Supplementary Material for "Minimax Optimal Regression over Sobolev Spaces via Laplacian Regularization on Neighborhood Graphs"

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

Carnegie Mellon University

A1 Preliminaries

In this supplement, we provide complete proofs of all results found in the paper "Minimax Optimal Regression over Sobolev Spaces via Laplacian Regularization on Neighborhood Graphs". Our main theorems (Theorems 1-5) all follow the same general proof strategy of first establishing bounds in the fixed-design setup. In Section A2, we establish (estimation or testing) error bounds which hold for any graph G; these bounds are stated with respect to (functionals of) the graph G, and allow us to upper bound the error of $\widehat{f}$ and $\widehat{\varphi}$ conditional on the design $\{X_1, \dots, X_n\} = \{x_1, \dots, x_n\}$. In Sections A3, A4, A5, and A6 we develop all the necessary probabilistic estimates on these functionals, for the particular random neighborhood graph $G = G_{n,r}$. It is in these sections where we invoke our various assumptions on the distribution P and regression function f_0. In Section A7, we prove our main theorems and some other results. In Section A8, we state a few concentration bounds that we use repeatedly in our proofs.

Pointwise evaluation of Sobolev functions. First, however, as promised in our main text we clarify what is meant by pointwise evaluation of the regression function f_0. Strictly speaking, each $f \in H^1(X)$ is really an equivalence class, defined only up to sets of Lebesgue measure 0. In order to make sense of the evaluation $x \mapsto f(x)$, one must therefore pick a representative $f^\star \in f$. When $d = 1$, this is resolved in a standard way—since $H^1(\mathcal{X})$ embeds continuously into $C^0(\mathcal{X})$, there exists a continuous version of every $f \in H^1(\mathcal{X})$, and we take this continuous version as the representative $f^\star$. On the other hand, when $d \geq 2$, the Sobolev space $H^1(\mathcal{X})$ does not continuously embed into $C^0(\mathcal{X})$, and we must choose representatives in a different manner. In this case we let $f^\star$ be the precise representative [Evans and Gariepy, 2015], defined pointwise at points $x \in \mathcal{X}$ as

$$f^\star(x) = \begin{cases} \lim\limits_{\varepsilon \to 0} \dfrac{1}{\nu(B(x,\varepsilon))} \displaystyle\int_{B(x,\varepsilon)} f(z)dz, & \text{if the limit exists,} \\ 0, & \text{otherwise.} \end{cases}$$

Note that when $d = 1$, the precise representative of any $f \in H^1(\mathcal{X})$ is continuous.

Now we explain why the particular choice of representative is not crucial, using the notion of a Lebesgue point. Recall that for a locally Lebesgue integrable function f, a given point $x \in \mathcal{X}$ is a *Lebesgue point* of f if the limit of $1/(\nu(B(x,\varepsilon))) \int_{B(x,\varepsilon)} f(x)dx$ as $\varepsilon \to 0$ exists, and satisfies

$$\lim_{\varepsilon \to 0} \frac{1}{\nu\big(B(x,\varepsilon)\big)} \int_{B(x,\varepsilon)} f(x)dx = f(x).$$

Let E denote the set of Lebesgue points of f. By the Lebesgue differentiation theorem [Evans and Gariepy, 2015], if $f \in L^1(\mathcal{X})$ then almost every $x \in \mathcal{X}$ is a Lebesgue point, $\nu(\mathcal{X} \setminus E) = 0$. Since $f_0 \in H^1(\mathcal{X}) \subseteq L^1(\mathcal{X})$, we can conclude that any function $g_0 \in f_0$ disagrees with the precise representative $f_0^\star$ only on a set of Lebesgue measure 0. Moreover, since we always assume the design distribution P has a continuous density, with probability 1 it holds that $g_0(X_i) = f_0^\star(X_i)$ for all $i = 1, \dots, n$. This justifies the notation $f_0(X_i)$ used in the main text.

A2 Graph-dependent error bounds

In this section, we adopt the fixed design perspective; or equivalently, condition on $X_i = x_i$ for $i = 1, \dots, n$. Let $G = \big([n], W\big)$ be a fixed graph on $\{1, \dots, n\}$ with Laplacian matrix $L = D - W$. The randomness thus all comes from the responses

$$Y_i = f_0(x_i) + \varepsilon_i \tag{A.1}$$

where the noise variables ε_i are independent $N(0,1)$. In the rest of this section, we will mildly abuse notation and write $f_0 = (f_0(x_1), \dots, f_0(x_n)) \in \mathbb{R}^n$. We will also write $\mathbf{Y} = (Y_1, \dots, Y_n)$.

Recall (2) and (3): the Laplacian smoothing estimator of f_0 on G is

$$\widehat{f} := \underset{f \in \mathbb{R}^n}{\operatorname{argmin}} \Bigg\{ \sum_{i=1}^{n} (Y_i - f_i)^2 + \rho \cdot f^\top L f \Bigg\} = (\rho L + I)^{-1}\mathbf{Y}.$$

and the Laplacian smoothing test statistic is

$$\widehat{T} := \frac{1}{n}\|\widehat{f}\|_2^2.$$

We note that in this section, many of the derivations involved in upper bounding the estimation error of $\widehat{f}$ are similar to those of Sadhanala et al. [2016], with the difference being that we seek bounds in high probability rather than in expectation. We keep the work here self-contained for purposes of completeness.

A2.1 Error bounds for linear smoothers

Let $S \in \mathbb{R}^{n \times n}$ be a fixed square, symmetric matrix, and let

$$\check{f} := SY$$

be a linear estimator of f_0. In Lemma 3 we upper bound the error $\frac{1}{n}\|\check{f} - f_0\|_2^2$ as a function of the eigenvalues of S. Let $\boldsymbol{\lambda}(S) = (\lambda_1(S), \dots, \lambda_n(S)) \in \mathbb{R}^n$ denote these eigenvalues, and let $v_k(S)$ denote the corresponding unit-norm eigenvectors, so that $S = \sum_{k=1}^{n} \lambda_k(S) \cdot v_k(S) v_k(S)^\top$. Denote $Z_k = v_k(S)^\top \varepsilon$, and observe that $\mathbf{Z} = (Z_1, \dots, Z_n) \sim N(0, I)$.

Lemma 3. *Let $\check{f} = SY$ for a square, symmetric matrix, $S \in \mathbb{R}^{n \times n}$. Then*

$$\mathbb{P}_{f_0}\left(\frac{1}{n}\big\|\check{f} - f_0\big\|_2^2 \geq \frac{10}{n}\big\|\boldsymbol{\lambda}(S)\big\|_2^2 + \frac{2}{n}\big\|(S - I)f_0\big\|_2^2\right) \leq 1 - \exp\Big(-\big\|\boldsymbol{\lambda}(S)\big\|_2^2\Big)$$

Here we have written $\mathbb{P}_{f_0}(\cdot)$ for the probability law under the regression "function" $f_0 \in \mathbb{R}^n$.

In Lemma 4, we upper bound the error of a test involving the statistic $\|\check{f}\|_2^2 = \mathbf{Y}^\top S^2 \mathbf{Y}$. We will require that S be a *contraction*, meaning that it has operator norm no greater than 1, $\|Sv\|_2 \leq \|v\|_2$ for all $v \in \mathbb{R}^n$.

Lemma 4. *Let $\check{T} = \mathbf{Y}^\top S^2 \mathbf{Y}$ for a square, symmetric matrix $S \in \mathbb{R}^{n \times n}$. Suppose S is a contraction. Define the threshold $\check{t}_\alpha$ to be*

$$\check{t}_\alpha := \|\boldsymbol{\lambda}(S)\|_2^2 + \sqrt{\frac{2}{\alpha}}\|\boldsymbol{\lambda}(S)\|_4^2. \tag{A.2}$$

It holds that:

- ***Type I error.***

$$\mathbb{P}_0\big(\check{T} > \check{t}_\alpha\big) \leq \alpha. \tag{A.3}$$

- ***Type II error.*** *Under the further assumption*

$$f_0^\top S^2 f_0 \geq \left(2\sqrt{\frac{2}{\alpha}} + 2b\right) \cdot \|\boldsymbol{\lambda}(S)\|_4^2, \tag{A.4}$$

  then

$$\mathbb{P}_{f_0}\big(\check{T} \leq \check{t}_\alpha\big) \leq \frac{1}{b^2} + \frac{16}{b\|\boldsymbol{\lambda}(S)\|_4^2}. \tag{A.5}$$

Proof of Lemma 3. The expectation $\mathbb{E}_{f_0}[\check{f}] = Sf_0$, and by the triangle inequality,

$$\begin{aligned}\frac{1}{n}\|\check{f} - f_0\|_2^2 &\le \frac{2}{n}\Big(\|\check{f} - \mathbb{E}_{f_0}[\check{f}]\|_2^2 + \|\mathbb{E}_{f_0}[\check{f}] - f_0\|_2^2\Big) \\ &= \frac{2}{n}\Big(\|S\varepsilon\|_2^2 + \|(S - I)f_0\|_2^2\Big).\end{aligned}$$

Writing $\|S\varepsilon\|_2^2 = \sum_{k=1}^n \lambda_k(S)^2 Z_k^2$, the claim follows from the result of Laurent and Massart [2000] on concentration of χ^2-random variables, which for completeness we restate in Lemma 16. To be explicit, taking $t = \|\boldsymbol{\lambda}(S)\|_2^2$ in Lemma 16 completes the proof of Lemma 3.

Proof of Lemma 4. We compute the mean and variance of T as a function of f_0, then apply Chebyshev's inequality.

Mean. We make use of the eigendecomposition $S = \sum_{k=1}^n \lambda_k(S) \cdot v_k(S) v_k(S)^\top$ to obtain

$$\begin{aligned}\check{T} &= f_0^\top S^2 f_0 + 2 f_0^\top S^2 \varepsilon + \varepsilon^\top S^2 \varepsilon \\ &= f_0^\top S^2 f_0 + 2 f_0^\top S^2 \varepsilon + \sum_{k=1}^n \big(\lambda_k(S)\big)^2 (\varepsilon^\top v_k(S))^2 \\ &= f_0^\top S^2 f_0 + 2 f_0^\top S^2 \varepsilon + \sum_{k=1}^n \big(\lambda_k(S)\big)^2 Z_k^2,\end{aligned} \tag{A.6}$$

implying

$$\mathbb{E}_{f_0}\big[\check{T}\big] = f_0^\top S^2 f_0 + \sum_{k=1}^n \big(\lambda_k(S)\big)^2. \tag{A.7}$$

Variance. We start from (A.6). Recalling that $\mathrm{Var}(Z_k^2) = 2$, it follows from the Cauchy-Schwarz inequality that

$$\mathrm{Var}_{f_0}\big[\check{T}\big] \le 8 f_0^\top S^4 f_0 + 4 \sum_{k=1}^n \big(\lambda_k(S)\big)^4. \tag{A.8}$$

Bounding Type I and Type II error. The upper bound (A.3) on Type I error follows immediately from (A.7), (A.8), and Chebyshev's inequality.

We now establish the upper bound (A.5) on Type II error. From assumption (A.4), we see that $f_0^\top S^2 f_0^\top - \check{t}_\alpha \le 0$. As a result,

$$\begin{aligned}\mathbb{P}_{f_0}\Big(\check{T} \le \check{t}_\alpha\Big) &= \mathbb{P}_{f_0}\Big(\check{T} - \mathbb{E}_{f_0}\big[\check{T}\big] \le \check{t}_\alpha - \mathbb{E}_{f_0}\big[\check{T}\big]\Big) \\ &\le \mathbb{P}_{f_0}\Big(\Big|\check{T} - \mathbb{E}_{f_0}\big[\check{T}\big]\Big| \ge \Big|\check{t}_\alpha - \mathbb{E}_{f_0}\big[\check{T}\big]\Big|\Big) \\ &\le \frac{\mathrm{Var}_{f_0}\big[\check{T}\big]}{\big(\check{t}_\alpha - \mathbb{E}_{f_0}\big[\check{T}\big]\big)^2},\end{aligned}$$

where the last line follows from Chebyshev's inequality. Plugging in the expressions (A.7) and (A.8) for the mean and variance of $\check{T}$, as well as the definition of $\check{t}_\alpha$ in (A.2), we obtain that

$$\mathbb{P}_{f_0}\Big(\check{T} \le \check{t}_\alpha\Big) \le \frac{4\|\boldsymbol{\lambda}(S)\|_4^4}{\big(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\big)^2} + \frac{8 f_0^\top S^4 f_0}{\big(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\big)^2}. \tag{A.9}$$

We now use the assumed lower bound $f_0^\top S^2 f_0 \ge (2\sqrt{2/\alpha} + 2b)\|\boldsymbol{\lambda}(S)\|_4^2$ to separately upper bound each of the two terms on the right hand side of (A.9). It follows immediately that

$$\frac{4\|\boldsymbol{\lambda}(S)\|_4^4}{\big(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\big)^2} \le \frac{1}{b^2}, \tag{A.10}$$

giving a sufficient upper bound on the first term. Now we upper bound the second term,

$$\frac{8f_0^\top S^4 f_0}{\bigl(f_0^\top S^2 f_0 - \sqrt{2/\alpha}\|\boldsymbol{\lambda}(S)\|_4^2\bigr)^2} \leq \frac{32 f_0^\top S^4 f_0}{\bigl(f_0^\top S^2 f_0\bigr)^2} \leq \frac{16}{b\|\boldsymbol{\lambda}(S)\|_4^2}\frac{f_0^\top S^4 f_0}{f_0^\top S^2 f_0} \leq \frac{16}{b\|\boldsymbol{\lambda}(S)\|_4^2}, \tag{A.11}$$

where the final inequality is satisfied because S is a contraction. Plugging (A.10) and (A.11) back into (A.9) then gives the desired result.

A2.2 Analysis of Laplacian smoothing

Upper bounds on the mean squared error of $\widehat{f}$, and Type I and Type II error of $\widehat{T}$, follow from setting $S = (\rho L + I)^{-1}$ in Lemmas 3 and 4. We give these results in Lemma 5 and 6, and prove them immediately. Recall that $\lambda_1, \dots, \lambda_n$ are the n eigenvalues of L (sorted in ascending order).

Lemma 5. *For any $\rho > 0$,*

$$\frac{1}{n}\bigl\|\widehat{f} - f_0\bigr\|_2^2 \leq \frac{2\rho}{n}\bigl(f_0^\top L f_0\bigr) + \frac{10}{n}\sum_{k=1}^{n}\frac{1}{\bigl(\rho\lambda_k + 1\bigr)^2}, \tag{A.12}$$

with probability at least $1 - \exp\Bigl(-\sum_{k=1}^n \bigl(\rho\lambda_k + 1\bigr)^{-2}\Bigr)$.

Recall that

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

Lemma 6. *For any $\rho > 0$ and any $b \geq 1$, it holds that:*

- ***Type I error.***

$$\mathbb{P}_0\Bigl(\widehat{T} > \widehat{t}_\alpha\Bigr) \leq \alpha. \tag{A.13}$$

- ***Type II error.*** *If*

$$\frac{1}{n}\|f_0\|_2^2 \geq \frac{2\rho}{n}\bigl(f_0^\top L f_0\bigr) + \frac{2\sqrt{2/\alpha} + 2b}{n}\left(\sum_{k=1}^{n}\frac{1}{(\rho\lambda_k + 1)^4}\right)^{1/2}, \tag{A.14}$$

  then

$$\mathbb{P}_{f_0}\Bigl(\widehat{T}(G) \leq \widehat{t}_\alpha\Bigr) \leq \frac{1}{b^2} + \frac{16}{b}\left(\sum_{k=1}^{n}\frac{1}{(\rho\lambda_k + 1)^4}\right)^{-1/2}. \tag{A.15}$$

Proof of Lemma 5. Let $\widehat{S} = (I + \rho L)^{-1}$, the estimator $\widehat{f} = \widehat{S}Y$, and

$$\bigl\|\boldsymbol{\lambda}(\widehat{S})\bigr\|_2^2 = \sum_{k=1}^{n}\frac{1}{\bigl(1 + \rho\lambda_k\bigr)^2}.$$

We deduce the following upper bound on the bias term,

$$\begin{aligned}
\bigl\|(\widehat{S} - I)f_0\bigr\|_2^2 &= f_0^\top L^{1/2} L^{-1/2}\bigl(\widehat{S} - I\bigr)^2 L^{-1/2} L^{1/2} f_0 \\
&\leq f_0^\top L f_0 \cdot \lambda_n\Bigl(L^{-1/2}\bigl(\widehat{S} - I\bigr)^2 L^{-1/2}\Bigr) \\
&= f_0^\top L f_0 \cdot \max_{k\in[n]}\left\{\frac{1}{\lambda_k}\Bigl(1 - \frac{1}{\rho\lambda_k + 1}\Bigr)^2\right\} \\
&\leq f_0^\top L f_0 \cdot \rho.
\end{aligned}$$

In the above, we have written $L^{-1/2}$ for the square root of the pseudoinverse of L, the maximum is over all indices k such that $\lambda_k > 0$, and the last inequality follows from the basic algebraic identity $1 - 1/(1+x)^2 \leq 2x$ for any $x > 0$. The claim of the Lemma then follows from Lemma 3.

Proof of Lemma 6. Let $\widehat{S} := (I + \rho L)^{-1}$, so that $\widehat{T} = \frac{1}{n}\mathbf{Y}^\top \widehat{S}^2 \mathbf{Y}$. Note that $\widehat{S}$ is a contraction, so that we may invoke Lemma 4. The bound on Type I error (A.13) follows immediately from (A.3). To establish the bound on Type II error, we must lower bound $f_0^\top \widehat{S}^2 f_0$. We first note that by assumption (A.14),

$$\begin{aligned}f_0^\top \widehat{S}^2 f_0 &= \|f_0\|_2^2 - f_0^\top (I - \widehat{S}^2) f_0 \\ &\geq 2\rho\big(f_0^\top L f_0\big) - f_0^\top\big(I - \widehat{S}^2\big) f_0 + \Big(2\sqrt{\frac{2}{\alpha}} + 2b\Big) \cdot \left(\sum_{k=1}^{n} \frac{1}{(\rho\lambda_k + 1)^4}\right)^{-1/2}.\end{aligned}$$

Upper bounding $f_0^\top\big(I - \widehat{S}^2\big) f_0$ as follows:

$$\begin{aligned}f_0^\top\Big(I - \widehat{S}^2\Big) f_0 &= f_0^\top L^{1/2} L^{-1/2}\Big(I - \widehat{S}^2\Big) L^{-1/2} L^{1/2} f_0 \\ &\leq f_0^\top L f_0 \cdot \lambda_n\Big(L^{-1/2}\Big(I - \widehat{S}^2\Big) L^{-1/2}\Big) \\ &= f_0^\top L f_0 \cdot \max_k \left\{\frac{1}{\lambda_k}\Big(1 - \frac{1}{(\rho\lambda_k + 1)^2}\Big)\right\} \\ &\leq f_0^\top L f_0 \cdot 2\rho,\end{aligned}$$

—where in the above the maximum is over all indices k such that $\lambda_k > 0$—we deduce that

$$f_0^\top \widehat{S}^2 f_0 \geq \Big(2\sqrt{\frac{2}{\alpha}} + 2b\Big) \cdot \left(\sum_{k=1}^{n} \frac{1}{(\rho\lambda_k + 1)^4}\right)^{-1/2}.$$

The upper bound on Type II error (A.15) then follows from Lemma 4.

A3 Neighborhood graph Sobolev semi-norm

In this section, we prove Lemma 1, which states an upper bound on $f^\top L f$ that holds when f is bounded in Sobolev norm. We also establish stronger bounds in the case when f has a bounded Lipschitz constant; this latter result justifies one of our remarks after Theorem 1.

Throughout this proof, we will assume that $f \in H^1(\mathcal{X})$ has zero-mean, meaning $\int_{\mathcal{X}} f(x)\,dx = 0$. This is without loss of generality—assuming for the moment that (14) holds for zero-mean functions, for any $f \in H^1(\mathcal{X})$, taking $a = \int_{\mathcal{X}} f(x)\,dx$ and $g = f - a$, we have that

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

Now, for any zero-mean function $f \in H^1(\mathcal{X})$ it follows by the Poincare inequality (see Section 5.8, Theorem 1 of Evans [2010]) that $\|f\|_{H^1(\mathcal{X})}^2 \leq C_8 |f|_{H^1(\mathcal{X})}^2$, for some constant C_8 that does not depend on f. Therefore, to prove Lemma 1, it suffices to show that

$$\mathbb{E}\Big[f^\top L f\Big] \leq C n^2 r^{d+2} \|f\|_{H^1(\mathcal{X})}^2,$$

since the high-probability upper bound then follows immediately by Markov's inequality. (Recall that L is positive semi-definite, and therefore $f^\top L f$ is a non-negative random variable).

Since

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

it follows that

$$\mathbb{E}\Big[f^\top L f\Big] = \frac{n(n-1)}{2} \mathbb{E}\bigg[\Big(f(X') - f(X)\Big)^2 K\Big(\frac{\|X' - X\|}{r}\Big)\bigg], \tag{A.16}$$

where X and X' are random variables independently drawn from P.

Now, take Ω to be an arbitrary bounded open set such that $B(x, c_0) \subseteq \Omega$ for all $x \in \mathcal{X}$. For the remainder of this proof, we will assume that (i) $f \in H^1(\Omega)$ and additionally (ii) $\|f\|_{H^1(\Omega)} \leq C_5 \|f\|_{H^1(\mathcal{X})}$ for a constant C_5

that does not depend on f. This is without loss of generality, since by Theorem 1 in Chapter 5.4 of Evans [2010] there exists an extension operator $E : H^1(\mathcal{X}) \to H^1(\Omega)$ for which the extension Ef satisfies both (i) and (ii). Additionally, we will assume $f \in C^\infty(\Omega)$. Again, this is without loss of generality, as $C^\infty(\Omega)$ is dense in $H^1(\Omega)$ and the expectation on the right hand side of (A.16) is continuous in $H^1(\Omega)$. The reason for dealing with a smooth extension $f \in C^\infty(\Omega)$ is so that we can make sense of the following equality for any x and x' in $\mathcal{X}$:

$$f(x') - f(x) = \int_0^1 \nabla f\big(x + t(x' - x)\big)^\top (x' - x)\, dt. \tag{A.17}$$

Obviously

$$\mathbb{E}\Big[\big(f(X') - f(X)\big)^2 K\Big(\frac{\|X' - X\|}{r}\Big)\Big] \le p_{\max}^2 \int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 K\Big(\frac{\|x' - x\|}{r}\Big)\, dx'\, dx, \tag{A.18}$$

so that it remains now to bound the double integral. Replacing difference by integrated derivative as in (A.17), we obtain

$$\begin{aligned}
\int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 K\Big(\frac{\|x' - x\|}{r}\Big)\, dx'\, dx &= \int_{\mathcal{X}}\int_{\mathcal{X}} \Big[\int_0^1 \nabla f\big(x + t(x' - x)\big)^\top (x' - x)\, dt\Big]^2 K\Big(\frac{\|x' - x\|}{r}\Big)\, dx'\, dx \\
&\overset{(i)}{\le} \int_{\mathcal{X}}\int_{\mathcal{X}}\int_0^1 \Big[\nabla f\big(x + t(x' - x)\big)^\top (x' - x)\Big]^2 K\Big(\frac{\|x' - x\|}{r}\Big)\, dt\, dx'\, dx \\
&\overset{(ii)}{\le} r^{d+2} \int_{\mathcal{X}}\int_{B(0,1)}\int_0^1 \Big[\nabla f\big(x + trz\big)^\top z\Big]^2 K\big(\|z\|\big)\, dt\, dz\, dx \\
&\overset{(iii)}{\le} r^{d+2} \int_{\Omega}\int_{B(0,1)}\int_0^1 \Big[\nabla f\big(\widetilde{x}\big)^\top z\Big]^2 K\big(\|z\|\big)\, dt\, dz\, d\widetilde{x},
\end{aligned} \tag{A.19}$$

where (i) follows by Jensen's inequality, (ii) follows by substituting $z = (x' - x)/r$ and (K1), and (iii) by exchanging integrals, substituting $\widetilde{x} = x + trz$, and noting that $x \in \mathcal{X}$ implies that $\widetilde{x} \in \Omega$.

Now, writing $\big(\nabla f(\widetilde{x})^\top z\big)^2 = \big(\sum_{i=1}^d z_i f^{(e_i)}(x)\big)^2$, expanding the square and integrating, we have that for any $\widetilde{x} \in \mathcal{X}$,

$$\begin{aligned}
\int_{B(0,1)} \Big[\nabla f\big(\widetilde{x}\big)^\top z\Big]^2 K\big(\|z\|\big)\, dz &= \sum_{i,j=1}^d f^{(e_i)}(\widetilde{x}) f^{(e_j)}(\widetilde{x}) \int_{\mathbb{R}^d} z_i z_j K\big(\|z\|\big)\, dz \\
&= \sum_{i=1}^d \big(f^{(e_i)}(\widetilde{x})\big)^2 \int_{B(0,1)} z_i^2 K\big(\|z\|\big)\, dz \\
&= \sigma_K \|\nabla f(\widetilde{x})\|^2,
\end{aligned}$$

where the last equality follows from the rotational symmetry of $K(\|z\|)$. Plugging back into (A.19), we obtain

$$\int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 K\Big(\frac{\|x' - x\|}{r}\Big)\, dx'\, dx \le r^{d+2}\sigma_K \|f\|_{H^1(\Omega)}^2 \le C_5 r^{d+2}\sigma_K \|f\|_{H^1(\mathcal{X})}^2,$$

proving the claim of Lemma 1 upon taking $C_2 := C_8 C_5 \sigma_K p_{\max}^2$ in the statement of the lemma.

A3.1 Stronger bounds under Lipschitz assumption

Suppose f satisfies $|f(x') - f(x)| \le M\|x - x\|$ for all $x, x' \in \mathcal{X}$. Then we can strengthen the high probability bound in Lemma 1 from $1 - \delta$ to $1 - \delta^2/n$, at the cost of only a constant factor in the upper bound on $f^\top L f$.

Proposition 1. *Let $r \ge C_0(\log n/n)^{1/d}$. For any f such that $|f(x') - f(x)| \le M\|x - x\|$ for all $x, x' \in \mathcal{X}$, with probability at least $1 - C\delta^2/n$ it holds that*

$$f^\top L f \le \Big(\frac{1}{\delta} + C_2\Big) n^2 r^{d+2} M^2.$$

Proof of Proposition 1. We will prove Proposition 1 using Chebyshev's inequality, so the key step is to upper bound the variance of $f^\top L f$. Putting $\Delta_{ij} := K(\|X_i - X_j\|/r) \cdot (f(X_i) - f(X_j))^2$, we can write the variance of $f^\top L f$ as a sum of covariances,

$$\mathrm{Var}\bigl[f^\top L f\bigr] = \frac{1}{4}\sum_{i,j=1}^{n}\sum_{\ell,m=1}^{n}\mathrm{Cov}\bigl[\Delta_{ij}, \Delta_{\ell m}\bigr].$$

Clearly $\mathrm{Cov}\bigl[\Delta_{ij}, \Delta_{\ell m}\bigr]$ depends on the cardinality of $I := \{i,j,k,\ell\}$; we divide into cases, and upper bound the covariance in each case.

$\bigl|I\bigr| = 4.$ In this case Δ_{ij} and $\Delta_{\ell m}$ are independent, and $\mathrm{Cov}\bigl[\Delta_{ij}, \Delta_{\ell m}\bigr] = 0$.

$\bigl|I\bigr| = 3.$ Taking $i = \ell$ without loss of generality, and noting that the expectation of Δ_{ij} and Δ_{im} is non-negative, we have

$$\begin{aligned}
\mathrm{Cov}\bigl[\Delta_{ij}, \Delta_{im}\bigr] &\leq \mathbb{E}\bigl[\Delta_{ij}\Delta_{im}\bigr] \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} \bigl(f(z) - f(x)\bigr)^2\bigl(f(x') - f(x)\bigr)^2 K\biggl(\frac{\|x' - x\|}{r}\biggr) K\biggl(\frac{\|z - x\|}{r}\biggr) p(z)p(x')p(x)\,dz\,dx'\,dx \\
&\leq p_{\max}^3 M^4 r^4 \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} K\biggl(\frac{\|x' - x\|}{r}\biggr) K\biggl(\frac{\|z - x\|}{r}\biggr)\,dz\,dx'\,dx \\
&\leq p_{\max}^3 M^4 r^{4+2d}.
\end{aligned}$$

$\bigl|I\bigr| = 2.$ Taking $i = \ell$ and $j = m$ without loss of generality, we have

$$\begin{aligned}
\mathrm{Var}\bigl[\Delta_{ij}\bigr] &\leq \mathbb{E}\bigl[\Delta_{ij}^2\bigr] \\
&\leq \int_{\mathcal{X}}\int_{\mathcal{X}} \bigl(f(x') - f(x)\bigr)^4 \biggl[K\biggl(\frac{\|x' - x\|}{r}\biggr)\biggr]^2 p(x')p(x)\,dx'\,dx \\
&\leq p_{\max}^2 M^4 r^4 K(0) \int_{\mathcal{X}}\int_{\mathcal{X}} K\biggl(\frac{\|x' - x\|}{r}\biggr)\,dx'\,dx \\
&\leq p_{\max}^2 M^4 r^{4+d} K(0).
\end{aligned}$$

$\bigl|I\bigr| = 1.$ In this case $\Delta_{ij} = \Delta_{\ell m} = 0$.

Therefore

$$\mathrm{Var}\bigl[f^\top L f\bigr] \leq n^3 p_{\max}^3 M^4 r^{4+2d} + n^2 p_{\max}^2 M^4 r^{4+d} K(0) \leq C M^4 n^3 r^{4+2d},$$

where the latter inequality follows since $nr^d \gg 1$. For any $\delta > 0$, it follows from Chebyshev's inequality that

$$\mathbb{P}\biggl(\Bigl|f^\top L f - \mathbb{E}\bigl[f^\top L f\bigr]\Bigr| \geq \frac{M^2}{\delta} n^2 r^{d+2}\biggr) \leq C\frac{\delta^2}{n},$$

and since we have already upper bounded $\mathbb{E}\bigl[f^\top L f\bigr] \leq C_2 M^2 n^2 r^{d+2}$, the proposition follows.

Note that the bound on $\mathrm{Var}[\Delta_{ij}]$ follows as long as we can control $\|\nabla f\|_{L^4(\mathcal{X})}$; this implies the Lipschitz assumption—which gives us control of $\|\nabla f\|_{L^\infty(\mathcal{X})}$—can be weakened. However, the Sobolev assumption—which gives us control only over $\|\nabla f\|_{L^2(\mathcal{X})}$—will not do the job.

A4 Bounds on neighborhood graph eigenvalues

In this section, we prove Lemma 2, following the lead of Burago et al. [2014], García Trillos et al. [2019], Calder and García Trillos [2019], who establish similar results with respect to a manifold without boundary. To prove this lemma, in Theorem 5 we give estimates on the difference between eigenvalues of the graph Laplacian L and eigenvalues of the weighted Laplace-Beltrami operator Δ_P. We recall Δ_P is defined as

$$\Delta_P f(x) := -\frac{1}{p(x)}\mathrm{div}\bigl(p^2 \nabla f\bigr)(x).$$

To avoid confusion, in this section we write $\lambda_k(G_{n,r})$ for the kth smallest eigenvalue of the graph Laplacian matrix L and $\lambda_k(\Delta_P)$ for the kth smallest eigenvalue of Δ_P [1]. Some other notation: throughout this section, we will write $A, A_0, A_1, \ldots$ and $a, a_0, a_1, \ldots$ for constants which may depend on $\mathcal{X}$, d, K, and p, but do not depend on n; we keep track of all such constants explicitly in our proofs. We let L_K denote the Lipschitz constant of the kernel K. Finally, for notational ease we set θ and $\widetilde{\delta}$ to be the following (small) positive numbers:

$$\widetilde{\delta} := \max\left\{n^{-1/d}, \min\left\{\frac{1}{2^{d+3}A_0}, \frac{1}{A_3}, \frac{K(1)}{8L_K A_0}, \frac{1}{8\max\{A_1, A\}c_0}\right\}r\right\}, \quad \text{and} \quad \theta := \frac{1}{8\max\{A_1, A\}}. \tag{A.20}$$

We note that each of $\widetilde{\delta}, \theta$ and $\widetilde{\delta}/r$ are of at most constant order.

Theorem 5. *For any $\ell \in \mathbb{N}$ such that*

$$1 - A\left(r\sqrt{\lambda_\ell(\Delta_P)} + \theta + \widetilde{\delta}\right) \geq \frac{1}{2} \tag{A.21}$$

with probability at least $1 - A_0 n \exp(-a_0 n\theta^2\widetilde{\delta}^d)$, it holds that

$$a\lambda_k(G_{n,r}) \leq nr^{d+2}\lambda_k(\Delta_P) \leq A\lambda_k(G_{n,r}), \quad \textit{for all } 1 \leq k \leq \ell \tag{A.22}$$

Before moving forward to the proofs of Lemma 2 and Theorem 5, it is worth being clear about the differences between Theorem 5 and the results of Burago et al. [2014], García Trillos et al. [2019], Calder and García Trillos [2019]. First of all, the reason we cannot directly use the results of these works in the proof of Lemma 2 is that they all assume the domain $\mathcal{X}$ is without boundary, whereas for our results in Section 4 we instead assume $\mathcal{X}$ has a (Lipschitz smooth) boundary. Fortunately, in this setting the high-level strategy shared by Burago et al. [2014], García Trillos et al. [2019], Calder and García Trillos [2019] can still be used—indeed we follow it closely, as we summarize in Section A4.1. However, many calculations need to be redone, in order to account for points x which are on or sufficiently close to the boundary of $\mathcal{X}$. For completeness and ease of reading, we provide a self-contained proof of Theorem 5, but we comment where appropriate on connections between the technical results we use in this proof, and those derived in Burago et al. [2014], García Trillos et al. [2019], Calder and García Trillos [2019].

On the other hand, we should also point out that unlike the results of Burago et al. [2014], García Trillos et al. [2019], Calder and García Trillos [2019], Theorem 5 does not imply that $\lambda_k(G_{n,r})$ is a consistent estimate of $\lambda_k(\Delta_P)$, i.e. it does not imply that $|(nr^{d+2})^{-1}\lambda_k(G_{n,r}) - \lambda_k(\Delta_P)| \to 0$ as $n \to \infty, r \to 0$. The key difficulty in proving consistency when $\mathcal{X}$ has a boundary can be summarized as follows: while at points $x \in \mathcal{X}$ satisfying $B(x,r) \subseteq \mathcal{X}$, the graph Laplacian L is a reasonable approximation of the operator Δ_P, at points x near the boundary L is known to approximate a different operator altogether [Belkin et al., 2012]. This is reminiscent of the boundary effects present in the analysis of kernel smoothing. We believe a more subtle analysis might imply convergence of eigenvalues in this setting. However, the conclusion of Theorem 5—that $\lambda_k(G_{n,r})/(nr^{d+2}\lambda_k(\Delta_P))$ is bounded above and below by constants that do not depend on k—suffices for our purposes.

The bulk of the remainder of this section is devoted to the proof of Theorem 5. First, however, we show that under our regularity conditions on p and $\mathcal{X}$, Lemma 2 is a simple consequence of Theorem 5. The link between the two is Weyl's Law.

Proposition 2 (Weyl's Law)**.** *Suppose the density p and the domain $\mathcal{X}$ satisfy (P1) and (P2). Then there exist constants a_2 and A_2 such that*

$$a_2 k^{2/d} \leq \lambda_k(\Delta_P) \leq A_2 k^{2/d} \quad \textit{for all } k \in \mathbb{N}, k > 1. \tag{A.23}$$

See Lemma 28 of Dunlop et al. [2020] for a proof that (P1) and (P2) imply Weyl's Law.

[1] Under the assumptions (P1) and (P2), the operator Δ_P has a discrete spectrum; see García Trillos and Slepčev [2018] for more details.

Proof of Lemma 2. Put

$$\ell_\star = \left\lfloor \left(\frac{\big(1/(2A) - (\theta + \widetilde{\delta})\big)}{rA_2^{1/2}} \right)^d \right\rfloor.$$

Let us verify that $\lambda_{\ell_\star}(\Delta_P)$ satisfies the condition (A.21) of Theorem 5. Setting $c_0 := 1/(2^{1/d}4A_2^{1/2})$, the assumed upper bound on the radius $r \leq c_0$ guarantees that $\ell_\star \geq 2$. Therefore, by Proposition 2 we have that

$$\sqrt{\lambda_{\ell_\star}(\Delta_P)} \leq A_2^{1/2}\ell_\star^{1/d} \leq \frac{1}{r}\left(\frac{1}{2A} - (\theta + \widetilde{\delta})\right).$$

Rearranging the above inequality shows that condition (A.21) is satisfied.

It is therefore the case that the inequalities in (A.22) hold with probability at least $1 - A_0 n\exp(-a_0 n\theta^2\widetilde{\delta}^d)$. Together, (A.22) and (A.23) imply the following bounds on the graph Laplacian eigenvalues:

$$\frac{a}{A_2}nr^{d+2}k^{2/d} \leq \lambda_k(G_{n,r}) \leq \frac{A}{a_2}nr^{d+2}k^{2/d} \quad \text{for all } 2 \leq k \leq \ell_\star.$$

It remains to bound $\lambda_k(G_{n,r})$ for those indices k which are greater than $\ell_\star$. On the one hand, since the eigenvalues are sorted in ascending order, we can use the lower bound on $\lambda_{\ell_\star}(G_{n,r})$ that we have just derived:

$$\lambda_k(G_{n,r}) \geq \lambda_{\ell_\star}(G_{n,r}) \geq \frac{a_2}{A}nr^{d+2}\ell_\star^{2/d} \geq \frac{a_2}{64A^3A_2}nr^d.$$

On the other hand, for any graph G the maximum eigenvalue of the Laplacian is upper bounded by twice the maximum degree [Chung and Graham, 1997]. Writing $D_{\max}(G_{n,r})$ for the maximum degree of $G_{n,r}$, it is thus a consequence of Lemma 19 that

$$\lambda_k(G_{n,r}) \leq 2D_{\max}(G_{n,r}) \leq 4p_{\max}nr^d,$$

with probability at least $1 - 2n\exp\Big(-nr^dp_{\min}/(3K(0)^2)\Big)$. In sum, we have shown that with probability at least $1 - A_0n\exp(-a_0n\theta^2\widetilde{\delta}^d) - 2n\exp\Big(-nr^dp_{\min}/(3K(0)^2)\Big)$,

$$\min\left\{\frac{a_2}{A}nr^{d+2}k^{2/d}, \frac{a_2}{A^364A_3}nr^d\right\} \leq \lambda_k(G_{n,r}) \leq \min\left\{\frac{A_2}{a}nr^{2+d}k^{2/d}, 4p_{\max}nr^d\right\} \quad \text{for all } 2 \leq k \leq n.$$

Lemma 2 then follows upon setting

$$C_1 := \max\{2A_0, 4\}, \qquad c_1 := \min\left\{\frac{p_{\min}}{3K(0)^2}, \frac{\theta^2\widetilde{\delta}}{r}\right\}$$

$$C_3 := \max\left\{\frac{A_2}{a}, 4p_{\max}\right\}, \qquad c_3 := \min\left\{\frac{a_2}{A}, \frac{a_2}{A^364A_3}\right\}.$$

in the statement of that Lemma.

A4.1 Proof of Theorem 5

In this section we prove Theorem 5, following closely the approach of Burago et al. [2014], García Trillos et al. [2019], Calder and García Trillos [2019]. As in these works, we relate $\lambda_k(\Delta_P)$ and $\lambda_k(G_{n,r})$ by means of the Dirichlet energies

$$b_r(u) := \frac{1}{n^2r^{d+2}}u^\top Lu$$

and

$$D_2(f) := \begin{cases} \int_{\mathcal{X}} \|\nabla f(x)\|^2 p^2(x)\,dx & \text{if } f \in H^1(\mathcal{X}) \\ \infty & \text{otherwise,} \end{cases}$$

Let us pause briefly to motivate the relevance of $b_r(u)$ and $D_2(f)$. In the following discussion, recall that for a function $u : \{X_1, \ldots, X_n\} \to \mathbb{R}$, the empirical norm is defined as $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n (u(X_i))^2$, and the class $L^2(P_n)$ consists of those $u \in \mathbb{R}^n$ for which $\|u\|_n < \infty$. Similarly, for a function $f : \mathcal{X} \to \mathbb{R}$, the $L^2(P)$ norm of f is

$$\|f\|_P^2 := \int_{\mathcal{X}} \big|f(x)\big|^2 p(x)\,dx,$$

and the class $L^2(P)$ consists of those f for which $\|f\|_P < \infty$. Now, suppose one could show the following two results:

(1) an upper bound of $b_r(u)$ by $D_2\big(\mathcal{I}(u)\big)$ for an appropriate choice of interpolating map $\mathcal{I} : L^2(P_n) \to L^2(\mathcal{X})$, and vice versa an upper bound of $D_2(f)$ by $b_r(\mathcal{P}(f))$ for an appropriate choice of discretization map $\mathcal{P} : L^2(\mathcal{X}) \to L^2(P_n)$,

(2) that $\mathcal{I}$ and $\mathcal{P}$ were near-isometries, meaning $\|\mathcal{I}(u)\|_P \approx \|u\|_n$ and $\|\mathcal{P}(f)\|_P \approx \|f\|_n$.

Then, by using the variational characterization of eigenvalues $\lambda_k(\Delta_P)$ and $\lambda_k(G_{n,r})$—i.e. the Courant-Fischer Theorem—one could obtain estimates on the error $\big|nr^{d+2}\lambda_k(\Delta_P) - \lambda_k(G_{n,r})\big|$.

We will momentarily define particular maps $\widetilde{\mathcal{I}}$ and $\widetilde{\mathcal{P}}$, and establish that they satisfy both (1) and (2). In order to define these maps, we must first introduce a particular probability measure $\widetilde{P}_n$ that, with high probability, is close in transportation distance to both P_n and P. This estimate on the transportation distance—which we now give—will be the workhorse that allows us to relate b_r to D_2, and $\|\cdot\|_n$ to $\|\cdot\|_P$.

Transportation distance between P_n and P. For a measure μ defined on $\mathcal{X}$ and map $T : \mathcal{X} \to \mathcal{X}$, let $T_\sharp\mu$ denote the *push-forward* of μ by T, i.e the measure for which

$$\big(T_\sharp\mu\big)(U) := \mu\big(T^{-1}(U)\big)$$

for any Borel subset $U \subseteq \mathcal{X}$. Suppose $T_\sharp\mu = P_n$; then the map T is referred to as transportation map between μ and P_n. The ∞-transportation distance between μ and P_n is then

$$d_\infty(\mu, P_n) := \inf_{T : T_\sharp\mu = P_n} \|T - \mathrm{Id}\|_{L^\infty(\mu)} \tag{A.24}$$

where $\mathrm{Id}(x) = x$ is the identity mapping.

Calder and García Trillos [2019] take $\mathcal{X}$ to be a smooth submanifold of $\mathbb{R}^d$ without boundary, i.e. they assume $\mathcal{X}$ satisfies (P3). In this setting, they exhibit an absolutely continuous measure $\widetilde{P}_n$ with density $\widetilde{p}_n$ that with high probability is close to P_n in transportation distance, and for which $\|p - \widetilde{p}_n\|_{L^\infty}$ is also small. In Proposition 3, we adapt this result to the setting of full-dimensional manifolds with boundary.

Proposition 3. *Suppose $\mathcal{X}$ satisfies (P1) and p satisfies (P2). Then with probability at least $1 - A_0 n \exp\{-a_0 n\theta^2\widetilde{\delta}^d\}$, the following statement holds: there exists a probability measure $\widetilde{P}_n$ with density $\widetilde{p}_n$ such that:*

$$d_\infty(\widetilde{P}_n, P_n) \leq A_0\widetilde{\delta} \tag{A.25}$$

and

$$\|\widetilde{p}_n - p\|_\infty \leq A_0\big(\widetilde{\delta} + \theta\big) \tag{A.26}$$

For the rest of this section, we let $\widetilde{P}_n$ be a probability measure with density $\widetilde{p}_n$, that satisfies the conclusions of Proposition 3. Additionally we denote by $\widetilde{T}_n$ an *optimal transport map* between $\widetilde{P}_n$ and P_n, meaning a transportation map which achieves the infimum in (A.24). Finally, we write $U_1, \ldots, U_n$ for the preimages of $X_1, \ldots, X_n$ under $\widetilde{T}_n$, meaning $U_i = \widetilde{T}_n^{-1}(X_i)$.

Interpolation and discretization maps. The discretization map $\widetilde{\mathcal{P}} : L^2(\mathcal{X}) \to L^2(P_n)$ is given by averaging over the cells $U_1, \ldots, U_n$,

$$\big(\widetilde{\mathcal{P}}f\big)(X_i) := n \cdot \int_{U_i} f(x)\widetilde{p}_n(x)\,dx.$$

On the other hand, the interpolation map $\widetilde{\mathcal{I}} : L^2(P_n) \to L^2(\mathcal{X})$ is defined as $\widetilde{\mathcal{I}}u := \Lambda_{r-2A_0\widetilde{\delta}}(\widetilde{\mathcal{P}}^\star u)$. Here, $\widetilde{\mathcal{P}}^\star = u \circ \widetilde{T}$ is the adjoint of $\widetilde{\mathcal{P}}$, i.e.

$$\big(\widetilde{\mathcal{P}}^\star u\big)(x) = \sum_{j=1}^{n} u(x_i)\mathbf{1}\{x \in U_i\}$$

and $\Lambda_{r-2A_0\widetilde{\delta}}$ is a kernel smoothing operator, defined with respect to a carefully chosen kernel ψ. To be precise, for any $h > 0$,

$$\Lambda_h(f) := \frac{1}{h^d\tau_h(x)}\int_{\mathcal{X}} \eta_h(x', x)f(x')\,dx', \quad \eta_h(x', x) := \psi\bigg(\frac{\|x' - x\|}{r}\bigg)$$

where $\psi(t) := (1/\sigma_K) \int_t^\infty sK(s)\,ds$ and $\tau_h(x) := (1/h^d) \int_{\mathcal{X}} \eta_h(x',x)\,dx'$ is a normalizing constant.

Propositions 4 and 5 establish our claims regarding $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{I}}$: first, that they approximately preserve the Dirichlet energies b_r and D_2, and second that they are near-isometries for functions $u \in L^2(P_n)$ (or $f \in L^2(P)$) of small Dirichlet energy $b_r(u)$ (or $D_2(f)$).

Proposition 4 (cf. Proposition 4.1 of Calder and García Trillos [2019]). *With probability at least $1 - A_0 n \exp(-a_0 n\theta^2\widetilde{\delta}^d)$, we have the following.*

(1) For every $u \in L^2(P_n)$,

$$\sigma_K D_2(\widetilde{\mathcal{I}}u) \leq A_8\Big(1 + A_1(\theta + \widetilde{\delta})\Big)\cdot\Big(1 + A_3\frac{\widetilde{\delta}}{r}\Big)b_r(u). \tag{A.27}$$

(2) For every $f \in L^2(\mathcal{X})$,

$$b_r(\widetilde{\mathcal{P}}f) \leq \Big(1 + A_1(\theta + \widetilde{\delta})\Big)\cdot\Big(1 + A_9\frac{\widetilde{\delta}}{r}\Big)\cdot\Big(\frac{C_5 p_{\max}^2}{p_{\min}^2}\Big)\cdot\sigma_K D_2(f). \tag{A.28}$$

Proposition 5 (cf. Proposition 4.2 of Calder and García Trillos [2019]). *With probability at least $1 - A_0 n \exp(-a_0 n\theta^2\widetilde{\delta}^d)$, we have the following.*

(1) For every $f \in L^2(\mathcal{X})$,

$$\Big|\|f\|_P^2 - \|\widetilde{\mathcal{P}}f\|_n^2\Big| \leq A_5 r\|f\|_P\sqrt{D_2(f)} + A_1\big(\theta + \widetilde{\delta}\big)\|f\|_P^2. \tag{A.29}$$

(2) For every $u \in L^2(P_n)$,

$$\Big|\|\widetilde{\mathcal{I}}u\|_P^2 - \|u\|_n^2\Big| \leq A_6 r\|u\|_n\sqrt{b_r(u)} + A_7\big(\theta + \widetilde{\delta}\big)\|u\|_n^2. \tag{A.30}$$

We will devote most of the rest of this section to the proofs of Propositions 3, 4, and 5. First, however, we use these propositions to prove Theorem 5.

Proof of Theorem 5. Throughout this proof, we assume that inequalities (A.27)-(A.30) are satisfied. We take A and a to be positive constants such that

$$\frac{1}{a} \geq 2\Big(1 + A_1(\theta + \widetilde{\delta})\Big)\Big(1 + A_9\frac{\widetilde{\delta}}{r}\Big)\Big(\frac{C_5 p_{\max}^2}{p_{\min}^2}\Big), \quad \text{and} \quad A \geq \max\left\{A_1, A_5, \frac{1}{\sqrt{a}}A_6, A_7, 2A_8\Big(1 + A_1(\theta + \widetilde{\delta})\Big)\Big(1 + A_3\frac{\widetilde{\delta}}{r}\Big)\right\}.$$

Let k be any number in $1,\ldots,\ell$. We start with the upper bound in (A.22), proceeding as in Proposition 4.4 of Burago et al. [2014]. Let $f_1,\ldots,f_k$ denote the first k eigenfunctions of Δ_P and set $W := \mathrm{span}\{f_1,\ldots,f_k\}$, so that by the Courant-Fischer principle $D_2(f) \leq \lambda_k(\Delta_P)\|f\|_P^2$ for every $f \in W$. As a result, by Part (1) of Proposition 5 we have that for any $f \in W$,

$$\big\|\widetilde{\mathcal{P}}f\big\|_n^2 \geq \Big(1 - A_5 r\sqrt{\lambda_k(\Delta_P)} - A_1(\theta + \widetilde{\delta})\Big)\|f\|_P^2 \geq \frac{1}{2}\|f\|_P^2,$$

where the second inequality follows by assumption (A.21).

Therefore $\widetilde{\mathcal{P}}$ is injective over W, and $\widetilde{\mathcal{P}}W$ has dimension ℓ. This means we can invoke the Courant-Fischer Theorem, along with Proposition 4, and conclude that

$$\begin{aligned}
\frac{\lambda_k(G_{n,r})}{nr^{d+2}} &\leq \max_{\substack{u \in \widetilde{\mathcal{P}}W \\ u \neq 0}} \frac{b_r(u)}{\|u\|_n^2} \\
&= \max_{\substack{f \in W \\ f \neq 0}} \frac{b_r(\widetilde{\mathcal{P}}f)}{\big\|\widetilde{\mathcal{P}}f\big\|_n^2} \\
&\leq 2\Big(1 + A_1(\theta + \widetilde{\delta})\Big)\cdot\Big(1 + A_9\frac{\widetilde{\delta}}{r}\Big)\cdot\Big(\frac{C_5 p_{\max}^2}{p_{\min}^2}\Big)\sigma_K\lambda_k(\Delta_P),
\end{aligned}$$

establishing the lower bound in (A.22).

The upper bound follows from essentially parallel reasoning. Recalling that $v_1, \ldots, v_k$ denote the first k eigenvectors of L, set $U := \mathrm{span}\{v_1, \ldots, v_k\}$, so that $nr^{d+2} b_r(u) \leq \lambda_k(G_{n,r}) \|u\|_n^2$. By Proposition 5, Part (2), we have that for every $u \in U$,

$$\begin{aligned}
\|\widetilde{\mathcal{I}}u\|_P^2 &\geq \|u\|_n^2 - A_6 r \|u\|_n \sqrt{b_r(u)} - A_7\big(\theta + \widetilde{\delta}\big)\|u\|_n^2 \\
&\geq \|u\|_n^2 - A_6 r \|u\|_n^2 \sqrt{\frac{\lambda_k(G_{n,r})}{nr^{d+2}}} - A_7\big(\theta + \widetilde{\delta}\big)\|u\|_n^2 \\
&\geq \|u\|_n^2 - A_6 r \|u\|_n^2 \sqrt{\frac{1}{a}\lambda_k(\Delta_P)} - A_7\big(\theta + \widetilde{\delta}\big)\|u\|_n^2 \\
&\geq \frac{1}{2}\|u\|_n^2,
\end{aligned}$$

where the second to last inequality follows from the lower bound $a\lambda_k(G_{n,r}) \leq nr^{d+2}\lambda_k(\Delta_P)$ that we just derived, and the last inequality from assumption (A.21).

Therefore $\widetilde{\mathcal{I}}$ is injective over U, $\widetilde{\mathcal{I}}U$ has dimension k, and by Proposition 4 we conclude that

$$\begin{aligned}
\lambda_k(\Delta_P) &\leq \max_{u \in U} \frac{D_2(\widetilde{\mathcal{I}}u)}{\|u\|_P^2} \\
&\leq 2A_8\Big(1 + A_1(\theta + \widetilde{\delta})\Big)\Big(1 + A_3\frac{\widetilde{\delta}}{r}\Big) \max_{u \in U} \frac{b_r(u)}{\|u\|_n^2} \\
&\leq 2A_8\Big(1 + A_1(\theta + \widetilde{\delta})\Big)\Big(1 + A_3\frac{\widetilde{\delta}}{r}\Big) \frac{\lambda_k(G_{n,r})}{nr^{d+2}},
\end{aligned}$$

establishing the upper bound in (A.22).

Organization of this section. The rest of this section will be devoted to proving Propositions 3, 4 and 5. To prove the latter two propositions, it will help to introduce the intermediate energies

$$\widetilde{E}_r(f, \eta, V) := \frac{1}{r^{d+2}} \int_V \int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 \eta\Big(\frac{\|x' - x\|}{r}\Big) \widetilde{p}_n(x')\widetilde{p}_n(x)\, dx'\, dx$$

and

$$E_r(f, \eta, V) := \frac{1}{r^{d+2}} \int_V \int_{\mathcal{X}} \big(f(x') - f(x)\big)^2 \eta\Big(\frac{\|x' - x\|}{r}\Big) p(x')p(x)\, dx'\, dx.$$

Here $\eta : [0, \infty) \to [0, \infty)$ is an arbitrary kernel, and $V \subseteq \mathcal{X}$ is a measurable set. We will abbreviate $\widetilde{E}_r(f, \eta, \mathcal{X})$ as $\widetilde{E}_r(f, \eta)$ and $\widetilde{E}_r(f, K) = \widetilde{E}_r(f)$ (and likewise with E_r.)

The proof of Proposition 3 is given in Section A4.2. In Section A4.3, we establish relationships between the (non-random) functionals $E_r(f)$ and $D_2(f)$, as well as providing estimates on some assorted integrals. In Section A4.4, we establish relationships between the stochastic functionals $\widetilde{E}_r(f)$ and $E_r(f)$, between $\widetilde{E}_r\big(\widetilde{\mathcal{I}}(u)\big)$ and $b_r(u)$, and between $\widetilde{E}_r\big(f\big)$ and $b_r\big(\widetilde{\mathcal{P}}f\big)$. Finally, in Section A4.5 we use these various relationships to prove Propositions 4 and 5.

A4.2 Proof of Proposition 3

We start by defining the density $\widetilde{p}_n$, which will be piecewise constant over a particular partition $\mathcal{Q}$ of $\mathcal{X}$. Specifically, for each Q in $\mathcal{Q}$ and every $x \in Q$, we set

$$\widetilde{p}_n(x) := \frac{P_n(Q)}{\mathrm{vol}(Q)}, \tag{A.31}$$

where $\mathrm{vol}(\cdot)$ denotes the Lebesgue measure. Then $\widetilde{P}_n(U) = \int_U \widetilde{p}_n(x)\, dx$.

We now construct the partition $\mathcal{Q}$, in progressive degrees of generality on the domain $\mathcal{X}$.

- In the special case of the unit cube $\mathcal{X} = (0,1)^d$, the partition will simply be a collection of cubes,
$$\mathcal{Q} = \Big\{Q_k : k \in [\widetilde{\delta}^{-1}]^d\Big\},$$
where $Q_k = \widetilde{\delta}\Big([k_1 - 1, k_1] \otimes \cdots \otimes [k_d - 1, k_d]\Big)$ and we assume without loss of generality that $\widetilde{\delta}^{-1} \in \mathbb{N}$.
- If $\mathcal{X}$ is an open, connected set with smooth boundary, then by Proposition 3.2 of García Trillos and Slepčev [2015], there exist a finite number $N(\mathcal{X}) \in \mathbb{N}$ of disjoint polytopes which cover $\mathcal{X}$. Moreover, letting U_j denote the intersection of the jth of these polytopes with $\bar{\mathcal{X}}$, this proposition establishes that for each j there exists a bi-Lipschitz homeomorphism $\Phi_j : U_j \to [0,1]^d$. We take the collection
$$\mathcal{Q} = \Big\{\Phi_j^{-1}(Q_k) : j = 1, \dots, N(\mathcal{X}) \;\; \text{and} \;\; k \in [\widetilde{\delta}^{-1}]^d\Big\}$$
to be our partition. Denote by L_Φ the maximum of the bi-Lipschitz constants of $\Phi_1, \dots, \Phi_{N(\mathcal{X})}$.
- Finally, in the general case where $\mathcal{X}$ is an open, connected set with Lipschitz boundary, then there exists a bi-Lipschitz homeomorphism Ψ between $\mathcal{X}$ and a smooth, open, connected set with Lipschitz boundary. Letting Φ_j and $\widetilde{Q}_{j,k}$ be as before, we take the collection
$$\mathcal{Q} = \Big\{\widetilde{Q}_{j,k} = \Big(\Psi^{-1} \circ \Phi_j^{-1}\Big)(Q_k) : j = 1, \dots, N(\mathcal{X}) \;\; \text{and} \;\; k \in [\widetilde{\delta}^{-1}]^d\Big\}$$
to be our partition. Denote by L_Ψ the bi-Lipschitz constant of Ψ.

Let us record a few facts which hold for all $\widetilde{Q}_{j,k} \in \mathcal{Q}$, and which follow from the bi-Lipschitz properties of Φ_j and Ψ: first that
$$\mathrm{diam}(\widetilde{Q}_{j,k}) \leq L_\Psi \mathrm{L}_\Phi \widetilde{\delta}, \tag{A.32}$$
and second that
$$\mathrm{vol}(\widetilde{Q}_{j,k}) \geq \bigg(\frac{1}{L_\Psi L_\Phi}\bigg)^d \widetilde{\delta}^d. \tag{A.33}$$

We now use these facts to show that $\widetilde{P}_n$ satisfies the claims of Proposition 3. On the one hand for every $Q \in \mathcal{Q}$, letting $N(Q)$ denote the number of design points $\{X_1, \dots, X_k\}$ which fall in Q, we have
$$\widetilde{P}_n(Q) = \int_Q \widetilde{p}_n(x)\,dx = P_n(Q) = \frac{N(Q)}{n}.$$

Moreover, ignoring those cells for which $N(Q) = 0$ (since $\widetilde{P}_n(Q) = 0$ for such Q, and so they do not contribute to the essential supremum in (A.24)), appropriately dividing each remaining cell $Q \in \mathcal{Q}$ into $N(Q)$ subsets $S_1, \dots, S_{N(Q)}$ of equal volume, and mapping each S_ℓ to a different design point $X_i \in Q$, we can exhibit a transport map T from $\widetilde{P}_n$ to P_n for which
$$\|T - \mathrm{Id}\|_{L^\infty(\widetilde{P}_n)} \leq \max_{Q \in \mathcal{Q}} \mathrm{diam}(Q) \leq L_\Psi \mathrm{L}_\Phi \widetilde{\delta}.$$

On the other hand, applying the triangle inequality we have that for $x \in \widetilde{Q}_{j,k}$
$$|\widetilde{p}_n(x) - p(x)| \leq \bigg|\frac{P_n(\widetilde{Q}_{j,k}) - P(\widetilde{Q}_{j,k})}{\mathrm{vol}(\widetilde{Q}_{j,k})}\bigg| + \frac{1}{\mathrm{vol}(\widetilde{Q}_{j,k})}\int_{\widetilde{Q}_{j,k}} |p(x') - p(x)|\,dx,$$
and using the Lipschitz property of p we find that
$$\|\widetilde{p}_n - p\|_{L^\infty} \leq \max_{j,k}\bigg|\frac{P_n(\widetilde{Q}_{j,k}) - P(\widetilde{Q}_{j,k})}{\mathrm{vol}(\widetilde{Q}_{j,k})}\bigg| + L_p L_\Phi L_\Psi \widetilde{\delta}. \tag{A.34}$$

From Hoeffding's inequality and a union bound, we obtain that
$$\begin{aligned}\mathbb{P}\Big(\big|P_n(\widetilde{Q}) - P(\widetilde{Q})\big| \leq \theta P(\widetilde{Q}) \quad \forall \widetilde{Q} \in \mathcal{Q}\Big) &\geq 1 - 2\sharp(\mathcal{Q}) \cdot \exp\bigg\{-\frac{\theta^2 n \min\{P(\widetilde{Q})\}}{3}\bigg\} \\ &\geq 1 - \frac{2N(\mathcal{X})}{\widetilde{\delta}^d} \cdot \exp\bigg\{-\frac{\theta^2 n p_{\min} \widetilde{\delta}^d}{3\big(L_\Psi L_\Phi\big)^d}\bigg\}.\end{aligned}$$

Noting that by assumption $P(\widetilde{Q}) \leq p_{\max}\mathrm{vol}(\widetilde{Q})$ and $\widetilde{\delta}^{-d} \leq n$, the claim follows upon plugging back into (A.34), and setting

$$a_0 := \frac{1}{3\big(L_\Psi L_\Phi\big)^d} \quad \text{and} \quad A_0 := \max\Big\{2N(\mathcal{X}), L_p L_\Psi L_\Phi, L_\Psi L_\Phi\Big\}$$

in the statement of the proposition.

A4.3 Non-random functionals and integrals

Let us start by making the following observation, which we make use of repeatedly in this section. Let $\eta : [0,\infty) \to [0,\infty)$ be an otherwise arbitrary function. As a consequence of (P1), there exist constants c_0 and a_3 which depend on $\mathcal{X}$, such that for any $0 < \varepsilon \leq c_0$ it holds that

$$\int_{B(x,\varepsilon)\cap\mathcal{X}} \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx' \geq a_3 \cdot \int_{B(x,\varepsilon)} \eta\left(\frac{\|x'-x\|}{\varepsilon}\right) dx' \tag{A.35}$$

As a special case: when $\eta(x) = 1$, this implies $\mathrm{vol}\big(B(x,\varepsilon)\cap\mathcal{X}\big) \geq a_3\nu_d\varepsilon^d$ for any $0 < \varepsilon \leq c_0$.

We have already upper bounded $E_r(f)$ by (a constant times) $D_2(f)$ in the proof of Lemma 1. In Lemma 7, we establish the reverse inequality.

Lemma 7 **(cf. Lemma 9 of García Trillos et al. [2019], Lemma 5.5 of Burago et al. [2014]).** *For any $f \in L^2(\mathcal{X})$, and any $0 < h \leq c_0$, it holds that*

$$\sigma_K D_2(\Lambda_h f) \leq A_8 E_h(f).$$

To prove Lemma 7, we require upper and lower bounds on $\tau_h(x)$, as well as an upper bound on the gradient of τ_h. The lower bound here—$\tau_h(x) \geq a_3$—is quite a bit a looser than what can be shown when $\mathcal{X}$ has no boundary. The same is the case regarding the upper bound of the size of the gradient $\|\nabla\tau_h(x)\|$. However, the bounds as stated here will be sufficient for our purposes.

Lemma 8. *For any $0 < h \leq c_0$, for all $x \in \mathcal{X}$ it holds that*

$$a_3 \leq \tau_h(x) \leq 1.$$

and

$$\|\nabla\tau_h(x)\| \leq \frac{1}{\sqrt{d\sigma_K}h}.$$

Finally, to prove part (2) of Proposition 5, we require Lemma 9, which gives an estimate on the error $\Lambda_h f - f$ in $\|\cdot\|_P^2$ norm.

Lemma 9 **(c.f Lemma 8 of García Trillos et al. [2019], Lemma 5.4 of Burago et al. [2014]).** *For any $0 < h \leq c_0$,*

$$\big\|\Lambda_h f\big\|_P^2 \leq \frac{p_{\max}}{a_3 p_{\min}}\big\|f\big\|_P^2, \tag{A.36}$$

and

$$\big\|\Lambda_h f - f\big\|_P^2 \leq \frac{1}{a_3\sigma_K p_{\min}} h^2 E_h(f), \tag{A.37}$$

for all $f \in L^2(\mathcal{X})$.

Proof of Lemma 7. For any $a \in \mathbb{R}$, $\Lambda_h f$ satisfies the identity

$$\Lambda_h f(x) = a + \frac{1}{h^d\tau_h(x)}\int_{\mathcal{X}} \eta_h(x',x)\big(f(x') - a\big)\,dx',$$

and by differentiating with respect to x, we obtain

$$\big(\nabla\Lambda_h f\big)(x) = \frac{1}{h^d\tau_h(x)}\int_{\mathcal{X}} \big(\nabla\eta_h(x',\cdot)\big)(x)\big(f(x') - a\big)\,dx' + \nabla\left(\frac{1}{\tau_h}\right)(x)\cdot\frac{1}{h^d}\int_{\mathcal{X}} \eta_h(x',x)\big(f(x') - a\big)\,dx'$$

Plugging in $a = f(x)$, we get $\nabla\Lambda_h f(x) = J_1(x)/\tau_h(x) + J_2(x)$ for

$$J_1(x) := \frac{1}{h^d}\int_{\mathcal{X}} \bigl(\nabla\eta_h(x',\cdot)\bigr)(x)\bigl(f(x') - f(x)\bigr)\,dx', \quad J_2(x) := \nabla\Bigl(\frac{1}{\tau_h}\Bigr)(x)\cdot\frac{1}{h^d}\int_{\mathcal{X}}\eta_h(x',x)\bigl(f(x') - f(x)\bigr)\,dx'.$$

To upper bound $\bigl\|J_1(x)\bigr\|^2$, we first compute the gradient of $\eta_h(x',\cdot)$,

$$\begin{aligned}\bigl(\nabla\eta_h(x',\cdot)\bigr)(x) &= \frac{1}{h}\psi'\Bigl(\frac{\|x'-x\|}{h}\Bigr)\frac{(x-x')}{\|x'-x\|} \\ &= \frac{1}{\sigma_K h^2}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)(x'-x),\end{aligned}$$

and additionally note that $\|J_1(x)\|^2 = \sup_w\bigl(\langle J_1(x), w\rangle\bigr)^2$ where the supremum is over unit norm vector. Taking w to be a unit norm vector which achieves this supremum, we have that

$$\begin{aligned}\bigl\|J_1(x)\bigr\|^2 &= \frac{1}{\sigma_K^2 h^{4+2d}}\biggl[\int_{\mathcal{X}}\bigl(f(x') - f(x)\bigr)K\Bigl(\frac{\|x'-x\|}{h}\Bigr)(x'-x)^\top w\,dx'\biggr]^2 \\ &\leq \frac{1}{\sigma_K^2 h^{4+2d}}\biggl[\int_{\mathcal{X}}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)\bigl((x'-x)^\top w\bigr)^2\,dx'\biggr]\biggl[\int_{\mathcal{X}}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)\bigl(f(x') - f(x)\bigr)^2\,dx'\biggr].\end{aligned}$$

By a change of variables, we obtain

$$\int_{\mathcal{X}}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)\bigl((x'-x)^\top w\bigr)^2\,dx' \leq h^{d+2}\int_{\mathcal{X}}K\bigl(\|z\|\bigr)\bigl(z^\top w\bigr)^2\,dz \leq \sigma_K h^{d+2},$$

with the resulting upper bound

$$\bigl\|J_1(x)\bigr\|^2 \leq \frac{1}{\sigma_K h^{2+d}}\int_{\mathcal{X}}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)\bigl(f(x') - f(x)\bigr)^2\,dx'.$$

To upper bound $\bigl\|J_2(x)\bigr\|^2$, we use the Cauchy-Schwarz inequality along with the observation $\eta_h(x',x) \leq \frac{1}{\sigma_K}K\bigl(\|x'-x\|/h\bigr)$ to deduce

$$\begin{aligned}\bigl\|J_2(x)\bigr\|^2 &\leq \Bigl\|\nabla\Bigl(\frac{1}{\tau_h}\Bigr)(x)\Bigr\|^2\frac{1}{h^{2d}}\biggl[\int_{\mathcal{X}}\eta_h(x',x)\,dx'\biggr]\cdot\biggl[\int_{\mathcal{X}}\eta_h(x',x)\bigl(f(x') - f(x)\bigr)^2\,dx'\biggr] \\ &= \Bigl\|\nabla\Bigl(\frac{1}{\tau_h}\Bigr)(x)\Bigr\|^2\frac{\tau_h(x)}{h^d}\int_{\mathcal{X}}\eta_h(x',x)\bigl(f(x') - f(x)\bigr)^2\,dx' \\ &\leq \Bigl\|\nabla\Bigl(\frac{1}{\tau_h}\Bigr)(x)\Bigr\|^2\frac{\tau_h(x)}{\sigma_K h^d}\int_{\mathcal{X}}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)\bigl(f(x') - f(x)\bigr)^2\,dx' \\ &\leq \frac{1}{da_3^2\sigma_K^2 h^{2+d}}\int_{\mathcal{X}}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)\bigl(f(x') - f(x)\bigr)^2\,dx',\end{aligned}$$

where the last inequality follows from the estimates on τ_h and $\nabla\tau_h$ provided in Lemma 8. Combining our bounds on $\bigl\|J_1(x)\bigr\|^2$ and $\bigl\|J_2(x)\bigr\|^2$ along with the lower bound on $\tau_h(x)$ in Lemma 8 and integrating over $\mathcal{X}$, we have

$$\begin{aligned}\sigma_K D_2(\Lambda_h f) &= \sigma_K\int_{\mathcal{X}}\Bigl\|\bigl(\nabla\Lambda_h f)(x)\Bigr\|^2 p^2(x)\,dx \\ &\leq 2\sigma_K\int_{\mathcal{X}}\biggl(\frac{\bigl\|J_1(x)\bigr\|^2}{\tau_h^2(x)} + \bigl\|J_2(x)\bigr\|^2\biggr)p^2(x)\,dx \\ &\leq \biggl(\frac{1}{a_3^2} + \frac{1}{da_3^2\sigma_K}\biggr)\frac{2}{h^{d+2}}\int_{\mathcal{X}}\int_{\mathcal{X}}K\Bigl(\frac{\|x'-x\|}{h}\Bigr)\bigl(f(x') - f(x)\bigr)^2 p^2(x)\,dx'\,dx \\ &\leq 2\Bigl(1 + \frac{L_p h}{p_{\min}}\Bigr)\biggl(\frac{1}{a_3^2} + \frac{1}{da_3^2\sigma_K}\biggr)E_h(f),\end{aligned}$$

and taking $A_8 := 2\Bigl(1 + \frac{L_p c_0}{p_{\min}}\Bigr)\Bigl(\frac{1}{a_3^2} + \frac{1}{da_3^2\sigma_K}\Bigr)$ completes the proof of Lemma 7.

Proof of Lemma 8. We first establish our estimates of $\tau_h(x)$, and then upper bound $\|\nabla\tau_h(x)\|$. Using (A.35), we have that

$$\begin{aligned}\tau_h(x) &= \frac{1}{h^d}\int_{\mathcal{X}\cap B(x,h)} \psi\biggl(\frac{\|x'-x\|}{h}\biggr)\,dx' \\ &\geq \frac{a_3}{h^d}\int_{B(x,h)} \psi\biggl(\frac{\|x'-x\|}{h}\biggr)\,dx' \\ &= a_3\int_{B(0,1)} \psi(\|z\|)\,dz,\end{aligned}$$

and it follows from similar reasoning that $\tau_h(x) \leq \int_{B(0,1)} \psi(\|z\|)\,dz$.

We will now show that $\int_{B(0,1)} \psi(\|z\|)\,dz = 1$, from which we derive the estimates $a_3 \leq \tau_h(x) \leq 1$. To see the identity, note that on the one hand, by converting to polar coordinates and integrating by parts we obtain

$$\int_{B(0,1)} \psi\bigl(\|z\|\bigr)\,dz = d\nu_d\int_0^1 \psi(t)t^{d-1}\,dt = -\nu_d\int_0^1 \psi'(t)t^d\,dt = \frac{\nu_d}{\sigma_K}\int_0^1 t^{d+1}K(t)\,dt;$$

on the other hand, again converting to polar coordinates, we have

$$\sigma_K = \frac{1}{d}\int_{\mathbb{R}^d} \|x\|^2K(\|x\|)\,dx = \nu_d\int_0^1 t^{d+1}K(t)\,dt,$$

and so $\int_{B(0,1)} \psi(\|z\|)\,dz = 1$.

Now we upper bound $\|\nabla\tau_h(x)\|^2$. Exchanging derivative and integral, we have

$$\nabla\tau_h(x) = \frac{1}{h^d}\int_{\mathcal{X}} \bigl(\nabla\eta_h(x',\cdot)\bigr)(x)\,dx' = \frac{1}{\sigma_K h^{d+2}}\int_{\mathcal{X}} K\biggl(\frac{\|x'-x\|}{h}\biggr)(x'-x)\,dx',$$

whence by the Cauchy-Schwarz inequality,

$$\|\nabla\tau_h(x)\|^2 \leq \frac{1}{\sigma_K^2h^{2d+4}}\biggl[\int_{\mathcal{X}} K\biggl(\frac{\|x'-x\|}{h}\biggr)\,dx'\biggr]\biggl[\int_{\mathcal{X}} K\biggl(\frac{\|x'-x\|}{h}\biggr)\|x'-x\|^2\,dx',\biggr] \leq \frac{1}{d\sigma_K h^2},$$

concluding the proof of Lemma 8.

We remark that while $\nabla\tau(x) = 0$ when $B(x,r) \in \mathcal{X}$, near the boundary the upper bound we derived by using Cauchy-Schwarz appears tight.

Proof of Lemma 9. By Jensen's inequality and Lemma 8,

$$\begin{aligned}\Bigl|\Lambda_h f(x)\Bigr|^2 &\leq \frac{1}{h^d\tau_h(x)}\int_{\mathcal{X}} \eta_h(x',x)\bigl[f(x')\bigr]^2\,dx' \\ &\leq \frac{1}{a_3h^dp_{\min}}\int_{\mathcal{X}} \eta_h(x',x)\bigl[f(x')\bigr]^2p(x')\,dx'.\end{aligned}$$

Then, integrating over x, and recalling that$\int_{B(0,1)} \psi(\|z\|) = 1$ as shown in the proof of Lemma 8, we have

$$\begin{aligned}\bigl\|\Lambda_h f\bigr\|_P^2 &\leq \frac{1}{a_3h^dp_{\min}}\int_{\mathcal{X}}\int_{\mathcal{X}} \eta_h(x',x)\bigl[f(x')\bigr]^2p(x')p(x)\,dx'\,dx \\ &\leq \frac{p_{\max}}{a_3h^dp_{\min}}\int_{\mathcal{X}} \bigl[f(x')\bigr]^2p(x')\biggl(\int_{\mathcal{X}} \eta_h(x',x)\,dx\biggr)\,dx' \\ &\leq \frac{p_{\max}}{a_3p_{\min}}\int_{\mathcal{X}} \bigl[f(x')\bigr]^2p(x')\biggl(\int_{B(0,1)} \psi(\|z\|)\,dz\biggr)\,dx' \\ &= \frac{p_{\max}}{a_3p_{\min}}\|f\|_P^2.\end{aligned}$$

To establish (A.37), noting that $\Lambda_h a = a$ for any $a \in \mathbb{R}$, we have that

$$\begin{aligned}
\bigl|\Lambda_r f(x) - f(x)\bigr|^2 &= \biggl[\frac{1}{h^d \tau_h(x)} \int_{\mathcal{X}} \eta_h(x',x)\bigl(f(x') - f(x)\bigr)\,dx'\biggr]^2 \\
&\le \frac{1}{h^{2d}\tau_h^2(x)}\biggl[\int_{\mathcal{X}} \eta_h(x',x)\,dx'\biggr]\cdot\biggl[\int_{\mathcal{X}} \eta_h(x',x)\bigl(f(x') - f(x)\bigr)^2\,dx'\biggr] \\
&= \frac{1}{h^d \tau_h(x)} \int_{\mathcal{X}} \eta_h(x',x)\bigl(f(x') - f(x)\bigr)^2\,dx'. \\
&\le \frac{1}{h^d \tau_h(x) p_{\min}} \int_{\mathcal{X}} \eta_h(x',x)\bigl(f(x') - f(x)\bigr)^2 p(x')\,dx'.
\end{aligned}$$

From here, we can use the lower bound $\tau_h(x) \ge a_3$ stated in Lemma 8, as well as the upper bound $\eta_h(x',x) \le (1/\sigma_K)K(\|x'-x\|/h)$, to deduce

$$\bigl|\Lambda_r f(x) - f(x)\bigr|^2 \le \frac{1}{h^d a_3 \sigma_K p_{\min}} \int_{\mathcal{X}} K\biggl(\frac{\|x'-x\|}{h}\biggr)\bigl(f(x') - f(x)\bigr)^2 p(x')\,dx'.$$

Then integrating over $\mathcal{X}$ with respect to p yields (A.37).

A4.4 Random functionals

We will use Lemma 10 in the proof of Proposition 5.

Lemma 10 (cf. Lemma 3.4 of Burago et al. [2014]). *Let $U \subseteq \mathcal{X}$ be a measurable subset such that* $\mathrm{vol}(U) > 0$, *and* $\mathrm{diam}(U) \le 2A_0\widetilde{\delta}$. *Then, letting $a = (\widetilde{P}_n(U))^{-1} \cdot \int_U f(x)\widetilde{p}_n(x)\,dx$ be the average of f over U, it holds that*

$$\int_U \Bigl|f(x) - a\Bigr|^2 \widetilde{p}_n(x)\,dx \le A_3 r^2 \widetilde{E}_r(f,U).$$

Now we relate $\widetilde{E}_r(f)$ and $E_r(f)$. Some standard calculations show that for $A_1 := 3A_0/p_{\min}$,

$$\bigl(1 - A_1(\theta + \widetilde{\delta})\bigr)E_r(f) \le \widetilde{E}_r(f) \le \bigl(1 + A_1(\theta + \widetilde{\delta})\bigr)E_r(f), \tag{A.38}$$

as well as implying that the norms $\|f\|_P$ and $\|f\|_n$ satisfy

$$\bigl(1 - A_1(\theta + \widetilde{\delta})\bigr)\|f\|_P^2 \le \|f\|_{\widetilde{P}_n}^2 \le \bigl(1 + A_1(\theta + \widetilde{\delta})\bigr)\|f\|_P^2. \tag{A.39}$$

Lemma 11 relates the graph Sobolev semi-norm $b_r(\widetilde{\mathcal{P}}f)$ to the non-local energy $\widetilde{E}_r(f)$.

Lemma 11 (cf. Lemma 13 of García Trillos et al. [2019], Lemma 4.3 of Burago et al. [2014]). *For any $f \in L^2(\mathcal{X})$,*

$$b_r(\widetilde{\mathcal{P}}f) \le \Bigl(1 + A_9\frac{\widetilde{\delta}}{r}\Bigr)\widetilde{E}_{r+2A_0\widetilde{\delta}}(f).$$

In Lemma 12, we establish the reverse of Lemma 11.

Lemma 12 (cf. Lemma 14 of García Trillos et al. [2019]). *For any $u \in L^2(P_n)$,*

$$\widetilde{E}_{r-2A_0\widetilde{\delta}}(\widetilde{\mathcal{P}}^\star u) \le \biggl(1 + A_3\frac{\widetilde{\delta}}{r}\biggr)b_r(u).$$

Proof of Lemma 10. A symmetrization argument implies that

$$\int_U \Bigl|f(x) - a\Bigr|^2 \widetilde{p}_n(x)\,dx = \frac{1}{2\widetilde{P}_n(U)}\int_U\int_U \bigl|f(x') - f(x)\bigr|^2 \widetilde{p}_n(x')\widetilde{p}_n(x)\,dx'\,dx \tag{A.40}$$

Now, since x' and x belong to U, we have that $\|x' - x\| \le 2A_0\widetilde{\delta}$. Set $V = B(x,r) \cap B(x',r)$, and note that $B(x, r - 2A_0\widetilde{\delta}) \subseteq V$. Moreover, $r - 2A_0\widetilde{\delta} \le r \le c_0$ by assumption. Therefore by (A.35),

$$\mathrm{vol}\bigl(V \cap \mathcal{X}\bigr) \ge \mathrm{vol}\bigl(B(x, r - 2A_0\widetilde{\delta}) \cap \mathcal{X}\bigr) \ge a_3\nu_d(r - 2A_0\widetilde{\delta})^d \ge \frac{a_3\nu_d}{2^d}r^d$$

where the last inequality follows since $\widetilde{\delta} \leq \frac{1}{4A_0} r$. Using the triangle inequality

$$\left|f(x') - f(x)\right|^2 \leq 2\left(\left|f(x') - f(z)\right|^2 + \left|f(z) - f(x)\right|^2\right)$$

we have that for any x and x' in U,

$$\begin{aligned}
\left|f(x') - f(x)\right|^2 &\leq \frac{2}{\mathrm{vol}(V \cap \mathcal{X})} \int_{V \cap \mathcal{X}} \left|f(x') - f(z)\right|^2 + \left|f(z) - f(x)\right|^2 \, dz \\
&\leq \frac{2^{d+1}}{a_3 \nu_d r^d} \int_{V \cap \mathcal{X}} \left|f(x') - f(z)\right|^2 + \left|f(z) - f(x)\right|^2 \, dz \\
&\leq \frac{2^{d+2}}{K(1) a_3 \nu_d r^d p_{\min}} \Big(F(x') + F(x)\Big), \qquad \text{(A.41)}
\end{aligned}$$

where in the last inequality we set

$$F(x) := \int_{\mathcal{X}} K\left(\frac{\|z - x\|}{r}\right)\big(f(z) - f(x)\big)^2 \widetilde{p}_n(x) \, dx,$$

and use the facts that $\widetilde{p}_n(x) \geq p_{\min}/2$, that $K(\|z - x\|/r) \geq K(1)$ for all $z \in B(x, r)$.

Plugging the upper bound (A.41) back into (A.40), we have that

$$\begin{aligned}
\int_U \left|f(x) - a\right|^2 \widetilde{p}_n(x) \, dx &\leq \frac{2^{d+2}}{K(1) a_3 \nu_d r^d} \int_U F(x) \widetilde{p}_n(x) \, dx \\
&= \frac{2^{d+2}}{K(1) a_3 \nu_d} r^2 \widetilde{E}_r(f, U),
\end{aligned}$$

and Lemma 10 follows by taking $A_3 := 2^{d+2}/(K(1) a_3 \nu_d)$.

Proof of Lemma 11. Recalling that $\big(\widetilde{\mathcal{P}} f\big)(X_i) = n \cdot \int_{U_i} f(x) \widetilde{p}_n(x) \, dx$, by Jensen's inequality,

$$\Big(\big(\widetilde{\mathcal{P}} f\big)(X_i) - \big(\widetilde{\mathcal{P}} f\big)(X_j)\Big)^2 \leq n^2 \cdot \int_{U_i} \int_{U_j} \big(f(x') - f(x)\big)^2 \widetilde{p}_n(x') \widetilde{p}_n(x) \, dx' \, dx.$$

Additionally, the non-increasing and Lipschitz properties of K imply that for any $x \in U_i$ and $x' \in U_j$,

$$K\left(\frac{\|X_i - X_j\|}{r}\right) \leq K\left(\frac{\big(\|x' - x\| - 2A_0 \widetilde{\delta}\big)_+}{r}\right) \leq K\left(\frac{\|x' - x\|}{r + 2A_0 \widetilde{\delta}}\right) + \frac{2 L_K A_0 \widetilde{\delta}}{r} \mathbf{1}\Big\{\|x' - x\| \leq r + 2A_0 \widetilde{\delta}\Big\}.$$

As a result, the graph Dirichlet energy is upper bounded as follows:

$$\begin{aligned}
b_r(\widetilde{\mathcal{P}} f) &= \frac{1}{n^2 r^{d+2}} \sum_{i,j=1}^{n} \Big(\big(\widetilde{\mathcal{P}} f\big)(X_i) - \big(\widetilde{P} f\big)(X_j)\Big)^2 K\left(\frac{\|X_i - X_j\|}{r}\right) \\
&\leq \frac{1}{r^{d+2}} \sum_{i,j=1}^{n} \int_{U_i} \int_{U_j} \big(f(x') - f(x)\big)^2 \widetilde{p}_n(x') \widetilde{p}_n(x) K\left(\frac{\|X_i - X_j\|}{r}\right) dx' \, dx \\
&\leq \frac{1}{r^{d+2}} \sum_{i,j=1}^{n} \int_{U_i} \int_{U_j} \big(f(x') - f(x)\big)^2 \widetilde{p}_n(x') \widetilde{p}_n(x) \left[K\left(\frac{\|x' - x\|}{r + 2A_0 \widetilde{\delta}}\right) + \frac{2 L_K A_0 \widetilde{\delta}}{r} \mathbf{1}\Big\{\|x' - x\| \leq r + 2\widetilde{\delta}\Big\}\right] dx' \, dx \\
&= \left(1 + 2A_0 \frac{\widetilde{\delta}}{r}\right)^{d+2} \left[\widetilde{E}_{r + 2A_0 \widetilde{\delta}}(f) + \frac{2 L_K A_0 \widetilde{\delta}}{r} \widetilde{E}_{r + 2A_0 \widetilde{\delta}}(f; \mathbf{1}_{[0,1]})\right],
\end{aligned}$$

for $\mathbf{1}_{[0,1]}(t) = \mathbf{1}\{0 \leq t \leq 1\}$. But by assumption $\widetilde{E}_{r + 2A_0 \widetilde{\delta}}(f; \mathbf{1}_{[0,1]}) \leq 1/(K(1)) \widetilde{E}_{r + 2A_0 \widetilde{\delta}}(f)$, and so we obtain

$$b_r(\widetilde{\mathcal{P}} f) \leq \left(1 + 2A_0 \frac{\widetilde{\delta}}{r}\right)^{d+2} \left(1 + \frac{2 L_K A_0 \widetilde{\delta}}{r K(1)}\right) \widetilde{E}_{r + 2A_0 \widetilde{\delta}}(f);$$

the Lemma follows upon choosing $A_9 := A_0 (2^{d+4} + \frac{4 L_K}{K(1)})$.

Proof of Lemma 12. For brevity, we write $\widetilde{r} := r - 2A_0\widetilde{\delta}$. We begin by expanding the energy $\widetilde{E}_{\widetilde{r}}\bigl(\widetilde{\mathcal{P}}^\star u\bigr)$ as a double sum of double integrals,

$$\widetilde{E}_{\widetilde{r}}\bigl(\widetilde{\mathcal{P}}^\star u\bigr) = \frac{1}{\widetilde{r}^{d+2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{U_i}\int_{U_j}\Bigl(u(X_i) - u(X_j)\Bigr)^2 K\biggl(\frac{\|x' - x\|}{\widetilde{r}}\biggr)\widetilde{p}_n(x')\widetilde{p}_n(x)\,dx'\,dx.$$

We next use the Lipschitz property of the kernel K—in particular that for $x \in U_i$ and $x' \in U_j$,

$$K\biggl(\frac{\|x' - x\|}{\widetilde{r}}\biggr) \leq K\biggl(\frac{\|X_i - X_j\|}{r}\biggr) + \frac{2A_0L_K\widetilde{\delta}}{\widetilde{r}}\cdot\mathbf{1}\biggl\{\frac{\|x' - x\|}{\widetilde{r}} \leq 1\biggr\},$$

—to conclude that

$$\begin{aligned}
\widetilde{E}_{\widetilde{r}}\bigl(\widetilde{\mathcal{P}}^\star u\bigr) &\leq \frac{1}{n^2\widetilde{r}^{d+2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\Bigl(u(X_i) - u(X_j)\Bigr)^2 K\biggl(\frac{\|X_i - X_j\|}{r}\biggr) + \frac{2A_0L_K\widetilde{\delta}}{\widetilde{r}}\widetilde{E}_{\widetilde{r}}(\widetilde{\mathcal{P}}^\star u, \mathbf{1}_{[0,1]}) \\
&\leq \biggl(1 + 2^{d+2}A_0\frac{\widetilde{\delta}}{r}\biggr)b_r(u) + \frac{2A_0L_K\widetilde{\delta}}{\widetilde{r}}\widetilde{E}_{\widetilde{r}}(\widetilde{\mathcal{P}}^\star u, \mathbf{1}_{[0,1]}) \\
&\leq \biggl(1 + 2^{d+2}A_0\frac{\widetilde{\delta}}{r}\biggr)b_r(u) + \frac{4A_0L_K\widetilde{\delta}}{K(1)r}\widetilde{E}_{\widetilde{r}}(\widetilde{\mathcal{P}}^\star u).
\end{aligned}$$

In other words,

$$\begin{aligned}
\widetilde{E}_{\widetilde{r}}\bigl(\widetilde{\mathcal{P}}^\star u\bigr) &\leq \biggl(1 - \frac{4A_0L_K\widetilde{\delta}}{K(1)r}\biggr)^{-1}\biggl(1 + 2^{d+2}A_0\frac{\widetilde{\delta}}{r}\biggr)b_r(u) \\
&\leq \biggl(1 + \frac{\widetilde{\delta}}{r}\Bigl(\frac{8A_0L_K}{K(1)} + 2^{d+3}\Bigr)\biggr)b_r(u),
\end{aligned}$$

where the second inequality follows from the algebraic identities $(1-t)^{-1} \leq (1+2t)$ for any $0 < t < 1/2$ and $(1+s)(1+t) < 1+2s+t$ for any $0 < t < 1$ and $s > 0$. The Lemma follows upon choosing $A_3 := \frac{8A_0L_K}{K(1)} + 2^{d+3}$.

A4.5 Proof of Propositions 4 and 5

Proof of Proposition 4. Part (1) of Proposition 4 follows from

$$\begin{aligned}
\sigma_K D_2(\Lambda_{r-2A_0\widetilde{\delta}}\widetilde{\mathcal{P}}^\star u) &\overset{(i)}{\leq} A_8 E_{r-2A_0\widetilde{\delta}}(\widetilde{\mathcal{P}}^\star u) \\
&\overset{(ii)}{\leq} A_8\Bigl(1 + A_1(\theta + \widetilde{\delta})\Bigr)\widetilde{E}_{r-2A_0\widetilde{\delta}}(\widetilde{\mathcal{P}}^\star u) \\
&\overset{(iii)}{\leq} A_8\Bigl(1 + A_1(\theta + \widetilde{\delta})\Bigr)\cdot\Bigl(1 + A_3\frac{\widetilde{\delta}}{r}\Bigr)b_r(u),
\end{aligned}$$

where (i) follows from Lemma 7, (ii) follows from (A.38), and (iii) follows from Lemma 12.

Part (2) of Proposition 4 follows from

$$\begin{aligned}
b_r(\widetilde{\mathcal{P}}f) &\overset{(iv)}{\leq} \Bigl(1 + A_9\frac{\widetilde{\delta}}{r}\Bigr)\widetilde{E}_{r+2A_0\widetilde{\delta}}(f) \\
&\overset{(v)}{\leq} \Bigl(1 + A_1(\theta + \widetilde{\delta})\Bigr)\Bigl(1 + A_9\frac{\widetilde{\delta}}{r}\Bigr)E_{r+2A_0\widetilde{\delta}}(f) \\
&\overset{(vi)}{\leq} \Bigl(1 + A_1(\theta + \widetilde{\delta})\Bigr)\cdot\Bigl(1 + A_9\frac{\widetilde{\delta}}{r}\Bigr)\cdot\Bigl(\frac{C_5p_{\max}^2}{p_{\min}^2}\Bigr)\cdot\sigma_K D_2(f),
\end{aligned}$$

where (iv) follows from Lemma 11, (v) follows from (A.38), and (vi) follows from the proof of Lemma 1.

Proof of Proposition 5. *Proof of (1).* We begin by upper bounding $\big\|\widetilde{\mathcal{P}}f\big\|_n$. By the Cauchy-Schwarz inequality and the bound on $\|\widetilde{p}_n - p\|_\infty$ in (A.26),

$$\begin{aligned}
\Big|\widetilde{\mathcal{P}}f(X_i)\Big|^2 &= n^2\Big|\int_{U_i} f(x)\widetilde{p}_n(x)\,dx\Big|^2 \\
&\leq n\int_{U_i} \big|f(x)\big|^2\widetilde{p}_n(x)\,dx \\
&\leq n\Big(1 + A_1(\theta+\widetilde{\delta})\Big)\Big[\int_{U_i}\big|f(x)\big|^2 p(x)\,dx + A_1(\theta+\widetilde{\delta})\int_{U_i}\big|f(x)\big|^2 p(x)\,dx\Big],
\end{aligned}$$

and summing over $i = 1,\dots,n$, we obtain

$$\big\|\widetilde{\mathcal{P}}f\big\|_n^2 \leq \Big(1 + A_1(\theta+\widetilde{\delta})\Big)\big\|f\big\|_P^2. \tag{A.42}$$

Now, noticing that $\big\|\widetilde{\mathcal{P}}f\big\|_n = \big\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f\big\|_{\widetilde{P}_n}$, we can use the upper bound (A.42) to show that

$$\begin{aligned}
\Big|\big\|\widetilde{\mathcal{P}}f\big\|_n^2 - \big\|f\big\|_P^2\Big| &\leq \Big|\big\|\widetilde{\mathcal{P}}f\big\|_n^2 - \big\|f\big\|_{\widetilde{P}_n}^2\Big| + \Big|\big\|f\big\|_{\widetilde{P}_n}^2 - \big\|f\big\|_P^2\Big| \\
&\overset{(i)}{\leq} \Big|\big\|\widetilde{\mathcal{P}}f\big\|_n^2 - \big\|f\big\|_{\widetilde{P}_n}^2\Big| + A_1(\theta+\widetilde{\delta})\big\|f\big\|_P^2 \qquad \text{(A.43)} \\
&\overset{(ii)}{\leq} 2\sqrt{1 + A_1(\theta+\widetilde{\delta})}\Big|\big\|\widetilde{\mathcal{P}}f\big\|_n - \big\|f\big\|_{\widetilde{P}_n}\Big| \cdot \big\|f\big\|_P + A_1(\theta+\widetilde{\delta})\big\|f\big\|_P^2 \\
&\leq 2\sqrt{1 + A_1(\theta+\widetilde{\delta})}\big\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\big\|_{\widetilde{P}_n} \cdot \big\|f\big\|_P + A_1(\theta+\widetilde{\delta})\big\|f\big\|_P^2, \qquad \text{(A.44)}
\end{aligned}$$

where (i) follows from (A.39) and (ii) follows from (A.39) and (A.42).

It remains to upper bound $\big\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\big\|_{\widetilde{P}_n}^2$. Noting that $\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f$ is piecewise constant over the cells U_i, we have

$$\big\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\big\|_{\widetilde{P}_n}^2 = \sum_{i=1}^n \int_{U_i}\Big(f(x) - n\cdot\int_{U_i} f(x')\widetilde{p}_n(x')\,dx'\Big)^2\widetilde{p}_n(x)\,dx.$$

From Lemma 10, we have that for each $i = 1,\dots,n$,

$$\int_{U_i}\Big(f(x) - n\cdot\int_{U_i} f(x')\widetilde{p}_n(x')\,dx'\Big)^2\widetilde{p}_n(x)\,dx \leq A_3 r^2\widetilde{E}_r(f,U_i).$$

Summing up over i on both sides of the inequality gives

$$\big\|\widetilde{\mathcal{P}}^\star\widetilde{\mathcal{P}}f - f\big\|_{\widetilde{P}_n}^2 \leq A_3 r^2\widetilde{E}_r(f,\mathcal{X}) \leq A_3\Big(1 + A_1(\theta+\widetilde{\delta})\Big)\cdot\Big(\frac{C_5 p_{\mathrm{max}}^2}{p_{\mathrm{min}}^2}\Big)\cdot\sigma_K r^2 D_2(f),$$

where the latter inequality follows from the proof of Proposition 4, Part (2). Then Proposition 5, Part (1) follows by plugging this inequality into (A.44) and taking

$$A_5 := 2\sqrt{A_3}\Big(1 + A_1(\theta+\widetilde{\delta})\Big)\Big(\frac{\sqrt{C_5}p_{\mathrm{max}}}{p_{\mathrm{min}}}\Big)\cdot\sqrt{\sigma_K}.$$

Proof of (2). By the triangle inequality and (A.39),

$$\begin{aligned}
\Big|\|\widetilde{\mathcal{I}}u\|_P^2 - \|u\|_n^2\Big| &\leq \Big|\|\widetilde{\mathcal{I}}u\|_P^2 - \|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2\Big| + \Big|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 - \|u\|_n^2\Big| \\
&\leq A_1(\theta+\widetilde{\delta})\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 + \Big|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 - \|u\|_n^2\Big| \\
&= A_1(\theta+\widetilde{\delta})\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 + \Big(\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} + \|u\|_n\Big)\cdot\Big|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} - \|u\|_n\Big| \qquad \text{(A.45)}
\end{aligned}$$

To upper bound the second term in the above expression, we first note that $\|u\|_n = \|\widetilde{\mathcal{P}}^{\star} u\|_{\widetilde{P}_n}$, and thus

$$\begin{aligned}\Big|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} - \|u\|_n\Big| &= \Big|\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n} - \|\widetilde{\mathcal{P}}^{\star}u\|_{\widetilde{P}_n}\Big| \\ &\overset{(iii)}{\leq} \|\Lambda_{\widetilde{r}}\widetilde{\mathcal{P}}^{\star}u - \widetilde{\mathcal{P}}^{\star}u\|_{\widetilde{P}_n} \\ &\overset{(iv)}{\leq} \widetilde{r}\sqrt{\frac{1}{a_3\sigma_K p_{\min}} E_{\widetilde{r}}(\widetilde{\mathcal{P}}^{\star}u)} \\ &\overset{(v)}{\leq} \widetilde{r}\sqrt{\frac{1 + A_1(\theta+\widetilde{\delta})}{a_3\sigma_K p_{\min}}\Big(1 + A_3\frac{\widetilde{\delta}}{r}\Big) b_r(u)}, \end{aligned} \tag{A.46}$$

where (iii) follows by the triangle inequality, (iv) follows from Lemma 9, and (v) follows from (A.38) and Lemma 12. On the other hand, by (A.39) and Lemma 9,

$$\begin{aligned}\|\widetilde{\mathcal{I}}u\|_{\widetilde{P}_n}^2 &\leq \Big(1 + A_1(\theta+\widetilde{\delta})\Big)\|\widetilde{\mathcal{I}}u\|_P^2 \\ &\leq \frac{p_{\max}}{a_3 p_{\min}} \cdot \Big(1 + A_1(\theta+\widetilde{\delta})\Big)\|\widetilde{\mathcal{P}}^{\star}u\|_P^2 \\ &\leq \frac{p_{\max}}{a_3 p_{\min}} \cdot \Big(1 + A_1(\theta+\widetilde{\delta})\Big)^2\|\widetilde{\mathcal{P}}^{\star}u\|_{\widetilde{P}_n}^2 \\ &= \frac{p_{\max}}{a_3 p_{\min}} \cdot \Big(1 + A_1(\theta+\widetilde{\delta})\Big)^2\|u\|_n^2. \end{aligned}$$

Plugging this estimate along with (A.46) back into (A.45), we obtain part (2) of Proposition 5, upon choosing

$$A_6 := \left(3\sqrt{\frac{2p_{\max}}{p_{\min}}} + 1\right)\sqrt{\frac{4}{a_3\sigma_K p_{\min}}}, \quad A_7 := 4A_1\frac{p_{\max}}{a_3 p_{\min}}.$$

A5 Bound on the empirical norm

In Lemma 13, we lower bound $\|f_0\|_n^2$ by (a constant times) the $L^2(\mathcal{X})$ norm of f.

Lemma 13. *Fix $\delta \in (0,1)$ Suppose P satisfies (P2). If $f \in H^1(\mathcal{X}, M)$ is lower bounded in $L^2(\mathcal{X})$ norm,*

$$\|f\|_{L^2(\mathcal{X})} \geq \frac{C_6 M}{\delta} \cdot \max\Big\{n^{-1/2}, n^{-1/d}\Big\}. \tag{A.47}$$

Then with probability at least $1 - 5\delta$,

$$\|f\|_n^2 \geq \delta \cdot \mathbb{E}\Big[\|f\|_n^2\Big]. \tag{A.48}$$

Proof of Lemma 13. In this proof, we will find it more convenient to deal with the parameterization $b = 1/\delta$. To establish (A.48), it is sufficient to show that

$$\mathbb{E}\big[\|f\|_n^4\big] \leq \left(1 + \frac{1}{b^2}\right) \cdot \big(\mathbb{E}\big[\|f\|_n^2\big]\big)^2;$$

then (A.48) follows from the Paley-Zygmund inequality (Lemma 17). Since $p \leq p_{\max}$ is uniformly bounded, we can relate $\mathbb{E}\big[\|f\|_n^4\big]$ to the $L^4(\mathcal{X})$-norm,

$$\mathbb{E}\big[\|f\|_n^4\big] = \frac{(n-1)}{n}\Big(\mathbb{E}\Big[\|f\|_n^2\Big]\Big)^2 + \frac{\mathbb{E}\Big[\big(f(X_1)\big)^4\Big]}{n} \leq \Big(\mathbb{E}\Big[\|f\|_n^2\Big]\Big)^2 + p_{\max}\frac{\|f\|_{L^4(\mathcal{X})}^4}{n}.$$

We will use the Sobolev inequalities as a tool to show that $\|f\|_{L^4(\mathcal{X})}/n \leq \big(\mathbb{E}[\|f\|_n^2]\big)^2/(b^2 p_{\max})$, whence the claim of the Lemma is shown. The nature of the inequalities we use depend on the value of d. In particular, we will use the following relationships between norms: for any $f \in H^1(\mathcal{X}; M)$,

$$\left.\begin{array}{ll} \sup_{x\in\mathcal{X}}|f(x)|, & d = 1 \\ \|f\|_{L^q(\mathcal{X})}, & d = 2, \text{ for all } 0 < q < \infty \\ \|f\|_{L^q(\mathcal{X})}, & d \geq 3, \text{ for all } 0 < q \leq 2d/(d-2) \end{array}\right\} \leq C_7 \cdot M.$$

(See Theorem 6 in Section 5.6.3 of Evans [2010] for a complete statement and proof of the various Sobolev inequalities.)

As a result, we divide our analysis into three cases: (i) the case where $d < 2$, (ii) the case where $d > 2$, and (iii) the borderline case $d = 2$.

Case 1: $d < 2$. The $L^4(\mathcal{X})$-norm of f can be bounded in terms of the $L^2(\mathcal{X})$ norm,

$$\|f\|^4_{L^4(\mathcal{X})} \leq \left(\sup_{x \in \mathcal{X}} |f(x)|\right)^2 \cdot \int_{\mathcal{X}} [f(x)]^2 \, dx \leq C_7^2 M^2 \cdot \|f\|^2_{L^2(X)}.$$

Since by assumption

$$\|f\|^2_{L^2(\mathcal{X})} \geq C_6^2 \cdot b^2 \cdot M^2 \cdot \frac{1}{n},$$

we have

$$p_{\max}\frac{\|f\|^4_{L^4(\mathcal{X})}}{n} \leq C_7^2 M^2 p_{\max} \cdot \frac{\|f\|^2_{L^2(\mathcal{X})}}{n} \leq \frac{C_7 p_{\max}}{C_6^2 b^2}\|f\|^4_{L^2(\mathcal{X})} \leq \frac{\mathbb{E}\big[\|f\|_n^2\big]}{b^2},$$

where the last inequality follows by taking $C_6 \geq C_7\sqrt{p_{\max}/p_{\min}}$.

Case 2: $d > 2$. Let $\theta = 2 - d/2$ and $q = 2d/(d-2)$. Noting that $4 = 2\theta + (1-\theta)q$, Lyapunov's inequality implies

$$\|f\|^4_{L^4(\mathcal{X})} \leq \|f\|^{2\theta}_{L^2(\mathcal{X})} \cdot \|f\|^{(1-\theta)q}_{L^q(\mathcal{X})} \leq \|f\|^4_{L^2(\mathcal{X})} \cdot \left(\frac{C_7\|f\|_{H^1(\mathcal{X})}}{\|f\|_{L^2(\mathcal{X})}}\right)^d.$$

By assumption, $\|f\|_{L^2(\mathcal{X})} \geq C_6 b \|f\|_{H^1(\mathcal{X})} n^{-1/d}$, and therefore

$$p_{\max}\frac{\|f\|^4_{L^4(\mathcal{X})}}{n} \leq \|f\|^4_{L^2(\mathcal{X})} p_{\max} \cdot \left(\frac{C_7\|f\|_{H^1(\mathcal{X})}}{n^{1/d}\|f\|_{L^2(\mathcal{X})}}\right)^d \leq \frac{C_7^d p_{\max}\|f\|^4_{L^2(\mathcal{X})}}{C_6^d b^d} \leq \frac{\mathbb{E}\big[\|f\|_n^2\big]}{b^2}.$$

where the last inequality follows by taking $C_6 \geq C_7(p_{\max}/p_{\min})^{1/d}$, and keeping in mind that $d > 2$ and $b \geq 1$.

Case 3: $d = 2$. Fix $t \in (1/2, 1)$, and suppose that

$$\|f\|_{L^2(\mathcal{X})} \geq \frac{C_6 M}{\delta} \cdot n^{-t/2}. \tag{A.49}$$

Putting $q = 2/(1-t)$, we have that $\|f\|_{L^q(\mathcal{X})} \leq C_7 \cdot M$, and it follows from derivations similar to those in Case 2 that $\|f\|_{L^4(\mathcal{X})}/n \leq \big(\mathbb{E}[\|f\|_n^2]\big)^2/(b^2 p_{\max})$ when $C_6 \geq C_7\sqrt{p_{\max}/p_{\min}}$.

Now, suppose $f \in L^4(\mathcal{X})$ satisfies (A.49) only when $t = 1$. For each $k = 1, 2, \ldots$ let $f_k := n^{1/(2k)} f$, so that each f_k satisfies (A.49) with respect to $t = 1 - 1/k$. Clearly $\|f_k - f\|_{L^4(\mathcal{X})} \to 0$ as $k \to \infty$, and therefore

$$\frac{1}{n}\|f\|_{L^4(\mathcal{X})} = \frac{1}{n}\lim_{k\to\infty}\|f_k\|_{L^4(\mathcal{X})} \leq \frac{1}{b^2 p_{\max}}\lim_{k\to\infty}\big(\mathbb{E}[\|f_k\|_n^2]\big)^2 = \frac{1}{b^2 p_{\max}}\big(\mathbb{E}[\|f\|_n^2]\big)^2.$$

This establishes the claim when $d = 2$, and completes the proof of Lemma 13.

A6 Graph functionals under the manifold hypothesis

In this section, we restate a few results of García Trillos et al. [2019], Calder and García Trillos [2019], which are analogous to Lemmas 1 and 2 but cover the case where $\mathcal{X}$ is an m-dimensional submanifold without boundary. As such, the results in this section will hold under the assumption (P3). We refer to García Trillos et al. [2019], Calder and García Trillos [2019] for the proofs of these results.

Proposition 6 follows from Lemma 5 of García Trillos et al. [2019] and Markov's inequality.

Proposition 6. *For any $f \in H^1(\mathcal{X})$, with probability at least $1 - \delta$,*

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

In Proposition 7, it is assumed that r, $\widetilde{\delta}$ and θ satisfy the following smallness conditions.

(S1)
$$n^{-1/m} < \widetilde{\delta} \leq \frac{1}{4}r \quad \text{and} \quad C(\theta + \widetilde{\delta}) \leq \frac{1}{2}p_{\min} \quad \text{and} \quad C_4\bigl(\log(n)/n\bigr)^{1/m} \leq r \leq \min\{c_4, 1\}.$$

Proposition 7 **(c.f Theorem 2.4 of Calder and García Trillos [2019]).** *With probability at least* $1 - Cn\exp(-cn\theta^2\widetilde{\delta}^m)$*, the following statement holds. For any* $k \in \mathbb{N}$ *such that*

$$\sqrt{\lambda_k(\Delta_P)}r + C(\theta + \widetilde{\delta}) \leq \frac{1}{2},$$

it holds that

$$nr^{m+2}\lambda_k(\Delta_P)\biggl(1 - C\Bigl(r(\sqrt{\lambda_k(\Delta_P)}+1) + \frac{\widetilde{\delta}}{r} + \theta\Bigr)\biggr) \leq \lambda_k(G_{n,r}) \leq nr^{m+2}\lambda_k(\Delta_P)\biggl(1 + C\Bigl(r(\sqrt{\lambda_k(\Delta_P)}+1) + \frac{\widetilde{\delta}}{r} + \theta\Bigr)\biggr).$$

Proposition 8 follows from Lemma 3.1 of Calder and García Trillos [2019], along with a union bound.

Proposition 8. *With probability at least* $1 - 2Cn\exp(-cp_{\max}nr^m)$*, it holds that*

$$D_{\max}(G_{n,r}) \leq Cnr^m.$$

Finally, we note that a Weyl's Law holds for Riemmanian manifolds without boundary, i.e.

$$\lambda_k(\Delta_P) \asymp k^{2/m}.$$

Put $B_{n,r}(k) := \min\{nr^{m+2}k^{2/m}, nr^m\}$. Following parallel steps to the proof of Lemma 2, one can derive from Propositions 7 and 8, and Weyl's Law, that with probability at least $1 - Cn\exp(-cnr^m)$,

$$cB_{n,r}(k) \leq \lambda_k \leq CB_{n,r}(k), \quad \text{for all } 2 \leq k \leq n. \tag{A.50}$$

A7 Proofs of main results

We are now in a position to prove Theorems 1-5, as well as a few other claims from our main text. In Section A7.1 we prove all of our results regarding estimation and in Section A7.2 we prove all of our results regarding testing; in Section A7.3, Lemmas 14 and 15, we provide some useful estimates on a particular pair of sums that appear repeatedly in our proofs. Throughout, it will be convenient for us to deal with the normalization $\widetilde{\rho} := \rho n r^{d+2}$. We note that in each of our Theorems, the prescribed choice of ρ will always result in $\widetilde{\rho} \leq 1$.

A7.1 Proof of estimation results

Proof of Theorem 1. We have shown that the inequalities (14) and (15) are satisfied with probability at least $1 - \delta - C_1 n\exp(-c_1nr^d)$, and throughout this proof we take as granted that both of these inequalities hold.

Now, set $\widetilde{\rho} = M^{-4/(2+d)}n^{-2/(2+d)}$ as prescribed in Theorem 1, and note that $\widetilde{\rho}^{-d/2} \leq n$ is implied by the assumption $M \leq n^{1/d}$. Therefore from (15) and Lemma 14, it follows that

$$\sum_{k=1}^{n}\biggl(\frac{1}{\rho\lambda_k + 1}\biggr)^2 \geq 1 + \frac{1}{C_3^2}\sum_{k=2}^{n}\biggl(\frac{1}{\widetilde{\rho}k^{2/d} + 1}\biggr)^2 \geq \frac{1}{8C_3^2}\widetilde{\rho}^{-d/2}.$$

As a result, by Lemma 5 along with (14) and (15), with probability at least $1 - \delta - C_1n\exp(-c_1nr^d) - \exp(-\widetilde{\rho}^{-d/2}/8C_3^2)$ it holds that,

$$\begin{aligned}\|\widehat{f} - f_0\|_n^2 &\leq \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{n}\sum_{k=2}^{n}\biggl(\frac{1}{c_3\widetilde{\rho}\min\{k^{2/d}, r^{-2}\} + 1}\biggr)^2 \\ &\leq \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{nc_3^2}\sum_{k=2}^{n}\biggl(\frac{1}{\widetilde{\rho}k^{2/d} + 1}\biggr)^2 + \frac{10r^4}{c_3^2\widetilde{\rho}^2}.\end{aligned} \tag{A.51}$$

The first term on the right hand side of (A.51) is a bias term, while the second, third, and fourth terms each contribute to the variance. Of these, under our assumptions the third term dominates, as we show momentarily. First, we use Lemma 14 to get an upper bound on this variance term,

$$\sum_{k=2}^{n}\left(\frac{1}{\widetilde{\rho}k^{2/d}+1}\right)^2 \leq 4\widetilde{\rho}^{-d/2}.$$

Then plugging this upper bound back into (A.51), we have that

$$\begin{aligned}\|\widehat{f}-f_0\|_n^2 &\leq \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{40\widetilde{\rho}^{-d/2}}{c_3^2 n} + \frac{10r^4}{c_3^2\widetilde{\rho}^2} \\ &= \left(\frac{C_2}{\delta} + \frac{40}{c_3^2}\right)M^{2d/(2+d)}n^{-2/(2+d)} + \frac{10}{n} + \frac{10}{c_3^2}r^4 M^{8/(2+d)}n^{4/(2+d)} \\ &\leq \left(\frac{C_2}{\delta} + \frac{50}{c_3^2}\right)M^{2d/(2+d)}n^{-2/(2+d)},\end{aligned}$$

with the last inequality following from (R1) and the assumption $M \geq n^{-1/2}$. This completes the proof of Theorem 1.

Proof of Theorem 2. We first establish that $\widehat{f}$ achieves nearly-optimal rates when $d = 4$, and then establish the claimed sub-optimal rates when $d > 4$.

Nearly-optimal rates when $d = 4$.

Continuing on from (A.51), from Lemma 14 we have that

$$\|\widehat{f}-f_0\|_n^2 \leq \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{nc_3^2\widetilde{\rho}^2} + \frac{10\log n}{nc_3^2\widetilde{\rho}^2} + \frac{10r^4}{c_3^2\widetilde{\rho}^2}.$$

Setting $r = (C_0\log(n)/n)^{1/4}$, we obtain

$$\|\widehat{f}-f_0\|_n^2 \leq \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{nc_3^2\widetilde{\rho}^2} + \frac{10\log n}{nc_3^2\widetilde{\rho}^2} + \frac{10C_0\log n}{nc_3^2\widetilde{\rho}^2},$$

and choosing $\widetilde{\rho} = M^{-2/3}(\log n/n)^{1/3}$ yields

$$\|\widehat{f}-f_0\|_n^2 \leq \left(\frac{C_2}{\delta} + \frac{20}{c_3^2} + \frac{10C_0}{c_3^2}\right)M^{4/3}\left(\frac{\log n}{n}\right)^{1/3} + \frac{10}{n}.$$

Suboptimal rates when $d > 4$.

Once again continuing on from (A.51), from Lemma 14 we have that

$$\|\widehat{f}-f_0\|_n^2 \leq \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{nc_3^2\widetilde{\rho}^{d/2}} + \frac{10}{n^{4/d}\widetilde{\rho}^2 c_3^2} + \frac{10r^4}{\widetilde{\rho}^2 c_3^2}.$$

Setting $r = (C_0\log n/n)^{1/d}$, we obtain

$$\|\widehat{f}-f_0\|_n^2 \leq \frac{C_2}{\delta}\widetilde{\rho}M^2 + \frac{10}{n} + \frac{10}{n\widetilde{\rho}^{d/2}c_3^2} + \frac{10}{n^{4/d}\widetilde{\rho}^2 c_3^2} + \frac{10C_0^{4/d}(\log n)^{4/d}}{n^{4/d}\widetilde{\rho}^2 c_3^2},$$

and choosing $\widetilde{\rho} = M^{-2/3}n^{-4/(3d)}$ yields

$$\|\widehat{f}-f_0\|_n^2 \leq \left(\frac{C_2}{\delta} + \frac{10}{c_3^2} + \frac{10C_0^{4/d}}{c_3^2}\right)M^{4/3}\left(\frac{\log n}{n^{1/3}}\right)^{4/d} + \frac{10}{c_3^{d/2}}M^{d/3}n^{-1/3} + \frac{10}{n}.$$

Bounds on $L^2(\mathcal{X})$ error under Lipschitz assumption. Let $V_1, \ldots, V_n$ denote the Voronoi tesselation of $\mathcal{X}$ with respect to $X_1, \ldots, X_n$. Extend $\widehat{f}$ over $\mathcal{X}$ by taking it piecewise constant over the Voronoi cells, i.e.

$$\widehat{f}(x) := \sum_{i=1}^{n} \widehat{f}_i \cdot \mathbf{1}\{x \in V_i\}.$$

Note that we are abusing notation slightly by also using $\widehat{f}$ to refer to this extension.

In Proposition 9, we establish that the out-of-sample error $\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}$ will not be too much larger than the in-sample error $\|\widehat{f} - f_0\|_n$.

Proposition 9. *Suppose f_0 satisfies $|f_0(x') - f_0(x)| \leq M\|x' - x\|$ for all $x', x \in \mathcal{X}$. Then for all n sufficiently large, with probability at least $1 - \delta$ it holds that*

$$\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}^2 \leq C \log(1/\delta)\Bigg(\log(n) \cdot \|\widehat{f} - f_0\|_n^2 + M^2\Big(\frac{\log n}{n}\Big)^{2/d}\Bigg).$$

Note that $n^{-2/d} \ll n^{-2/(2+d)}$. Therefore Proposition 9 together with Theorem 1 implies that with high probability, $\widehat{f}$ achieves the nearly-optimal (up to a factor of $\log n$) estimation rates out-of-sample error—that is, $\|\widehat{f} - f_0\|_{L^2(\mathcal{X})}^2 \leq C\log(n)M^{2d/(2+d)}n^{-2/(2+d)}$—as long as $M \leq Cn^{1/d}$.

Proof of Proposition 9. Suppose $x \in V_i$, so that we can upper bound the pointwise squared error $|\widehat{f}(x) - f(x)|^2$ using the triangle inequality:

$$\big(\widehat{f}(x) - f_0(x)\big)^2 = \big(\widehat{f}(X_i) - f_0(x)\big)^2 \leq 2\big(\widehat{f}(X_i) - f_0(X_i)\big)^2 + 2\big(f_0(X_i) - f_0(x)\big)^2.$$

Integrating both sides of the inequality, we have

$$\begin{aligned}\int_{\mathcal{X}} \big(\widehat{f}(x) - f_0(x)\big)^2\,dx &\leq 2\sum_{i=1}^{n}\int_{V_i}\Big(\widehat{f}(X_i) - f_0(X_i)\Big)^2\,dx + 2\sum_{i=1}^{n}\int_{V_i}\Big(f_0(X_i) - f_0(x)\Big)^2\,dx \\ &= 2\sum_{i=1}^{n}\mathrm{vol}(V_i)\Big(\widehat{f}(X_i) - f_0(X_i)\Big)^2 + 2\sum_{i=1}^{n}\int_{V_i}\Big(f_0(X_i) - f_0(x)\Big)^2\,dx,\end{aligned}$$

and so by invoking the Lipschitz property of f_0, we obtain

$$\|\widehat{f} - f\|_{L^2(\mathcal{X})}^2 \leq 2\sum_{i=1}^{n}\mathrm{vol}(V_i)\Big(\widehat{f}(X_i) - f_0(X_i)\Big)^2 + 2M^2\sum_{i=1}^{n}\Big(\mathrm{diam}(V_i)\Big)^2. \tag{A.52}$$

Here we have written $\mathrm{diam}(V)$ for the diameter of a set V.

Now we will use some results of Chaudhuri and Dasgupta [2010] regarding uniform concentration of empirical counts, to upper bound $\mathrm{diam}(V_i)$ Set

$$\varepsilon_n := \left(\frac{2C_o\log(1/\delta)d\log n}{\nu_d p_{\min} a_3 n}\right)^{1/d},$$

where C_o is a constant given in Lemma 16 of Chaudhuri and Dasgupta [2010]. Note that for n sufficiently large, $\varepsilon_n \leq c_0$, and therefore by (A.35) we have that for every $x \in \mathcal{X}$, $P(B(x,\varepsilon_n)) \geq 2C_o\log(1/\delta)d\frac{\log n}{n}$. Consequently, by Lemma 16 of Chaudhuri and Dasgupta [2010] it holds that with probability at least $1 - \delta$,

$$\text{for all } x \in \mathcal{X}, \quad B(x,\varepsilon_n) \cap \{X_1, \ldots, X_n\} \neq \emptyset. \tag{A.53}$$

But if (A.53) is true, it must also be true that for each $i = 1, \ldots, n$ and for every $x \in V_i$, the distance $\|x - X_i\| \leq \varepsilon_n$. Thus by the triangle inequality, $\max_{i=1,\ldots,n}\mathrm{diam}(V_i) \leq 2\varepsilon_n$. Plugging back in to (A.52), and using the upper bound volume $\mathrm{vol}(V_i) \leq \nu_d\big(\mathrm{diam}(V_i)\big)^d$, we obtain the desired upper bound on $\|\widehat{f} - f\|_{L^2(\mathcal{X})}^2$.

Proof of Theorem 4. The proof of Theorem 4 follows exactly the same steps as the proof of Theorem 1, replacing the references to Lemma 1 and 2 by references to Proposition 6 and (A.50), and the ambient dimension d by the intrinsic dimension m.

A7.2 Proofs of testing results

Proof of Theorem 3. Let $\delta = 1/b$. Recall that we have shown that the inequalities (14) and (15) are satisfied with probability at least $1 - 1/b - C_1 n \exp(-c_1 n r^d)$, and throughout this proof we take as granted that both of these inequalities hold.

Now, we would like to invoke Lemma 6, and in order to do so, we must show that the inequality (A.14) is satisfied with respect to $G = G_{n,r}$. First, we upper bound the right hand side of this inequality. Setting $\widetilde{\rho} = M^{-8/(4+d)} n^{-4/(4+d)}$ as prescribed by Theorem 3, it follows from (14) and (15) that

$$\begin{aligned}\frac{2\rho}{n}\big(f_0^\top L f_0\big) + \frac{2\sqrt{2/\alpha}+2b}{n}\left(\sum_{k=1}^{n}\frac{1}{(\rho\lambda_k+1)^4}\right)^{1/2} &\le C_2 b\widetilde{\rho}M^2 + \frac{2\sqrt{2/\alpha}+2b}{n}\left[1 + \frac{1}{c_3^2}\left(\sum_{k=2}^{n}\frac{1}{(\widetilde{\rho}k^{2/d}+1)^4}\right)^{1/2} + \frac{r^4 n^{1/2}}{c_3^2\widetilde{\rho}^2}\right] \\ &\le C_2 b\widetilde{\rho}M^2 + \frac{2\sqrt{2/\alpha}+2b}{n}\left(1 + \frac{\sqrt{2}}{c_3^2}\widetilde{\rho}^{-d/4} + \frac{r^4 n^{1/2}}{c_3^2\widetilde{\rho}^2}\right) \\ &\le \left(C_2 + 2 + \frac{2\sqrt{2}}{c_3^2} + \frac{2}{c_3^2}\right)\cdot\left(\sqrt{\frac{2}{\alpha}} + b\right)\cdot M^{2d/(4+d)} n^{-4/(4+d)}.\end{aligned}$$

The second inequality in the above is justified by Lemma 15, keeping in mind that $M \le M_{\max}(d)$ implies that $\widetilde{\rho}^{-d/2} \le n$. The third inequality follows from the upper bound on r assumed in (R2) as well as the fact that $M \ge n^{-1/2}$.

Next we lower bound the left hand side of the inequality (A.14)—i.e. we lower bound the empirical norm $\|f_0\|_n^2$—using Lemma 13. Recall that by assumption, $M \le M_{\max}(d)$. Therefore, taking $C \ge C_6$ in (11) implies that the lower bound on $\|f\|_{L^2(\mathcal{X})}$ in (A.47) is satisfied. As a result, it follows from (A.48) that

$$\|f\|_n^2 \ge \frac{\mathbb{E}[\|f\|_n^2]}{b} \ge \frac{p_{\min}}{b}\|f\|_{L^2(\mathcal{X})}^2 \ge C\left(\sqrt{\frac{1}{\alpha}} + b\right) M^{2d/(4+d)} n^{-4/(4+d)},$$

with probability at least $1 - 5/b$. Taking $C \ge C_2 + 2 + (2\sqrt{2})/c_3^2 + 2/c_3^2$ in (11) thus implies (A.14), and we may therefore use Lemma 6 to upper bound the type II error the Laplacian smoothing test $\widehat{\varphi}$. Observe that by (15) and the lower bound in Lemma 15,

$$\sum_{k=1}^{n}\left(\frac{1}{\rho\lambda_k+1}\right)^4 \ge 1 + \frac{1}{C_3^4}\sum_{k=2}^{n}\left(\frac{1}{\widetilde{\rho}k^{2/d}+1}\right)^4 \ge \frac{1}{32C_3^4}\widetilde{\rho}^{-d/2}.$$

We conclude that

$$\begin{aligned}\mathbb{P}_{f_0}\big(\widehat{T} \le \widehat{t}_\alpha\big) &\le \frac{6}{b} + \frac{1}{b^2} + \frac{16}{b}\left(\sum_{k=1}^{n}\frac{1}{(\rho\lambda_k+1)^4}\right)^{-1/2} + C_1 n\exp(-c_1 n r^d) \\ &\le \frac{7}{b} + \frac{64\sqrt{2}}{b}C_3^2\widetilde{\rho}^{d/4} + C_1 n\exp(-c_1 n r^d),\end{aligned}$$

establishing the claim of Theorem 3.

Proof of Theorem 5. The proof of Theorem 5 follows exactly the same steps as the proof of Theorem 3, replacing the references to Lemma 1 and 2 by references to Propositions 6 and (A.50), and the ambient dimension d by the intrinsic dimension m.

Proof of (12). When $\rho = 0$, the Laplacian smoother $\widehat{f} = \mathbf{Y}$, the test statistic $\widehat{T} = \frac{1}{n}\|\mathbf{Y}\|_2^2$, and the threshold $\widehat{t}_\alpha = 1 + n^{-1/2}\sqrt{2/\alpha}$. The expectation of $\widehat{T}$ is

$$\mathbb{E}\big[\widehat{T}\big] = \mathbb{E}\big[f_0^2(X)\big] + 1 \ge p_{\min}\|f_0\|_{L^2(\mathcal{X})}^2 + 1.$$

When $f_0 \in L^4(\mathcal{X}, M)$, the variance can be upper bounded

$$\mathrm{Var}\big[\widehat{T}\big] \le \frac{1}{n}\Big(3 + p_{\max} M^4 + p_{\max}\|f_0\|_{L^2(\mathcal{X})}^2\Big).$$

Now, let us assume that

$$\|f_0\|_{L^2(X)}^2 \ge \frac{2\sqrt{2/\alpha} + 2b}{p_{\min}} n^{-1/2},$$

so that $E[\widehat{T}] - \widehat{t}_\alpha \ge E[f_0^2(X)]/2$. Hence, by Chebyshev's inequality

$$\begin{aligned}
\mathbb{P}_{f_0}\Big(\widehat{T} \le \widehat{t}_\alpha\Big) &\le 4\frac{\mathrm{Var}_{f_0}\big[\widehat{T}\big]}{\mathbb{E}[f_0^2(X)]^2} \\
&\le \frac{4}{n} \cdot \frac{3 + p_{\max}\big(M^4 + \|f_0\|_{L^2(\mathcal{X})}^2\big)}{p_{\min}^2\|f_0\|_{L^2(\mathcal{X})}^4} \\
&\le \frac{1}{b^2}\Big(3 + \frac{4bp_{\max}}{p_{\min}n^{1/2}} + p_{\max}M^4\Big).
\end{aligned}$$

A7.3 Two convenient estimates

The following Lemmas provides convenient upper and lower bounds on our estimation variance term (Lemma 14) and testing variance term (Lemma 15).

Lemma 14. *For any $t > 0$ such that $1 \le t^{-d/2} \le n$,*

$$\frac{1}{8}t^{-d/2} - 1 \le \sum_{k=2}^{n}\Big(\frac{1}{tk^{2/d}+1}\Big)^2 \le t^{-d/2} + \begin{cases} 3t^{-d/2}, & \text{if } d < 4 \\ \frac{1}{t^2}\log n, & \text{if } d = 4 \\ \frac{1}{t^2}n^{1-4/d}, & \text{if } d > 4. \end{cases}$$

Lemma 15. *Suppose $d \le 4$. Then for any $t > 0$ such that $1 \le t^{-d/2} \le n$,*

$$\frac{1}{32}t^{-d/2} - 1 \le \sum_{k=2}^{n}\Big(\frac{1}{tk^{2/d}+1}\Big)^4 \le 2t^{-d/2}.$$

Proof of Lemma 14. We begin by proving the upper bounds. Treating the sum over k as a Riemann sum of a non-increasing function, we have that

$$\sum_{k=2}^{n}\Big(\frac{1}{tk^{2/d}+1}\Big)^2 \le \int_1^n \Big(\frac{1}{tx^{2/d}+1}\Big)^2 dx \le t^{-d/2} + \int_{t^{-d/2}}^n \Big(\frac{1}{tx^{2/d}+1}\Big)^2 dx \le t^{-d/2} + \frac{1}{t^2}\int_{t^{-d/2}}^n x^{-4/d}\,dx.$$

The various upper bounds (for $d < 4$, $d = 4$, and $d > 4$) then follow upon computing the integral.

For the lower bound, we simply recognize that for each $k = 2, \ldots, n$ such that $k \le \lfloor t^{-d/2}\rfloor$, it holds that $1/(tk^{2/d}+1)^2 \ge 1/4$, and there are at least $\min\Big\{\lfloor t^{-d/2}\rfloor - 1, n-1\Big\} > \frac{1}{2}t^{-d/2} - 1$ such values of k.

Proof of Lemma 15. The upper bound follows similarly to that of Lemma 14:

$$\sum_{k=1}^{n}\Big(\frac{1}{tk^{2/d}+1}\Big)^4 \le t^{-d/2} + \frac{1}{t^4}\sum_{k=t^{-d/2}+1}^{n}\frac{1}{k^{8/d}} \le t^{-d/2} + \frac{1}{t^4}\int_{t^{-d/2}}^n x^{-8/d}\,dx \le 2t^{-d/2}.$$

The lower bound follows from the same logic as we used to derive the lower bound in Lemma 14.

A8 Concentration inequalities

Lemma 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 $t > 0$,*

$$\mathbb{P}\Big[U \geq 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\Big] \leq \exp(-t).$$

In particular if $a_k = 1$ for each $k = 1, \ldots, N$, then

$$\mathbb{P}\Big[U \geq 2\sqrt{Nt} + 2t\Big] \leq \exp(-t).$$

The proof of Lemma 13 relies on (a variant of) the Paley-Zygmund Inequality.

Lemma 17. *Let f satisfy the following moment inequality for some $b \geq 1$:*

$$\mathbb{E}\big[\|f\|_n^4\big] \leq \left(1 + \frac{1}{b^2}\right) \cdot \left(\mathbb{E}\big[\|f\|_n^2\big]\right)^2. \tag{A.54}$$

Then,

$$\mathbb{P}\left[\|f\|_n^2 \geq \frac{1}{b}\mathbb{E}\big[\|f\|_n^2\big]\right] \geq 1 - \frac{5}{b}. \tag{A.55}$$

Proof. Let Z be a non-negative random variable such that $\mathbb{E}(Z^q) < \infty$. The Paley-Zygmund inequality says that for all $0 \leq \lambda \leq 1$,

$$\mathbb{P}(Z > \lambda\mathbb{E}(Z^p)) \geq \left[(1 - \lambda^p)\frac{\mathbb{E}(Z^p)}{(\mathbb{E}(Z^q))^{p/q}}\right]^{\frac{q}{q-p}}. \tag{A.56}$$

Applying (A.56) with $Z = \|f\|_n^2$, $p = 1$, $q = 2$ and $\lambda = \frac{1}{b}$, by assumption (A.54) we have

$$\mathbb{P}\Big(\|f\|_n^2 > \frac{1}{b}\mathbb{E}[\|f\|_n^2]\Big) \geq \Big(1 - \frac{1}{b}\Big)^2 \cdot \frac{\left(\mathbb{E}[\|f\|_n^2]\right)^2}{\mathbb{E}[\|f\|_n^4]} \geq \frac{\left(1 - \frac{2}{b}\right)}{\left(1 + \frac{1}{b^2}\right)} \geq 1 - \frac{5}{b}.$$

□

Let $Z_1, \ldots, Z_n$ be independently distributed and bounded random variables, such that $\mathbb{E}[Z_i] = \mu_i$. Let $S_n = Z_1 + \ldots + Z_n$ and $\mu = \mu_1 + \ldots + \mu_n$. The multiplicative form of Hoeffding's inequality gives sharp bounds when $\mu \ll 1$.

Lemma 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*

$$\mathbb{P}\Big(\Big|S_n - \mu\Big| \geq \delta\mu\Big) \leq 2\exp\Big(-\frac{\delta^2\mu}{3B^2}\Big).$$

We use Lemma 18, along with properties of the kernel K and density p, to upper bound the maximum degree in our neighborhood graph, which we denote by $D_{\max}(G_{n,r}) := \max_{i=1,\ldots,n} D_{ii}$.

Lemma 19. *Under the conditions of Lemma 2,*

$$D_{\max}(G_{n,r}) \leq 2p_{\max}nr^d,$$

with probability at least $1 - 2n\exp\Big(-nr^da_3p_{\min}/(3[K(0)]^2)\Big)$.

Proof of Lemma 19. Fix $x \in \mathcal{X}$, and set

$$D_{n,r}(x) := \sum_{i=1}^{n} K\biggl(\frac{\|X_i - x\|}{r}\biggr);$$

note that $D_{n,r}(X_i)$ is just the degree of X_i in $G_{n,r}$. By Hoeffding's inequality

$$\mathbb{P}\Bigl(\Bigl|D_{n,r}(x) - \mathbb{E}\bigl[D_{n,r}(x)\bigr]\Bigr| \geq \delta \mathbb{E}\bigl[D_{n,r}(x)\bigr]\Bigr) \leq 2\exp\biggl(-\frac{\delta^2 \mathbb{E}\bigl[D_{n,r}(x)\bigr]}{3[K(0)]^2}\biggr). \tag{A.57}$$

Now we lower bound $\mathbb{E}[D_{n,r}(x)]$ using the boundedness of the density p, and the fact that $\mathcal{X}$ has Lipschitz boundary:

$$\begin{aligned}
\mathbb{E}\bigl[D_{n,r}(x)\bigr] &= n \int_{\mathcal{X}} K\biggl(\frac{\|x' - x\|}{r}\biggr) p(x)\,dx \\
&\geq n p_{\min} \int_{\mathcal{X}} K\biggl(\frac{\|x' - x\|}{r}\biggr)\,dx \\
&\geq n p_{\min} a_3 \int_{\mathcal{X}} K\biggl(\frac{\|x' - x\|}{r}\biggr)\,dx \\
&\geq n r^d p_{\min},
\end{aligned}$$

with the second inequality following from (A.35), and the final inequality from the normalization $\int_{\mathbb{R}^d} K(\|z\|)\,dz = 1$. Similar derivations yield the upper bound

$$\mathbb{E}\bigl[D_{n,r}(x)\bigr] \leq n r^d p_{\max},$$

and plugging these bounds in to (A.57), we determine that

$$\mathbb{P}\Bigl(D_{n,r}(x) \geq (1+\delta) n r^d p_{\max}\Bigr) \leq 2\exp\biggl(-\frac{\delta^2 n r^d a_0 p_{\min}}{3[K(0)]^2}\biggr).$$

Applying a union bound, we get that

$$\mathbb{P}\Bigl(\max_{i=1,\ldots,n} D_{n,r}(X_i) \geq (1+\delta) n r^d p_{\max}\Bigr) \leq 2n\exp\biggl(-\frac{\delta^2 n r^d a_0 p_{\min}}{3[K(0)]^2}\biggr),$$

and taking $\delta = 1$ gives the claimed upper bound.

REFERENCES

Mikhail Belkin, Qichao Que, Yusu Wang, and Xueyuan Zhou. Toward understanding complex spaces: Graph laplacians on manifolds with singularities and boundaries. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 36.1–36.26, Edinburgh, Scotland, 25–27 Jun 2012. JMLR Workshop and Conference Proceedings.

Dmitri Burago, Sergei Ivanov, and Yaroslav Kurylev. A graph discretization of the Laplace-Beltrami operator. *Journal of Spectral Theory*, 4(4):675–714, 2014.

Jeff Calder and Nicolás García Trillos. Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs. *arXiv preprint arXiv:1910.13476*, 2019.

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

Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. American Mathematical Soc., 1997.

Matthew M Dunlop, Dejan Slepčev, Andrew M Stuart, and Matthew Thorpe. Large data and zero noise limits of graph-based semi-supervised learning algorithms. *Applied and Computational Harmonic Analysis*, 49(2): 655–697, 2020.

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

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

Nicolás García Trillos and Dejan Slepčev. On the rate of convergence of empirical measures in infinity-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

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

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepcev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*, 20:1–61, 2019.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, volume 29, 2016.