r), n/(8\log(K/\delta))\}$, then with probability at least $1 - 2\delta$,

$$\|\widetilde{f}_n\|^2_{C^1(\mathcal{X})} \leq 10\pi^2 K^3 \|\widetilde{f}_n\|^2_n. \tag{H.17}$$

We defer the proof of Lemma H.15 until after we complete the proof of (H.11).

Now, if $\|\widetilde{f}_n\|_n^2 \geq \frac{3}{2}\|f_0\|_n^2$, then by the triangle inequality

$$\|\widetilde{f}_n - f_0\|_n \geq \|\widetilde{f}_n\|_n - \|f_0\|_n \geq \sqrt{\frac{3}{2}} \cdot \|f_0\|_n = \sqrt{\frac{3}{2}} \cdot \theta.$$

Otherwise $\|\widetilde{f}_n\|_n^2 \geq \frac{3}{2}\|f_0\|_n^2$. In this case, we show that $|\widetilde{f}_n(X_i) - f_0(X_i)|$ must be large (of the order of θ) for many points X_i which are close to $x = 1/2$. Let us suppose without loss of generality that $\widetilde{f}_n(1/2) \leq \theta/2$ 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 H.15 we have that with probability at least $1 - 2\delta$,

$$|\widetilde{f}_n(X_i) - \widetilde{f}_n(1/2)| \leq CK^{3/2}\|\widetilde{f}_n\|_n \cdot |X_i - 1/2| \leq \sqrt{10}\pi K^{3/2}\theta \cdot |X_i - 1/2|.$$

Since $\widetilde{f}_n(1/2) \leq \theta/2$ and $f_0(X_i) = \theta/2$ for all $X_i \in Q_1$ it follows that

$$|\widetilde{f}_n(X_i) - f_0(X_i)| \geq \theta - \sqrt{10}\pi K^{3/2}\theta \cdot |X_i - 1/2|,$$

and consequently

$$|\widetilde{f}_n(X_i) - f_0(X_i)| \geq \theta/2, \quad \text{for any } X_i \in Q_1 \text{ such that } |X_i - 1/2| \leq 1/(\sqrt{40}\pi K^{3/2}).$$

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

$$\|\widetilde{f}_n - f_0\|_n \geq \frac{\theta}{2} \cdot P_n\left(Q_k\right).$$

Then as long as $K^{-3/2} \geq \sqrt{160}\pi r$, from the multiplicative form of Hoeffding's inequality (Lemma I.18)

$$P^{(n)}(Q_K) \geq \frac{1}{\sqrt{160}\pi K^{3/2}} \geq 2r \Longrightarrow \mathbb{P}\left(P_n(Q_K) \geq \frac{1}{\sqrt{640}\pi K^{3/2}}\right) \geq 1 - \exp(-nr/4) \geq 1 - \frac{4}{n^2}.$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

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 H.15. Proof of (H.14). 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

$$|\widetilde{f}_n'(x)|^2 \le \pi^2 K^2 \sum_{k=1}^{K} (\sin(k\pi x))^2 \, \|\beta\|_2^2 \le \pi^2 K^3 \|\beta\|_2^2.$$

On the other hand, we can also relate the empirical norm $\|\widetilde{f}_n\|_n^2$ to the ℓ^2 norm of β. Specifically,

$$\|\widetilde{f}_n\|_n^2 = \beta^\top \widehat{\varSigma}\beta \ge \frac{\|\beta\|_2^2}{\|\widehat{\varSigma}^{-1}\|_{\mathrm{op}}} \ge \frac{\|\beta\|_2^2}{\|\varSigma^{-1}\|_{\mathrm{op}} \cdot \|\varSigma^{1/2}\widehat{\varSigma}^{-1}\varSigma^{1/2}\|_{\mathrm{op}}} = \frac{\|\beta\|_2^2 \|\varSigma\|_{\mathrm{op}}}{\|\varSigma^{1/2}\widehat{\varSigma}^{-1}\varSigma^{1/2}\|_{\mathrm{op}}}$$

Rearranging, we see that

$$\sup_{x\in[0,1]} |\widetilde{f}_n'(x)|^2 \le \frac{\pi^2 K^3}{\|\varSigma\|_{\mathrm{op}}} \|\varSigma^{1/2}\widehat{\varSigma}^{-1}\varSigma^{1/2}\|_{\mathrm{op}} \le \frac{\pi^2 K^3}{1 - \|I_K - \varSigma\|_F} \|\varSigma^{1/2}\widehat{\varSigma}^{-1}\varSigma^{1/2}\|_{\mathrm{op}}$$

with the latter inequality following since $\|\varSigma\|_{\mathrm{op}} \ge \|I_K\|_{\mathrm{op}} - \|I_K - \varSigma\|_{\mathrm{op}} \ge 1 - \|I_K - \varSigma\|_F$.

Proof of (H.15). We will show that for all $1 \le k < \ell \le K$,

$$(1 - \langle \phi_k, \phi_k \rangle_{P^{(n)}})^2 \le 32r^2, \quad \text{and} \quad |\langle \phi_k, \phi_\ell \rangle_{P^{(n)}}| \le 64r^2. \tag{H.18}$$

This implies $\|I - \varSigma\|_F^2 \le 32K^2r^2$, so that $\|I - \varSigma\|_F \le 1/2$ so long as $K \le 1/(16r)$.

The proof of (H.18) 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

$$\langle \phi_k, \phi_\ell \rangle_{P^{(n)}} = \frac{-2\sqrt{2}}{(1-2r)} \cdot \frac{\cos(\ell\pi/2)\sin(\ell\pi r)}{\ell\pi},$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

and therefore

$$\left[\langle \phi_k, \phi_\ell \rangle_{P^{(n)}}\right]^2 \leq \frac{8}{(1-2r)^2} \cdot \left(\frac{\sin(\ell \pi r)}{\ell \pi}\right)^2 \leq \frac{8}{(1-2r)^2} r^2 \leq 16 r^2,$$

where in the second-to-last inequality follows because $\sin(x)/x \leq 1$, and the last inequality follows by our assumed upper bound on r.

Case 2: $k > 1$. When $k > 1$,

$$\langle \phi_k, \phi_k \rangle_{P^{(n)}} = 1 - \frac{2}{(1-2r)} \frac{\cos(k\pi)\sin(2k\pi r)}{k\pi} \Longrightarrow \left[1 - \langle \phi_k, \phi_k \rangle_{P^{(n)}}\right]^2$$
$$\leq \frac{4}{(1-2r)^2} \cdot \left(\frac{\sin(2k\pi r)}{k\pi}\right)^2 \leq 32 r^2.$$

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

$$\left[\langle \phi_k, \phi_\ell \rangle_{P^{(n)}}\right]^2 \leq \frac{16}{(1-2r)^2} \left(\left[\frac{\sin((k+\ell)\pi r)}{(k+\ell)\pi}\right]^2 + \left[\frac{\sin((k-\ell)\pi r)}{(k-\ell)\pi}\right]^2\right) \leq 64 r^2.$$

Proof of (H.16). Denote $\Phi(x) = (\phi_1, \dots, \phi_K(x)) \in \mathbb{R}^K$ for any $x \in [0,1]$. Then for any $x \in [0,1]$,

$$\|\Sigma^{-1/2}\Phi(x)\| \leq \|\Sigma^{-1}\|_{\mathrm{op}}^{1/2} \|\Phi(x)\|_2 \leq \|\Sigma^{-1}\|_{\mathrm{op}}^{1/2} \sqrt{2K} \leq 2\sqrt{K}$$

with the second-to-last inequality following from (H.15), and the last inequality following since $|\phi_k(x)| \leq \sqrt{2}$ for all k. Thus $\|\Sigma^{-1/2}\Phi(x)\|/\sqrt{K} \leq 2$, and (H.16) follows from Theorem 1 of Hsu et al. [37].

Proof of (H.17). Follows immediately.

I. 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.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

I.1 *Concentration Inequalities*

Lemma I.16 controls the deviation of a chi-squared random variable. It is from [45].

LEMMA I.16. 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 $t > 0$,

$$\mathbb{P}\left[U \geq 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\right] \leq \exp(-t).$$

In particular if $a_k = 1$ for each $k = 1, \ldots, N$, then

$$\mathbb{P}\left[U \geq 2\sqrt{Nt} + 2t\right] \leq \exp(-t).$$

Lemma I.17 is an immediate consequence of the one-sided Bernstein's inequality (14.23) in [70].

LEMMA I.17. (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 $1 - \exp\left(-n/8 \cdot \|f\|_P^4/\mathbb{E}[f^4(X)]\right)$.

Lemma I.18 is a multiplicative form of Hoeffding's inequality.

LEMMA I.18. (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

$$\mathbb{P}\left(\left|S_n - \mu\right| \geq \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{3B^2}\right).$$

The following Lemma gives a 'balls-in-bins' result. More precisely, it gives a lower bound on the probability that every bin

$$Q_{i1} = [i/m, (i+1)/m] \cdot (1/2 - r), \quad Q_{i2} = 1/2 + [i/m, (i+1)/m] \cdot (1/2 - r).$$

will contain at least one ball.

LEMMA I.19. Suppose $(X_1, Y_1), \ldots (X_n, Y_n)$ are sampled according to (H.1), and suppose $r \leq 1/4$. We have that

$$\mathbb{P}\left(\sharp\{Q_{ij} \cup \mathbf{X}\} > 0 \ \text{ for all } i = 1, \ldots, m-1 \text{ and } j = 1, 2\right) \geq 1 - 2m\exp\{-n/2m\}. \tag{I.1}$$

Proof of Lemma I.19. For each Q_{ij}, we have that $P(Q_{ij}) = (1/2 - r)/m \geq 1/(2m)$. Therefore

$$\mathbb{P}\left(\sharp\{Q_{ij} \cup \mathbf{X}\} = 0\right) = (1 - 1/(2m))^n \leq \exp\{-n/2m\}.$$

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

By a union bound,

$$\mathbb{P}\left(\sharp\{Q_{ij} \cup \mathbf{X}\} = 0 \text{ for any } i = 1, \ldots, m-1 \text{ and } j = 1, 2\right) \leq 2m \exp\{-n/2m\}.$$

□

Let $\varepsilon = 2/m$. Note that by construction, (I.1) 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 $x \in Q_1 \cup Q_2$.

I.2 *Taylor expansion*

We begin with some notation that allows us to concisely derivatives. For a given $z \in \mathbb{R}^d$ and s-times differentiable function $f : \mathcal{X} \to \mathbb{R}$, we denote $\left(d_x^s f\right)(z) := \sum_{|\alpha|=s} D^\alpha f(x) z^\alpha$. 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 $k \in \mathbb{N}\backslash\{0\}$. 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

$$u(x') = u(x) + \sum_{j=1}^{s-1} \frac{1}{j!} \left(d_x^j u\right)(x' - x) + r_{x'}^s(x; u)$$

For notational convenience we have adopted the convention that $\sum_{j=1}^{0} a_j = 0$. Thus $\left(d_x^j f\right)(z)$ is a degree-j polynomial—and so a j-homogeneous function—in z, meaning for any $t \in \mathbb{R}$,

$$\left(d_x^j f\right)(tz) = t^j \cdot \left(d_x^j f\right)(z).$$

The remainder term $r_{x'}$ is given by

$$r_{x'}^s(x; f) = \frac{1}{(j-1)!} \int_0^1 (1-t)^{j-1} \left(d_{x+t(x'-x)}^s f\right)(x' - x)\, dt,$$

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

$$\sup_{x \in \mathcal{X}_h} |r_{x+hz}^j(x; f)| \leq C h^j \|f\|_{C^j(\mathcal{X})},$$

whereas in $L^2(\mathcal{X}_h)$ norm we have

$$\int_{\mathcal{X}_h} \left|r_{x+thz}^j(x; f)\right|^2 dx \leq h^{2j} \int_{\mathcal{X}_h} \int_0^1 |d_{x+thz}^j f(z)|^2\, dt\, dx \leq h^{2j} \|d^j f\|_{L^2(\mathcal{X})}^2. \tag{I.2}$$

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

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023

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

$$\int_{B(0,1)} q_j(z)\eta(\|z\|)\,dz = 0.$$

J. Sparsification

Recall that when $s = 1$, we have shown that PCR-LE 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 PCR-LE 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. In this dense-graph setting, we now discuss a procedure to more efficiently compute an approximation to the PCR-LE estimate: *edge sparsification*.

By now there exist various methods see (e.g. the seminal papers of Spielman and Teng [61, 62, 63], or the overview by Vishnoi [68] 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 $\check{G}$, $E(\check{G}) \subseteq E(G)$, with a Laplacian $\check{L}_{n,\varepsilon}$ satisfying

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

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

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

with probability at least $1 - \exp(-K)$. Consequently $\|\widetilde{f} - f_0\|_n^2$is at most $\sigma^{2s} \cdot \|\widehat{f} - f_0\|_n^2$, and for any choice of σ that is constant in n the estimator $\check{f}$ 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 very close to 1, whereas in order for $\check{f}$ 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. [52], in this special case there are far simpler and faster methods for sparsification, which at least empirically seem to do the job.

Downloaded from https://academic.oup.com/imaiai/article/12/3/2423/7265386 by guest on 17 October 2023