Adaptive and non-adaptive minimax rates for weighted Laplacian-Eigenmap based nonparametric regression

Zhaoyang Shi
University of California, Davis,
zysshi@ucdavis.edu

Krishnakumar Balasubramanian
University of California, Davis,
kbala@ucdavis.edu

Wolfgang Polonik
University of California, Davis,
wpolonik@ucdavis.edu

Abstract

We show both adaptive and non-adaptive minimax rates of convergence for a family of weighted Laplacian-Eigenmap based nonparametric regression methods, when the true regression function belongs to a Sobolev space and the sampling density is bounded from above and below. The adaptation methodology is based on extensions of Lepski's method and is over both the smoothness parameter ($s \in \mathbb{N}_+$) and the norm parameter ($M > 0$) determining the constraints on the Sobolev space. Our results extend the non-adaptive result in Green et al. (2023), established for a specific normalized graph Laplacian, to a wide class of weighted Laplacian matrices used in practice, including the unnormalized Laplacian and random walk Laplacian.

1 INTRODUCTION

Consider the following regression model,

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is the true regression function, $X_i \overset{\text{i.i.d.}}{\sim} g$, where g is a density on $\mathcal{X} \subset \mathbb{R}^d$, and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ is the noise (independent of the X_i's). The goal is to estimate the regression function f given pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. Our main contribution in this work is to develop non-adaptive and adaptive estimators that achieve minimax optimal estimation rates, when f lies in Sobolev spaces.

The estimators we study are based on performing principal components regression using the estimated eigenfunctions of a family of weighted Graph Laplacian operators. Various versions of Graph Laplacian matrices have been considered in the literature. Recently, Hoffmann et al. (2022) proposed a unifying framework describing a family of Graph Laplacian matrices, parametrized by $w \in \mathbb{R}^3$; see (2) and (3) for details. This captures Laplacian matrices used widely in practice, including the normalized, unnormalized and the random walk Laplacian.

Green et al. (2023) analyzed principal components regression specifically using unnormalized graph Laplacian matrices constructed over ϵ-graphs, and established non-adaptive minimax rates when f lies in Sobolev spaces. In this paper, we first extend this result to the entire family of weighted Laplacian matrices from (2) and (3); Theorem 3.1. These results are established by assuming a sampling density bounded from above and below and a true regression function belonging to a Sobolev space.

Note that technically, the weighted Laplacian matrices correspond to a family of weighted Sobolev spaces which all become equivalent under the above-mentioned boundedness assumption on the sampling density. However, the parameters of the corresponding Sobolev spaces, in particular the smoothness parameter ($s \in \mathbb{N}_+$) and the norm parameter $M > 0$ determining the constraints on the Sobolev space, both change on w.

While the minimax rate optimal non-adaptive estimator depends on the knowledge of the smoothness and norm parameters of the true regression function, these parameters are unknown in practice. Tuning parameters, such as ϵ, the graph radius (or the bandwidth for the kernel) and K, the number of eigenvectors considered, require knowledge of the smoothness and the norm parameters. Hence, in order to apply the Laplacian-based regression methodology in practice, we develop an adaptive estimator, based on Lepski's method, and show in Theorem 3.2 that the developed estimator achieves minimax rates (up to log factors) without requiring the knowledge of either the smooth-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

ness or the norm parameters.

The main technical contributions we make in this work towards establishing the aforementioned both adaptive and non-adaptive results include the following:

- As a part of the proofs of our main results in Theorem 3.1, we rigorously prove the idea roughly outlined by Hoffmann et al. (2022) on showing the convergence of the discrete weighted graph Laplacian matrices to their continuum counterparts (in appropriately well-defined sense) by leveraging the concentration result established by Giné and Guillou (2002) for kernel density estimators.
- We generalize the convergence property of the eigenvalues of the Laplacian matrices in Calder and Trillos (2022) to the weighted Laplacian matrices by providing an analogous bound for the eigenvalues combined with Weyl's law.
- We formulate a simultaneous two-parameter Lepski's procedure and obtain the adaptive minimax rate (see Theorem 3.2) through deriving a high-order-moment-based concentration inequality of the weighted Sobolev semi-norm.

Our contributions not only highlight the significance of utilizing the weighted graph Laplacians for nonparametric regression but also establish a solid statistical foundation for this method, offering a robust framework that underpins the reliability and effectiveness of this approach.

1.1 Related Works

Graph Laplacians are widely used in many data science problems for feature learning and spectral clustering (Weiss, 1999, Shi and Malik, 2000, Ng et al., 2001, von Luxburg, 2007), extracting heat kernel signatures for shape analysis (Sun et al., 2009, Andreux et al., 2015, Dunson et al., 2021), reinforcement learning (Mahadevan and Maggioni, 2007, Wu et al., 2019) and dimensionality reduction (Belkin and Niyogi, 2003, Coifman and Lafon, 2006), among other applications. There is an ever-growing literature on further applications of graph Laplacian in data science topic, and we also refer to Belkin et al. (2006), Wang et al. (2015), Chun et al. (2016) for more discussions.

As mentioned above, we consider the application of the weighted graph Laplacian for achieving minimax optimal rates in nonparametric regression. Other works focusing on this problem (including the semi-supervised setting) include Green et al. (2021) and Green et al. (2023) using unnormalized Laplacian based on the Laplacian eigenmaps (see Belkin and Niyogi, 2003), Bousquet et al. (2003) with Laplacian smoothing, Rice (1984) adopting spectral series regression on the Sobolev spaces, Trillos et al. (2022) applying the graph Poly-Laplacians (see Remark 3.4 for specific comparison to this method) and Hacquard et al. (2022) using topological data analysis. We also refer to Zhu et al. (2003), Zhou and Srebro (2011), Lee and Izbicki (2016), Dicker et al. (2017) and Garcia Trillos and Murray (2020) for related analysis in the context of regression problems.

In recent years, there has been a great deal of progress on obtaining theoretical rates of convergence in the context of Laplacian operator estimation and related eigenvalue and/or eigenfunction estimation. Early work on consistency of graph Laplacians focused on pointwise consistency results for ϵ-graphs, see Belkin and Niyogi (2005), Hein et al. (2005), Giné and Koltchinskii (2006), Hein et al. (2007) and references therein for more details. For fixed neighborhood size ϵ, von Luxburg et al. (2008) and Rosasco et al. (2010) considered spectral convergence of graph Laplacians. Furthermore, Trillos and Slepčev (2018) established conditions on connectivity for the above spectral convergence with no specific error estimates. Later on, the convergence of Laplacian matrices to Laplacian operators has been considered, includes unnormalized, random walk Laplacians and k-NN graph based Laplacians in various work including Shi (2015), Calder and Trillos (2022). There, rates of convergences of Laplacian eigenvalues and eigenvectors to population counterparts with explicit error estimates are derived. Following the above literature, Hoffmann et al. (2022) developed a framework for extending the above convergence results to a general Laplacian family, the weighted Laplacians, and presented some heuristic asymptotic analysis.

To the best of our knowledge, the only work that considers adaptivity in the context of Laplacian estimation is Chazal et al. (2016). They use Lepski's method for adaptive estimation of the unnormalized Laplace-Beltrami operators, focusing on bandwidth parameters. Also, they adopted a more flexible version of Lepski's method introduced in Lacour and Massart (2016) that involves certain multiplicative coefficients introduced in the variance and bias terms to develop the method. Therefore, their proof technique is to consider the trade-off between the bounds on the approximation error and the variance of Laplacian estimators. However, in this paper, we apply Lepski's method in the context of regression problem by using weighted Laplacians instead of just the unnormalized Laplacians (as in Chazal et al. (2016)). Additionally, besides the bandwidth parameter, our method is also adaptive to the smoothness parameter and the norm

parameter of the Sobolev space under consideration, i.e., in our work, we use Lepski's method for *simultaneous* adaptation to the unknown parameters of the function class under consideration.

2 PRELIMINARIES

In this section, we first describe the data-based weighted graph Laplacian matrices, and the corresponding nonparametric regression estimator. We then introduce the associated limiting operators and the weighted Sobolev spaces.

2.1 Weighted Graph Laplacian Matrices

Given i.i.d data $X_1, \ldots, X_n$ from a distribution G on $\mathcal{X} \subseteq \mathbb{R}^d$ with the density g, consider a graph G with vertex set $\{X_1, \ldots, X_n\}$ and adjacency matrix $\tilde{W}$ given by

$$\tilde{w}_{i,j}^{\epsilon} := \frac{1}{n\epsilon^d}\eta\left(\frac{\|X_i - X_j\|}{\epsilon}\right), \quad i,j = 1,\ldots,n, \tag{2}$$

where $\|\cdot\|$ denotes the standard Euclidean norm. Here $\eta \geq 0$ is a kernel function with support $[0,1]$, and ε is the bandwidth parameter. In other words, G is constructed by placing an edge $X_i \sim X_j$, when $\|X_i - X_j\| \leq \epsilon$, and this edge is given the weight $\tilde{w}_{i,j}^{\epsilon}$. The term $(n\epsilon^d)^{-1}$ is a convenient normalization factor. The degree matrix is then given by a diagonal matrix $\tilde{D}$ with the i-th diagonal element as

$$\tilde{d}_i := \sum_{j=1}^{n} \tilde{w}_{i,j}^{\epsilon}, \quad i = 1,\ldots,n,$$

which can also be thought of as the kernel density estimator (KDE) of the density g at X_i.

The weighted graph Laplacian matrices are a family of graph Laplacians consisting of various types of normalizations characterized by a parameter $w = (p,q,r) \in \mathbb{R}^3$ constructed as follows. First define a re-weighted adjacency matrix W with (i,j)-th element as

$$w_{i,j}^{\epsilon} := \frac{\tilde{w}_{i,j}^{\epsilon}}{\tilde{d}_i^{1-\frac{q}{2}}\tilde{d}_j^{1-\frac{q}{2}}}, \quad i,j = 1,\ldots,n,$$

so that the corresponding diagonal degree matrix D as entries

$$d_i := \sum_{j=1}^{n} w_{i,j}^{\epsilon}, \quad i = 1,\ldots,n.$$

Then, the weighted graph Laplacian after re-weighting is defined in Hoffmann et al. (2022) as follows: for a tuple $w = (p,q,r) \in \mathbb{R}^3$,

$$L_{w,n,\epsilon} := \begin{cases} \frac{1}{\epsilon^2} D^{\frac{1-p}{q-1}}(D-W)D^{-\frac{r}{q-1}}, & \text{if } q \neq 1, \\ \frac{1}{\epsilon^2}(D-W), & \text{if } q = 1, \end{cases} \tag{3}$$

where $1/\epsilon^2$ is also a normalization factor. For $u \in \mathbb{R}^n$, the i-th coordinate of the vector $L_{w,n,\epsilon}u$ is given by

$$(L_{w,n,\epsilon}u)_i = \frac{1}{\epsilon^2}\sum_{j=1}^{n} d_i^{\frac{1-p}{q-1}} w_{i,j}^{\epsilon}\left(d_i^{-\frac{r}{q-1}}u_i - d_j^{-\frac{r}{q-1}}u_j\right). \tag{4}$$

The above weighted graph Laplacian (3) generalizes many commonly used graph Laplacian. For $(p,q,r) = (1,2,0)$, it recovers the unnormalized graph Laplacian L_u; if $(p,q,r) = (3/2,2,1/2)$, it gives the normalized graph Laplacian L_n; if $(p,q,r) = (2,2,0)$, it corresponds to a non-symmetric matrix but can be interpreted as a transition probability of a random walk on a graph denoted by L_r:

$$\begin{aligned} L_u &:= D - W, \\ L_n &:= D^{-1/2}(D-W)D^{1/2}, \\ L_r &:= D^{-1}(D-W), \end{aligned}$$

While the main focus is on ϵ-graphs, we highlight that the above formulation also captures the limits of graphs constructed based on the k-nearest neighbor graphs. In particular, when $(p,q,r) = (1, 1-2/d, 0)$, one can call the related normalization as the *near k-NN* normalization; see Calder and Trillos (2022) and Hoffmann et al. (2022) for details.

Note that the weighted Laplacian matrix $L_{w,n,\epsilon}$ is actually not self-adjoint with respect to the Euclidean inner product $\langle\cdot,\cdot\rangle$ since it is in general not symmetric. However, it is self-adjoint with respect to the following weighted inner product $\langle\cdot,\cdot\rangle_{g^{p-r}}$:

$$\langle\cdot,\cdot\rangle_{g^{p-r}} := \begin{cases} \langle\cdot,\cdot\rangle_{D^{\frac{p-1-r}{q-1}}} & \text{if } q \neq 1, \\ \langle\cdot,\cdot\rangle & \text{if } q = 1, \end{cases}$$

where for a given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and vectors $u, v \in \mathbb{R}^n$, define

$$\langle u, v\rangle_A := u^T A v.$$

We also define the normalized weighted inner product: $\langle\cdot,\cdot\rangle_{w,n} := n^{-1}\langle\cdot,\cdot\rangle_{g^{p-r}}$ and the normalized Euclidean inner product: $\langle\cdot,\cdot\rangle_n := n^{-1}\langle\cdot,\cdot\rangle$ and denote by $\|\cdot\|_{w,n}$ and $\|\cdot\|_n$ their respective corresponding norms. Here, our estimation results are measured in $\|\cdot\|_{w,n}$ and under our assumptions in Section 3.1, it can be shown to be equivalent to the classic norm $\|\cdot\|_n$.

2.2 Weighted Laplacian-Eigenmap Based Nonparametric Regression

Following the ideas in Belkin and Niyogi (2003) and Green et al. (2023), we propose the following principal components regression with the weighted Laplacian eigenmaps (PCR-WLE) algorithm:

(1) For a given parameter $\epsilon > 0$ and a kernel function η, construct the ϵ-graph according to Section 2.1.

(2) Construct the weighted Laplacian matrix given by (3) and take its eigendecomposition $L_{w,n,\epsilon} = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ with respect to $\langle \cdot, \cdot \rangle_{w,n}$, where (λ_i, v_i) are the eigenpairs with eigenvalues $0 = \lambda_1 \leq \ldots \leq \lambda_n$ in an ascending order and eigenvectors normalized to satisfy $\|v_i\|_{w,n} = 1$.

(3) Project the response vector $Y = (Y_1, \ldots, Y_n)^T$ onto the space spanned by the first K eigenvectors, i.e., denote by $V_K \in \mathbb{R}^{n \times K}$ the matrix with j-th column as $V_{K,j} = v_j$ for $j = 1, \ldots, K$ and define

$$\hat{f} := V_K V_K^T Y,$$

as the estimator.

The entries of the vector $\hat{f}$ are the in-sample values of the estimator of the regression function f. Green et al. (2023) considered the special case of the above approach for the case when $(p, q, r) = (1, 2, 0)$ corresponding to the unnormalized graph Laplacian. Here, we consider the entire family of graph Laplacians for various choices of the parameters (p, q, r), the generalization from Hoffmann et al. (2022).

2.3 Weighted Laplacians And Weighted Sobolev Spaces

Hoffmann et al. (2022) showed a heuristic framework for the convergence of the weighted graph Laplacian $L_{w,n,\epsilon}$ defined in (3) to the following weighted Laplace-Beltrami operators, in the large sample limit, in terms of the eigenvalues and eigenvectors/eigenfunctions:

$$\begin{cases} \mathcal{L}_w u := -\dfrac{1}{2g^p} \operatorname{div}\left(g^q \nabla\left(\dfrac{u}{g^r}\right)\right), & \text{in } \mathcal{X}, \\ g^q \dfrac{\partial}{\partial n}\left(\dfrac{u}{g^r}\right) = 0, & \text{on } \partial\mathcal{X}. \end{cases} \tag{5}$$

Special cases of this convergence, including convergences of L_u, L_n, have been studied in Calder and Trillos (2022), Trillos et al. (2020) as mentioned before in Section 1.1. Although our focus is not directly on the convergence of the weighted Laplacians but on the regression problems, the proof arguments in our paper can be applied to show the convergence of the weighted Laplacians by rigorously proving the heuristic idea in Hoffmann et al. (2022) via the concentration properties of kernel density estimation in Giné and Guillou (2002) when the domain is considered without boundary as it is well-known that the convergence of the Laplacian matrices to the Laplacian operators is problematic at the boundary (Belkin et al., 2012).

The weighted Laplacian operators are a generalization of the classical Laplacian operator with different values of $w = (p, q, r)$. Similar to the fact that the Laplacian operator is linked with the Sobolev space, the weighted Laplacian operators in (5) share a close connection with the following so-called weighted Sobolev spaces; see Triebel (1983) for a general introduction. Define the weighted L^2 space for $\ell > 0$ on $\mathcal{X}$ with a density g as

$$L^2(\mathcal{X}, g^\ell) := \left\{ u : \int_{\mathcal{X}} |u(x)|^2 g(x)^\ell dx < \infty \right\},$$

with inner product

$$\langle u, v \rangle_{g^\ell} := \int_{\mathcal{X}} u(x) v(x) g(x)^\ell dx.$$

Then, for $w := (p, q, r) \in \mathbb{R}^3$ and $s \in \mathbb{N}_+$, we define the weighted Sobolev space as:

$$H^s(\mathcal{X}, g) := \left\{ \frac{u}{g^r} \in L^2(\mathcal{X}, g^{p+r}) : \|u\|_{H^s(\mathcal{X},g)} < \infty \right\},$$

where the weighted Sobolev norm $\|u\|_{H^s(\mathcal{X},g)}$ is

$$\|u\|^2_{H^s(\mathcal{X},g)} := \sum_{j=1}^{s} |u|^2_{H^j(\mathcal{X},g)} + \left\| \frac{u}{g^r} \right\|^2_{L^2(\mathcal{X}, g^{p+r})},$$

with the j-th order semi-norm $|\cdot|_{H^j(\mathcal{X},g)}$ defined as $|u|_{H^j(\mathcal{X},g)} := \sum_{|\alpha|=j} \|D^\alpha (u g^{-r})\|_{L^2(\mathcal{X}, g^p)}$ and using multi-index notation with $x = (x^{(1)}, \ldots, x^{(d)}) \in \mathbb{R}^d$, $D^\alpha f(x) := \partial^{|\alpha|} f / \partial (x^{(1)})^{\alpha_1} \ldots \partial (x^{(d)})^{\alpha_d}$ and $|\alpha| = \alpha_1 + \ldots + \alpha_d$. When g is uniform or $r = 0$ and g is bounded from above and below, the weighted Sobolev space $H^s(\mathcal{X}, g)$ becomes (or is equivalent to) the classic Sobolev space $H^s(\mathcal{X})$. However, when f/g^r is s-times differentiable but f is not, the weighted Sobolev space differs from the classic Sobolev space. See Evans (2022) for more details regarding Sobolev spaces. For $M > 0$, the class of all functions u such that $\|u\|_{H^s(\mathcal{X},g)} \leq M$ is a weighted Sobolev ball $H^s(\mathcal{X}, g; M)$ of radius M.

Furthermore, we say a function $u \in H^s(\mathcal{X}, g)$ belongs to the zero-trace weighted Sobolev space $H^s_0(\mathcal{X}, g)$ if there exists a sequence $u_1 g^{-r}, \ldots, u_m g^{-r}$ of $C^\infty_c(\mathcal{X})$ functions such that

$$\lim_{m \to \infty} \|u_m - u\|_{H^s(\mathcal{X},g)} = 0,$$

where $C_c^\infty(\mathcal{X})$ stands for the C^∞ functions with compact support contained in $\mathcal{X}$.

Similar to the weighted Laplacian matrix $L_{w,n,\epsilon}$, the weighted Laplacian operators (5) are self-adjoint with respect to the following weighted inner product (Hoffmann et al. (2022)):

$$\langle u, v\rangle_{g^{p-r}} := \int_{\mathcal{X}} u(x)v(x)g^{p-r}(x)dx.$$

Note the following connection between the weighted norms and inner products:

$$\left\| \frac{u}{g^r} \right\|^2_{L^2(\mathcal{X}, g^{p+r})} = \|u\|^2_{L^2(\mathcal{X}, g^{p-r})} = \langle u, u\rangle_{g^{p-r}}.$$

A simple example showing the dependency of the choice M on p, q, r is as follows. Consider u/g^r is a constant function and is assumed to be 1 for simplicity and take $s = 1$. Then, we have

$$\|u\|^2_{H^1(\mathcal{X},g)} = \int_{\mathcal{X}} g(x)^{p+r}dx.$$

Clearly, the power $p+r$ of the density function g determines the size of the weighted Sobolev ball, and thus M. In other words, say for example, assuming $g \geq 1$ for simplicity, larger configurations of $p + r$ will result in large weighted Sobolev norm, thus requiring a large norm parameter M. For generic u/g^r, the situation is more intricate and depends on the geomtry of u and g and choices of $p + r$.

3 MAIN RESULTS

We now present our main results on adaptive and non-adaptive rates for estimating the regression function f as in (1) under some smoothness assumptions. Before that, we recall that the minimax estimation error over $H^s(\mathcal{X}; M)$, a standard Sobolev ball of radius M, is given by

$$\inf_{\hat{f}} \sup_{f \in H^s(\mathcal{X};M)} \|\hat{f} - f\|_n^2 \asymp M^2(M^2 n)^{-\frac{2s}{2s+d}},$$

with high probability (Györfi et al., 2002, Wasserman, 2006, Tsybakov, 2008). Moreover, there are other methods that can achieve the above minimax rate such as kernel smoothing, local polynomial regression, thin-plate splines, etc. In this context, Green et al. (2023) showed that PCR-WLE method with the unnormalized Laplacian[1] L_u achieves the minimax rate, provided that $n^{-1/2} \lesssim M \lesssim n^{s/d}$ under appropriate assumptions, where for two real-valued quantities, A, B, the notation $A \lesssim B$ means that there exists a constant $C > 0$ not depending on f, M or n such that $A \leq CB$ and $A \asymp B$ stands for $A \lesssim B$ and $B \lesssim A$.

[1]This procedure is refered to as PCR-LE in Green et al. (2023).

3.1 Assumptions

We now list the major assumptions that are needed for our theoretical results.

(A1) The distribution G is supported on $\mathcal{X}$, which is an open, connected, and bounded subset of $\mathbb{R}^d$ with Lipschitz boundary.

(A2) The distribution G has a density g on $\mathcal{X}$ such that

$$0 < g_{\min} \leq g(x) \leq g_{\max} < \infty, \text{ for all } x \in \mathcal{X},$$

for some $g_{min}, g_{\max} > 0$. Additionally, g is Lipschitz on $\mathcal{X}$ with Lipschitz constant $L_g > 0$.

(A3) The kernel η is a non-negative, monotonically non-decreasing function supported on the interval $[0, 1]$ and its restriction on $[0, 1]$ is Lipschitz and for convenience, we assume $\eta(1/2) > 0$ and define

$$\sigma_0 := \int_{\mathbb{R}^m} \eta(\|x\|)dx, \quad \sigma_1 := \frac{1}{d}\int_{\mathbb{R}^m} \|y\|^2\eta(\|y\|)dy.$$

Without loss of generality, we will assume $\sigma_0 = 1$ from now on.

(A4) The kernel η satisfies a kernel VC-type condition as follows. Let

$$\mathscr{K} := \left\{ y \to \eta\left(\frac{x-y}{\epsilon}\right) : \epsilon > 0, x \in \mathbb{R} \right\}$$

be the collection of kernel functions indexed by x and ϵ. For a density ρ, let the $L^2(\mathcal{X}, \rho)$-covering number $N(\epsilon, \mathcal{K}, \|\cdot\|_{L^2(\mathcal{X},\rho)})$ of $\mathcal{K}$ be the smallest number of $L^2(\mathcal{X}, \rho)$-balls of radius ϵ needed to cover $\mathcal{K}$. With that we say that η satisfies the kernel VC-type condition if there exist constants $A, \nu > 0$ such that

$$\sup_{\rho} N(\zeta, \mathscr{K}, \|\cdot\|_{L^2(\mathcal{X},\rho)}) \leq \left(\frac{A}{\zeta}\right)^\nu, \quad (6)$$

See Remark 3.2 for some examples.

Assumptions (A1) and (A2) are mild assumption on the density function, which are also made in Green et al. (2023). In particular (A2) is important for us, as it gives us the norm equivalence between the various families of weighted Sobolev spaces. Assumption (A3) is a standard normalization condition made on the smoothing kernel, also made in Green et al. (2023). Assumption (A4) is *not* used in Green et al. (2023). It is used here because the general family of weighted Laplacian matrices that we work with involve kernel density estimation normalization, with which the normalization in (3) will not tend to either infinity or

zero. Also note that in general condition (6) involves the $L^2(\mathcal{X},\rho)$-norm of an envelope function η_0 for $\mathcal{K}$, i.e. of a function $\eta_0 \le h$ for all $h \in \mathcal{K}$. Since, by our assumptions, η is bounded, we can use the maximum of η as an envelope, for which the $L^2(\mathcal{X},\rho)$-norm obviously does not depend on ρ and can thus be absorbed by the constant A.

3.2 Non-adaptive Rates

In the following, we present the non-adaptive minimax optimal rate of convergence of the PCR-WLE estimator in Section 2.2 for $s = 1$ and $s > 1$ separately. These rates are non-adaptive as the choice of K and ϵ depends on unknown problem parameters, the smoothness parameter s and the norm parameter M. In the following, we make a remark here that the norm $\|\cdot\|_{w,n}$ of f is empirically evaluated at $X_1,\dots,X_n$, i.e., we consider in-sample errors.

Theorem 3.1 (Non-adaptive minimax rate of PCR-WLE algorithm)**.** *Assume (A1)-(A4).*

(a) For $s \in \mathbb{N}_+\backslash\{1\}$, assume $f \in H_0^s(\mathcal{X},g;M)$, $f \in H^1(\mathcal{X},g;M)$ and $g \in C^{s-1}(\mathcal{X})$. Suppose there exist constants $c_0, C_0 > 0$ such that

$$c_0\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d}} \vee (M^2 n)^{-\frac{1}{2(s-1)+d}}\right) \le \epsilon \le C_0 K^{-\frac{1}{d}},$$

and

$$\sqrt{\frac{|\log\epsilon|}{n\epsilon^d}} \to 0, \tag{7}$$

where

$$K = \min\left\{\lfloor (M^2 n)^{\frac{d}{2s+d}}\rfloor \vee 1, n\right\}. \tag{8}$$

Then, there exist constants $c, C > 0$ not depending on f, M or n such that for n large enough and any $0 < \delta < 1$, we have:

$$\|\hat{f} - f\|_{w,n}^2 \le C\left\{\left(\delta^{-1} M^2 (M^2 n)^{-\frac{2s}{2s+d}} \wedge 1\right) \vee n^{-1}\right\},$$

with probability at least $1 - \delta - Cne^{-cn\epsilon^d} - e^{-K}$.

(b) For $s = 1$, assume $f \in H^1(\mathcal{X},g;M)$. Suppose there exist constants $c_0, C_0 > 0$ such that

$$c_0\left(\frac{\log n}{n}\right)^{\frac{1}{d}} \le \epsilon \le C_0 K^{-\frac{1}{d}},$$

and (7), where K is given in (8) for $s = 1$. Then, the assertion in part (a) also holds for $s = 1$.

Remark 3.1. *Notably, the above theorems do not require the assumption that $s > d/2$. As we mentioned before in Section 2.2, this condition is commonly appeared in the literature as in the sub-critical regime, i.e., $s \le d/2$, the (weighted) Sobolev space H^s is not a Reproducing Kernel Hilbert Space (RKHS) and cannot be continuously embedded into the space of continuous functions $C^0(\mathcal{X})$. Theorem 3.1 highlights the point that PCR-WLE algorithm obtains the minimax optimal rate when $n^{-1/2} \lesssim M \lesssim n^{s/d}$ and the error is measured by the weighted empirical norm $\|\cdot\|_{w,n}$.*

Remark 3.2. *The kernel VC-type condition was first proposed in Giné and Guillou (2002). A simple sufficient condition for this condition to hold is that η is of bounded variation; see Nolan and Pollard (1987) or Giné and Nickl (2021). Clearly, many common kernels are of this type, including Gaussian, Epanechnikov and cosine kernels. Furthermore, Matérn family of kernels is also of practical importance but is not considered a natural choice for kernels satisfying the required bounded variation condition due to the oscillatory behavior of the Bessel function while the bounded variation condition is generally associated with functions that do not oscillate too wildly. However, some special cases of the family indeed satisfy the bounded variation condition. For smoothness parameters $\nu = p + 1/2$ of Matérn kernels with positive integer p, the Matérn kernels can be expressed as a product of an exponential and a polynomial whose derivative is absolutely integrable, and thus they satisfy the condition. Moreover, the Matérn kernel becomes the Gaussian (RBF) kernel in the limit of $\nu \to \infty$, thereby satisfying the VC-type condition.*

Remark 3.3. *For practical consideration, there are two tuning parameters: the graph radius (the bandwidth for the kernel η) ϵ and the number of eigenvalues K. The lower bound for ϵ makes sure that with this smallest radius, the resulting weighted graph will still be connected with high probability and the upper bound for ϵ ensures the eigenvalue of the weighted graph Laplacian (3) to be of the same order as its continuum version, the eigenvalue of the weighted Laplacian operator (5) (Weyl's law). The asymptotic assumption on ϵ is from the concentration of the KDE. The condition on K is set to trade-off bias and variance. Both ϵ and K depend on the true smoothness parameter $s \in \mathbb{N}_+$.*

3.3 Adaptive Rates Via Lepski's Method

Despite the minimax optimality of the PCR-WLE algorithm shown in Section 3.2, the main practical difficulties are the choice of several tuning parameters including the bandwidth parameter (or the graph radius) ϵ and the number of eigenvalues K, because optimal choices depend on the *unknown* true smoothness parameter s of the regression function f in the model 1. Moreover, K also relies on the bound of the weighted Sobolev norm M. This naturally brings

about the issue of adaptation, which we address using Lepski's method. Note that, as we are concerned with in-sample estimation error, other techniques like cross-validation are not directly applicable to set the tuning parameters.

Since its introduction in Lepskii (1991), Lepski's method has been widely used for adaptive estimation and testing in various statistical contexts; e.g. see Birgé (2001), Chichignoud et al. (2016), Bellec et al. (2018), Balasubramanian et al. (2021), Lacour and Massart (2016). In the following, we consider Lepski's method on the product space of the smoothness parameter $s \in \mathbb{N}_+$ and the constraint on the weighted Sobolev norm $M \in \mathbb{R}_+$.

Recall that s and M denote the true smoothness parameter and the norm parameter, respectively for the weighted Sobolev norm of f. Here, we actually take M as the minimum over all bounds of the weighted Sobolev norm. We start by picking $s_{\min}, s_{\max} \in \mathbb{N}_+$; here we can set $s_{\min} = 1$ under no availability of further information[2] regarding the knowledge of s. The goal is that $s_{\max}$ is large enough that $s \in \mathbb{N}_+$ satisfies $s \in [s_{\min}, s_{\max}]$. Similarly, we pick $M_{\min}, M_{\max}$ satisfying $0 < M_{\min} < M_{\max} < \infty$, where $M_{\min}$ and $M_{\max}$ are small and large enough respectively such that $M \in [M_{\min}, M_{\max}]$. Next, define the grid $\mathcal{B} \times \mathcal{D} := \{(s_j, M_j)\}_{j=1}^{N_l}$ given by:

$$\begin{aligned} \mathcal{B} &:= [s_{\min}, s_{\max}] \cap \mathbb{N}_+ \\ &= \{s_{\min} =: s_1 < s_2 < \ldots < s_{N_l} := s_{\max}\}, \end{aligned} \tag{9}$$

and

$$\begin{aligned} \mathcal{D} &:= [M_{\min}, M_{\max}] \\ &= \{M_{\min} =: M_1 < M_2 < \ldots < M_{N_l} := M_{\max}\}, \end{aligned}$$

where $N_l \asymp \log n$.

For any pair $(\tilde{s}, \tilde{M})$ in the above grid, let $\hat{f}_{\tilde{s},\tilde{M}}$ be the PCR-WLE estimator in Section 2.2 corresponding to the parameters $\tilde{s}$ and $\tilde{M}$. We define the Lepski's estimator as

$$\hat{f}_{\mathsf{adapt}} := \hat{f}_{\hat{s},\hat{M}},$$

where $\hat{s}$ is given by

$$\begin{aligned} \hat{s} := \max\{\tilde{s} \in \mathcal{B} : &\|\hat{f}_{\tilde{s},\tilde{M}} - \hat{f}_{\tilde{s}',\tilde{M}'}\|_{w,n} \\ &\leq c_0 \tilde{M}'((\tilde{M}'^2 n/\log n)^{-\frac{\tilde{s}'}{2\tilde{s}'+d}}, \forall \tilde{s}' \leq \tilde{s}, \tilde{s}' \in \mathcal{B}\}, \end{aligned}$$

and $\hat{M}$ is the corresponding couple of $\hat{s}$ in the grid, where $c_0 > 0$ is some finite constant. Here, we formulate the above simultaneous Lepski's method by coupling the smoothness parameter and the norm parameter and only maximize through the smoothness parameter instead of dealing with a joint maximization, which is not needed for our purpose of showing the adaptive minimax rate in the following result as our focus is its convergence rate in n.

The following result presents a near minimax optimal rate of convergence of the Lepski's estimator $\hat{f}_{\mathsf{adapt}}$ up to a logarithmic factor in n.

Theorem 3.2. *Assume (A1)-(A4) and $g \in C^{s-1}(\mathcal{X})$. Also, assume $f \in H^1(\mathcal{X}, g; M) \cap H_0^s(\mathcal{X}, g; M)$ and $f_g := f/g^r$ is M-Lipschitz, i.e., $\|f_g(x) - f_g(x')\| \leq M\|x - x'\|$ for any $x, x' \in \mathcal{X}$. Furthermore, assume that (for large enough n) we have $s \in [s_{\min}, s_{\max}]$ and $M \in [M_{\min}, M_{\max}]$. Then, under the minimax optimal setting in Theorem 3.1 for M, i.e., $n^{-1/2} \lesssim M \lesssim n^{s/d}$, the estimator $\hat{f}_{\mathsf{adapt}}$ satisfies: For n large enough and any $\delta \in (0,1)$, there exists some constant $C > 0$ such that*

$$\|\hat{f}_{\mathsf{adapt}} - f\|_{w,n}^2 \leq C\delta^{-1}M^2(M^2 n/\log n)^{-\frac{2s}{2s+d}},$$

with probability at least

$$\begin{aligned} 1 - \delta \log^{-2s/(2s+d)} n - Cne^{-Cn\epsilon^d} \log^2 n \\ - 16Cc_0^{-4} n^{-1} \log^{2-2s_{\min}/(2s_{\min}+d)} n \\ - e^{-\lfloor M_{\min}^2 n \rfloor^{d/(2s+d)}} \log^2 n. \end{aligned}$$

Remark 3.4. *Trillos et al. (2022) proposed a graph poly-Laplacian regularization approach, where integer powers of the Laplacian matrices are used as regularization in a least-squares context. They showed that the proposed method achieves rate of convergence of order $n^{-s/(d+4s)}$. While the rate is not optimal, in comparison to the Green et al. (2023) their estimator does not require the knowledge of the norm parameter M to achieve the derived rate (althogh they require the knowledge of s). In comparison to both the above works, our result in Theorem 3.2 achieves the optimal rate, up to* log *factors, without requiring the knowledge of either s or M.*

Remark 3.5. *As a part of our proof, a better concentration inequality for the non-adaptive PCR-WLE estimator $\hat{f}$ is required compared to Theorem 3.1, for which the assumption that f_g is Lipschitz is required. As also discussed in Green et al. (2023, see below Theorem 1), it remains open whether a weaker assumption or even the weighted Sobolev condition $\|\nabla f_g\|_{L^2} < \infty$ alone might be sufficient establish the required concentration result for developing adaptive procedures.*

[2] If there is additional information, like $s > 10$, one can pick $s_{\min} = 10$. Hence, we present out result with a generic $s_{\min}$.

4 CONCLUSION

In this work, we provide adaptive and non-adaptive rates of convergence, in Theorem 3.1 and 3.2 respectively, for estimating a true regression function lying belonging to the Sobolev space. Our estimators are based on performing principal components regression based on the eigenvectors of the weighted graph Laplacian matrix, and using Lepski's method for deriving the adaptive results. Our contributions expand upon the non-adaptive outcome outlined in Green et al. (2023), which was originally established for a particular normalized graph Laplacian. This extension encompasses a broad spectrum of weighted Laplacian matrices commonly employed in practical applications, including the unnormalized Laplacian and the random walk Laplacian among them.

Future works include (i) relaxing the assumption that the density g is bounded from below, (ii) developing confidence intervals for the estimators by establishing asymptotic normality results and developing related bootstrap procedures, and (iii) developing estimators that are instance-optimal in the sense of Hoffman and Lepski (2002), i.e., estimators that achieve the best possible rate for a given combination of the true regression function f and the sampling density g by adaptively picking the parameters p, q and r in the weighted graph Laplacian matrix.

References

Mathieu Andreux, Emanuele Rodolà, Mathieu Aubry, and Daniel Cremers. Anisotropic Laplace-Beltrami Operators for Shape Analysis. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, Lecture Notes in Computer Science, pages 299–312, 2015. (Cited on page 2.)

Krishnakumar Balasubramanian, Tong Li, and Ming Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *The Journal of Machine Learning Research*, 22(1):1–45, 2021. (Cited on page 7.)

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. (Cited on pages 2 and 4.)

Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *International Conference on Computational Learning Theory*, pages 486–500. Springer, 2005. (Cited on page 2.)

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006. (Cited on page 2.)

Mikhail Belkin, Qichao Que, Yusu Wang, and Xueyuan Zhou. Toward understanding complex spaces: Graph Laplacians on manifolds with singularities and boundaries. In *Conference on learning theory*, pages 36–1. JMLR Workshop and Conference Proceedings, 2012. (Cited on page 4.)

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets Lasso: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018. (Cited on page 7.)

Lucien Birgé. An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, pages 113–133, 2001. (Cited on page 7.)

Olivier Bousquet, Olivier Chapelle, and Matthias Hein. Measure based regularization. *Advances in Neural Information Processing Systems*, 16, 2003. (Cited on page 2.)

Jeff Calder and Nicolas Garcia Trillos. Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs. *Applied and Computational Harmonic Analysis*, 60:123–175, 2022. (Cited on pages 2, 3, 4, 31, and 32.)

Frédéric Chazal, Ilaria Giulini, and Bertrand Michel. Data driven estimation of Laplace-Beltrami operator. *Advances in Neural Information Processing Systems*, 29, 2016. (Cited on page 2.)

Michael Chichignoud, Johannes Lederer, and Martin J Wainwright. A practical scheme and fast algorithm to tune the Lasso with optimality guarantees. *The Journal of Machine Learning Research*, 17(1):8162–8181, 2016. (Cited on page 7.)

Yongwan Chun, Daniel A Griffith, Monghyeon Lee, and Parmanand Sinha. Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. *Journal of Geographical Systems*, 18:67–85, 2016. (Cited on page 2.)

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006. (Cited on page 2.)

Lee H Dicker, Dean P Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. 2017. (Cited on page 2.)

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. (Cited on page 34.)

David B Dunson, Hau-Tieng Wu, and Nan Wu. Spectral convergence of graph Laplacian and heat kernel

reconstruction in L_∞ from random samples. *Applied and Computational Harmonic Analysis*, 55:282–336, 2021. (Cited on page 2.)

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022. (Cited on pages 4 and 12.)

Nicolas Garcia Trillos and Ryan W Murray. A maximum principle argument for the uniform convergence of graph Laplacian regressors. *SIAM Journal on Mathematics of Data Science*, 2(3):705–739, 2020. (Cited on page 2.)

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002. (Cited on pages 2, 4, 6, and 30.)

Evarist Giné and Vladimir Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. *Lecture Notes-Monograph Series*, pages 238–259, 2006. (Cited on page 2.)

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021. (Cited on page 6.)

Alden Green, Sivaraman Balakrishnan, and Ryan Tibshirani. Minimax optimal regression over Sobolev spaces via Laplacian regularization on neighborhood graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 2602–2610. PMLR, 2021. (Cited on pages 2, 12, 28, 31, 32, 33, and 34.)

Alden Green, Sivaraman Balakrishnan, and Ryan J Tibshirani. Minimax optimal regression over Sobolev spaces via Laplacian Eigenmaps on neighbourhood graphs. *Information and Inference: A Journal of the IMA*, 12(3):2423–2502, 2023. (Cited on pages 1, 2, 4, 5, 7, 8, 12, 16, 19, 20, and 23.)

László Györfi, Michael Köhler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002. (Cited on page 5.)

Olympio Hacquard, Krishnakumar Balasubramanian, Gilles Blanchard, Clément Levrard, and Wolfgang Polonik. Topologically penalized regression on manifolds. *The Journal of Machine Learning Research*, 23(1):7233–7271, 2022. (Cited on page 2.)

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005. (Cited on page 2.)

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(6), 2007. (Cited on page 2.)

M Hoffman and Oleg Lepski. Random rates in anisotropic regression (with a discussion and a rejoinder by the authors). *The Annals of Statistics*, 30(2):325–396, 2002. (Cited on page 8.)

Franca Hoffmann, Bamdad Hosseini, Assad A Oberai, and Andrew M Stuart. Spectral analysis of weighted Laplacians arising in data clustering. *Applied and Computational Harmonic Analysis*, 56:189–249, 2022. (Cited on pages 1, 2, 3, 4, 5, and 13.)

Claire Lacour and Pascal Massart. Minimal penalty for Goldenshluger–Lepski method. *Stochastic Processes and their Applications*, 126(12):3774–3789, 2016. (Cited on pages 2 and 7.)

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000. (Cited on page 22.)

Ann B Lee and Rafael Izbicki. A spectral series approach to high-dimensional nonparametric regression. 2016. (Cited on page 2.)

Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017. (Cited on page 19.)

OV Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991. (Cited on page 7.)

Sridhar Mahadevan and Mauro Maggioni. Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes. *Journal of Machine Learning Research*, 8(10), 2007. (Cited on page 2.)

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 849–856, January 2001. (Cited on page 2.)

Deborah Nolan and David Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15, 1987. (Cited on page 6.)

John Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, pages 1215–1230, 1984. (Cited on page 2.)

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(2), 2010. (Cited on page 2.)

Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 22(8):888–905, August 2000. ISSN 1939-3539. (Cited on page 2.)

Zuoqiang Shi. Convergence of Laplacian spectra from random samples. *arXiv preprint arXiv:1507.00151*, 2015. (Cited on page 2.)

Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009. (Cited on page 2.)

Hans Triebel. *Theory of Function Spaces*. Springer–Modern Birkhäuser Classics, 1983. (Cited on page 4.)

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

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(4): 827–887, 2020. (Cited on page 4.)

Nicolás García Trillos, Ryan Murray, and Matthew Thorpe. Rates of Convergence for Regression with the Graph Poly-Laplacian. *arXiv preprint arXiv:2209.02305*, 2022. (Cited on pages 2 and 7.)

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008. (Cited on page 5.)

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007. (Cited on page 2.)

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008. (Cited on page 2.)

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan Tibshirani. Trend filtering on graphs. In *Artificial Intelligence and Statistics*, pages 1042–1050. PMLR, 2015. (Cited on page 2.)

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006. (Cited on page 5.)

Yair Weiss. Segmentation using eigenvectors: A unifying view. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 975–982. IEEE, 1999. (Cited on page 2.)

Yifan Wu, George Tucker, and Ofir Nachum. The Laplacian in RL: Learning Representations with Efficient Approximations. In *International Conference on Learning Representations*, 2019. (Cited on page 2.)

Xueyuan Zhou and Nathan Srebro. Error analysis of laplacian eigenmaps for semi-supervised learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 901–908. JMLR Workshop and Conference Proceedings, 2011. (Cited on page 2.)

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003. (Cited on page 2.)

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Adaptive and non-adaptive minimax rates for weighted Laplacian-Eigenmap based nonparametric regression: Supplementary Materials

5 PROOF

5.1 Proof Of Theorem 3.1

In this section, we will prove both Theorem 3.1 for $s = 1$ and $s > 1$ together. We first present and prove some auxiliary lemmas. We will denote by $B_x(r)$ a closed Euclidean ball with midpoint x and radius $r \geq 0$.

Define the weighted Sobolev seminorm $\langle L_{w,\epsilon} f, f\rangle_{g^{p-r}}$ given by the following non-local operator:

$$L_{w,\epsilon} f(x) := \frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}} g(x)^{1-p} \frac{\eta\big(\frac{\|x-z\|}{\epsilon}\big)}{g(x)^{1-q/2} g(z)^{1-q/2}} (g(x)^{-r} f(x) - g(z)^{-r} f(z)) g(z) dz,$$

where according to (4), $L_{w,\epsilon}$ can be viewed as a population counterpart of the discrete graph weighted Laplacian $L_{w,n,\epsilon}$. As in Green et al. (2023), we also call it a 'non-local' version. Note that the above non-local weighted Sobolev seminorm and non-local operator generalize the definitions in Green et al. (2023) as the latter belong to a special case when $(p, q, r) = (1, 2, 0)$. The following Lemmas 5.1-5.6 therefore extend their counterparts in Green et al. (2023) to the weighted Laplacians and the weighted Sobolev seminorm. Note that in our proofs, we also highlight and fix several important typos and errors that appeared in Green et al. (2023). Despite the errors, the final results in Green et al. (2023) remain true.

Lemma 5.1. *For $f \in H^1(\mathcal{X}, g; M)$, we have*

$$\langle L_{w,\epsilon} f, f\rangle_{g^{p-r}} \lesssim M^2.$$

Proof of Lemma 5.1. Following the idea of Green et al. (2021, Proof of Lemma 1), take Ω as an arbitrary bounded open set such that $B_x(c_0) \subseteq \Omega$ for all $x \in \mathcal{X}$ for some $c_0 > 0$ and we can assume that $f \in H^1(\Omega, g)$ and $\|f\|_{H^1(\Omega,g)} \lesssim \|f\|_{H^1(\mathcal{X},g)}$ without loss of generality due to the existence of an extension operator $E : H^1(\mathcal{X}, g) \to H^1(\Omega, g)$ such that Ef satisfies these properties, see Theorem 1 in Chapter 5.4 in Evans (2022). Also, since $C^\infty(\Omega)$ is dense in $H^1(\Omega, g)$ and the integral in Lemma 5.1 is continuous in $H^1(\Omega, g)$, we can assume $f_g := f/g^r \in C^\infty(\Omega)$ so that

$$f_g(x') - f_g(x) = \int_0^1 \nabla f_g(x + t(x' - x))^T (x' - x) dt.$$

Then, we have by symmetry in the first step:

$$\begin{aligned}
&2\langle L_{w,\epsilon} f, f\rangle_{g^{p-r}} \\
&= \frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{\eta\left(\frac{\|x-y\|}{\epsilon}\right)}{g(x)^{1-q/2} g(y)^{1-q/2}} \left| \frac{f(x)}{g(x)^r} - \frac{f(y)}{g(y)^r} \right|^2 g(x) g(y) dx dy \\
&= \frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{\eta\left(\frac{\|x-y\|}{\epsilon}\right)}{g(x)^{1-q/2} g(y)^{1-q/2}} \left(\int_0^1 \nabla f_g(y + t(x - y))^T (x - y) dt \right)^2 g(x) g(y) dx dy \\
&\leq \frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}} \int_{\mathcal{X}} \int_0^1 \frac{\eta\left(\frac{\|x-y\|}{\epsilon}\right)}{g(x)^{1-q/2} g(y)^{1-q/2}} \left(\nabla f_g(y + t(x - y))^T (x - y) \right)^2 g(x) g(y) dt dx dy \\
&\leq \int_{\mathcal{X}} \int_{B_{\mathbf{0}}(1)} \int_0^1 \left(\nabla f_g(y + \epsilon t z)^T z \right)^2 \frac{\eta(\|z\|)}{g(y + \epsilon z)^{1-q/2} g(y)^{1-q/2}} g(y + \epsilon z) g(y) dt dz dy, \\
&\quad \text{with } (x - y)/\epsilon = z \\
&\lesssim \int_{\mathcal{X}} \int_{B_{\mathbf{0}}(1)} \int_0^1 \left(\nabla f_g(y + \epsilon t z)^T z \right)^2 \eta(\|z\|) g(y + \epsilon t z)^q dt dz dy
\end{aligned}$$

$$\leq \int_{\Omega}\int_{B_{\mathbf{0}}(1)}\int_0^1 \left(\nabla f_g(\tilde{y})^T z\right)^2 \eta(\|z\|) g(\tilde{y})^q dt dz d\tilde{y}, \quad \tilde{y} = y + \epsilon t z \in \Omega. \tag{10}$$

Since we have $\left(\nabla f_g(\tilde{y})^T z\right)^2 = \left(\sum_{i=1}^d (\nabla f_g(\tilde{y}))^{(i)} z^{(i)}\right)^2$ and $\eta(\|z\|)$ is invariant with respect to the rotation, it yields that

$$\begin{aligned}
\int_{B_{\mathbf{0}}(1)} \left(\nabla f_g(\tilde{y})^T z\right)^2 \eta(\|z\|) dz &= \sum_{i,j=1}^d (\nabla f_g(\tilde{y}))^{(i)} (\nabla f_g(\tilde{y}))^{(j)} \int_{B_{\mathbf{0}}(1)} z^{(i)} z^{(j)} \eta(\|z\|) dz \\
&= \sum_{i=1}^d \left((\nabla f_g(\tilde{y}))^{(i)}\right)^2 \int_{B_{\mathbf{0}}(1)} \left(z^{(i)}\right)^2 \eta(\|z\|) dz \\
&= \sigma_1 \left\| \nabla \left(\frac{f(\tilde{y})}{g(\tilde{y})^r} \right) \right\|^2 .
\end{aligned} \tag{11}$$

Plugging (11) in (10), we conclude

$$2\langle L_{w,\epsilon} f, f \rangle_{g^{p-r}} \lesssim \sigma_1 M^2.$$

This finishes the proof. □

Note that the proof of Lemma 5.1 also utilized the heuristic arguments given in Hoffmann et al. (2022) while we provide a rigorous proof here.

Lemma 5.2. *Suppose $f_g \in L^2(\mathcal{U}, g^{p+r}; M)$ for a Borel set $\mathcal{U} \subseteq \mathcal{X}$. Then, there exists a constant C which does not depend on f or M such that*

$$\|L_{w,\epsilon} f\|_{L^2(\mathcal{U}, g^{p+r})} \leq \frac{C}{\epsilon^2} \|f_g\|_{L^2(\mathcal{U}, g^{p+r})}.$$

Proof. By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|L_{w,\epsilon} f(x)|^2 &= \frac{1}{\epsilon^{2(d+2)}} \left(\int_{\mathcal{U}} g(x)^{1-p} \frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2} g(z)^{1-q/2}} (g(x)^{-r} f(x) - g(z)^{-r} f(z)) g(z) dz \right)^2 \\
&\lesssim \frac{1}{\epsilon^{2(d+2)}} g(x)^{2(1-p)} \int_{\mathcal{U}} \frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2} g(z)^{1-q/2}} (f_g(x) - f_g(z))^2 dz \cdot \int_{\mathcal{X}} \frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2} g(z)^{1-q/2}} dz \\
&\lesssim \frac{2\sigma_0}{\epsilon^{4+d}} g(x)^{2(q-p)} \int_{\mathcal{U}} \eta\left(\frac{\|x-z\|}{\epsilon}\right) (|f_g(x)|^2 + |f_g(z)|^2) dz.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\|L_{w,\epsilon} f\|^2_{L^2(\mathcal{U}, g^{p+r})} &= \int_{\mathcal{U}} g^{p-r}(x) |L_{w,\epsilon} f(x)|^2 dx \\
&\lesssim \frac{2}{\epsilon^{4+d}} \int_{\mathcal{U}}\int_{\mathcal{U}} g(x)^{2(q-p)+p-r}(x) \eta\left(\frac{\|x-z\|}{\epsilon}\right) (|f_g(x)|^2 + |f_g(z)|^2) dz dx \\
&\lesssim \frac{2}{\epsilon^{4+d}} \int_{\mathcal{U}}\int_{\mathcal{U}} \eta\left(\frac{\|x-z\|}{\epsilon}\right) (|f_g(x)|^2 + |f_g(z)|^2) dz dx \\
&\lesssim \frac{4}{\epsilon^{4+d}} \int_{\mathcal{U}}\int_{\mathcal{U}} \eta\left(\frac{\|x-z\|}{\epsilon}\right) |f_g(x)|^2 dz dx \\
&\leq \frac{4}{\epsilon^4} \int_{\mathcal{U}} |g(x)^{p+r}(x) f_g(x)|^2 dx \\
&\lesssim \frac{4}{\epsilon^4} \|f_g\|^2_{L^2(U, g^{p+r})}
\end{aligned}$$

□

Lemma 5.3. *Suppose $f_g \in L^2(\mathcal{U}, g^{p+r}; M)$ for a Borel set $\mathcal{U} \subseteq \mathcal{X}$. Then, there exists a constant $C > 0$ such that*

$$E_{w,\epsilon}(f;\mathcal{U}) \leq \frac{C}{\epsilon^2}\|f_g\|^2_{L^2(\mathcal{U},g^{p+r})},$$

where we define the Dirichlet energy for the set $\mathcal{U}$ as

$$E_{w,\epsilon}(f,\mathcal{U}) := \frac{1}{\epsilon^{d+2}}\int_{\mathcal{U}}\int_{\mathcal{U}}(g(x)^{-r}f(x) - g(z)^{-r}f(z))^2\frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2}g(z)^{1-q/2}}g(x)g(z)dxdz.$$

Proof. Note that

$$\begin{aligned}
E_{w,\epsilon}(f;\mathcal{U}) &= \frac{1}{\epsilon^{d+2}}\int_{\mathcal{U}}\int_{\mathcal{U}}(g(x)^{-r}f(x) - g(z)^{-r}f(z))^2\frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2}g(z)^{1-q/2}}g(x)g(z)dxdz \\
&\leq \frac{2}{\epsilon^{d+2}}\int_{\mathcal{U}}\int_{\mathcal{U}}(|g(x)^{-r}f(x)|^2 + |g(z)^{-r}f(z)|^2)\frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2}g(z)^{1-q/2}}g(x)g(z)dxdz \\
&= \frac{4}{\epsilon^{d+2}}\int_{\mathcal{U}}\int_{\mathcal{U}}|g(x)^{-r}f(x)|^2\eta\left(\frac{\|x-z\|}{\epsilon}\right)g(x)^{q/2}g(z)^{q/2}dxdz \\
&\lesssim \frac{4}{\epsilon^{d+2}}\int_{\mathcal{U}}\int_{\mathcal{U}}|g(x)^{-r}f(x)|^2\eta\left(\frac{\|x-z\|}{\epsilon}\right)g(x)^{p+r}dxdz \\
&\lesssim \frac{4}{\epsilon^2}\int_{\mathcal{U}}|g(x)^{-r}f(x)|^2 g(x)^{p+r}dx.
\end{aligned}$$

□

We denote by $\mathcal{X}_{t\epsilon}$ a subset of $\mathcal{X}$ such that for any $x \in \mathcal{X}_{t\epsilon}$, $B_x(t\epsilon) \in \mathcal{X}$ consisting of points sufficiently far away from the boundary and $\partial_{t\epsilon}\mathcal{X}$ by its complement within $\mathcal{X}$ consisting of points close enough to the boundary.

Lemma 5.4. *For $f \in H^1(\mathcal{X}, g; M) \cap H^s_0(\mathcal{X}, g; M)$ with $s \in \mathbb{N}_+$ and $g \in C^{s-1}(\mathcal{X})$, there exist constants $C_1, C_2 > 0$ such that*

(1) If s is odd, then we have with $t = (s-1)/2$:

$$\|L^t_{w,\epsilon}f - \sigma_1^t\mathcal{L}^t_w f\|_{L^2(\mathcal{X}_{t\epsilon},g^{p+r})} \leq C_1 M\epsilon.$$

(2) If s is even, then we have with $t = (s-2)/2$:

$$\|L^t_{w,\epsilon}f - \sigma_1^t\mathcal{L}^t_w f\|_{L^2(\mathcal{X}_{t\epsilon},g^{p+r})} \leq C_2 M\epsilon^2.$$

Proof. Without loss of generality, we assume both g and f are $C^\infty(\mathcal{X})$ due to the fact that $C^\infty(\mathcal{X})$ is dense in both $H^s(\mathcal{X}, g)$ and $C^{s-1}(\mathcal{X})$ and the norm in the statements is continuous with respect to $\|\cdot\|_{H^s(\mathcal{X},g)}$ and $\|\cdot\|_{C^{s-1}(\mathcal{X})}$.

Actually, we claim the following stronger result: for $t < s/2$ and every $x \in \mathcal{X}_{t\epsilon}$,

$$L^t_{w,\epsilon}f(x) = \sigma_1^t\mathcal{L}_w f(x) + \sum_{j=1}^{\lfloor(s-1)/2\rfloor - t} r_{2(j+t)}(x)\epsilon^{2j} + r_s(x)\epsilon^{s-2t}, \tag{12}$$

for some functions r_j such that

$$\|r_j\|_{H^{s-j}(\mathcal{X}_{t\epsilon},g)} \leq C\|g\|^t_{C^{s-1}(\mathcal{X})}M. \tag{13}$$

Note that the dependence of the functions r_j on t is suppressed in the notation.

The key idea underlying the proof of (12) is to consider the following Taylor expansion. For an s-times differentiable function $F : \mathcal{X} \to \mathbb{R}$ and $x \in \mathcal{X}$, define the following operator d_x^s:

$$(d_x^s F)(z) := \sum_{|\alpha|=s} D^\alpha F(x) z^\alpha.$$

Also, define $d^s F := \sum_{|\alpha|=s} D^\alpha F$. Then, for $\phi \in C^s(\mathcal{X})$ and some $h > 0$, $z \in \mathcal{X}_h$, $x \in B_z(h)$, the Taylor expansion at z is given as:

$$\phi(x) = \phi(z) + \sum_{j=1}^{s-1} \frac{1}{j!} (d_x^j \phi)(x - z) + R_s(x, z; \phi).$$

Here, we note that $(d_x^j \phi)(z)$ is a polynomial of degree j and we have for any $y \in \mathbb{R}$:

$$(d_x^j \phi)(yz) = y^j (d_x^j \phi)(z).$$

The remainder term $R_j(x, z; \phi)$ is

$$R_j(x, z; \phi) := \frac{1}{(j-1)!} \int_0^1 (1-\theta)^{j-1} (d^j_{z+\theta(x-z)} \phi)(x - z) d\theta,$$

such that for any $x^* \in B_{\mathbf{0}}(1)$,

$$\sup_{x \in \mathcal{X}_h} |R_j(x, x + hx^*; \phi)| \le C h^j \|\phi\|_{C^j(\mathcal{X})},$$

and

$$\int_{\mathcal{X}_h} |R_j(z + \theta h x, z; \phi)|^2 dz \le h^{2j} \int_{\mathcal{X}_h} \int_0^1 |(d^j_{z+\theta h x} \phi)(z)|^2 d\theta dz \le h^{2j} \|d^j \phi\|^2_{L^2(\mathcal{X})}.$$

Now, we apply the above Taylor expansion on the function $f_g(x) := f(x)/g(x)^r$ up to order s and the function $g^{q/2}(x)$ up to order S in $L_{w,\epsilon} f(x)$, where $S = 1$ if $s = 1$ and otherwise $S = s - 1$ and obtain:

$$\begin{aligned}
L_{w,\epsilon} f(x) = &\frac{1}{\epsilon^{d+2}} \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{S-1} \frac{1}{j_1! j_2!} \int_{\mathcal{X}} g(x)^{q/2-p} \eta\left(\frac{\|x - z\|}{\epsilon}\right) (d_x^{j_1} f_g)(x - z)(d_x^{j_2} g^{q/2})(z - x) dz \\
&+ \frac{1}{\epsilon^{d+2}} \sum_{j=1}^{s-1} \frac{1}{j!} \int_{\mathcal{X}} g(x)^{q/2-p} \eta\left(\frac{\|x - z\|}{\epsilon}\right) (d_x^{j_1} f_g)(x - z) R_S(x, z; g^{q/2}) dz \\
&+ \frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}} g(x)^{p/2-p} \eta\left(\frac{\|x - z\|}{\epsilon}\right) R_s(x, z; f_g) g(z)^{q/2} dz.
\end{aligned}$$

Now, with the transformation $y = (z - x)/\epsilon$, we have

$$\begin{aligned}
L_{w,\epsilon} f(x) = &-\frac{1}{\epsilon^2} \sum_{j_1=1}^{s-1} \sum_{j_2=0}^{S-1} \frac{\epsilon^{j_1+j_2}}{j_1! j_2!} \int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p} \eta(\|y\|) (d_x^{j_1} f_g)(y)(d_x^{j_2} g^{q/2})(y) dy \\
&- \frac{1}{\epsilon^2} \sum_{j=1}^{s-1} \frac{\epsilon^j}{j!} \int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p} \eta(\|y\|) (d_x^{j_1} f_g)(y) R_S(x, \epsilon y + x; g^{q/2}) dy \\
&+ \frac{1}{\epsilon^2} \int_{B_{\mathbf{0}}(1)} g(x)^{p/2-p} \eta(\|y\|) R_s(x, \epsilon y + x; f_g) g(\epsilon y + x)^{q/2} dy \\
=: &\, L_1(x) + L_2(x) + L_3(x).
\end{aligned}$$

We will now prove (12) by induction on t, and throughout this proof, with a slight abuse of notation, the functions r_j in (12) may vary from line to line depending on t at the induction step but they will always satisfy the condition (13) as we are only interested in the bounds.

Firstly, we start with $L_1(x)$. If $s = 1$, we can see $L_1(x) = 0$. Therefore, in the following, we only focus on $s \geq 2$. Now, we define

$$l_{j_1,j_2}(x) := \int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p}\eta(\|y\|)(d_x^{j_1}f_g)(y)(d_x^{j_2}g^{q/2})(y)dy,$$

such that

$$L_1(x) = -\frac{1}{\epsilon^2}\sum_{j_1=1}^{s-1}\sum_{j_2=0}^{S-1}\frac{\epsilon^{j_1+j_2}}{j_1!j_2!}l_{j_1,j_2}(x).$$

Since $(d_x^j f_g)(y)$ is a polynomial of degree j, l_{j_1,j_2} actually depends on the sum $j_1 + j_2$ and $d_x^{j_1}d_x^{j_2}$ is an order $j_1 + j_2$ multivariate monomial. Therefore, when $j_1 + j_2$ is odd, we have

$$l_{j_1,j_2}(x) = 0.$$

Then, when $s = 2$, we have $j_1 + j_2 = 1$ and $L_1(x) = 0$. As for $s \geq 3$, we notice that the lowest order term of $L_{11}(x)$ is from $j_1 + j_2 = 2$, which means either $j_1 = 1, j_2 = 1$ or $j_1 = 2, j_2 = 0$. We have

$$\begin{aligned} l_{1,1}(x) &= \int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p}\eta(\|y\|)(d_x^1 f_g)(y)(d_x^1 g^{q/2})(y)dy \\ &= \sum_{i_1=1,i_2=1}^{d} g(x)^{q/2-p}(Df_g(x))^{(i_1)}(Dg^{q/2}(x))^{(i_2)}\int_{B_{\mathbf{0}}(1)}\|y\|^2\eta(\|y\|)dy, \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2}l_{2,0}(x) &= \frac{1}{2}\int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p}\eta(\|y\|)(d_x^2 f_g)(y)(d_x^2 g^{q/2})(y)dy \\ &= \frac{1}{2}\sum_{i=1}^{d} g(x)^{q/2-p}((Df_g(x))^{(i)})^2 g(x)^{q/2}\int_{B_{\mathbf{0}}(1)}\|y\|^2\eta(\|y\|)dy. \end{aligned}$$

Therefore, we have by definition:

$$\mathcal{L}_w f(x) = -\frac{1}{2g(x)^p}\left(\nabla g(x)^q \cdot \nabla\left(\frac{f(x)}{g(x)^r}\right) + g(x)^q\Delta\left(\frac{f(x)}{g(x)^r}\right)\right),$$

and

$$-(l_{1,1}(x) + \frac{1}{2}l_{2,0}(x)) = \sigma_1\mathcal{L}_w f(x).$$

This is exactly the leading term. We remark here that in Green et al. (2023, Section D.2), the negative sign is missing, which does not actually give the Laplacian operator by the leading term. Now, it remains to bound the higher order terms with $j_1 + j_2 > 2$. We will show that

$$L_1(x) = \sigma_1\mathcal{L}_w + \sum_{j=1}^{\lfloor(s-1)/2\rfloor-1} r_{2(j+1)}(x)\epsilon^{2j} + r_s(x)\epsilon^{s-2}.$$

It suffices to show for $j_1 + j_2 > 2$, l_{j_1,j_2} satisfies (13) for $j = \min\{j_1 + j_2 - 2, s - 2\}$. Through the multi-index notation, we write that

$$l_{j_1,j_2}(x) = g(x)^{q/2-p}\sum_{|\alpha_1|=j_1,|\alpha_2|=j_2} D^{\alpha_1}f_g(x)D^{\alpha_2}g^{q/2}(x)\int_{B_{\mathbf{0}}(1)} y^{\alpha_1}y^{\alpha_2}\eta(\|y\|)dy,$$

where $|\int_{B_{\mathbf{0}}(1)} y^{\alpha_1}y^{\alpha_2}\eta(\|y\|)dy| < \infty$ for all α_1, α_2. Then, by Hölder's inequality, we have for $|\alpha_1| = j_1$, $|\alpha_2| = j_2$,

$$\|g(x)^{q/2-p}D^{\alpha_1}f_g D^{\alpha_2}g^{q/2}\|_{H^{s-(j+2)}(\mathcal{X},g)} \lesssim \|D^{\alpha_1}f_g\|_{H^{s-(j+2)}(\mathcal{X},g)}\|g(x)^{q/2-p}D^{\alpha_2}g^{q/2}\|_{C^{s-(j+2)}(\mathcal{X})}$$

$$\lesssim \|D^{\alpha_1} f_g\|_{H^{s-j_1}(\mathcal{X},g)} \|D^{\alpha_2} g^{q/2}\|_{C^{s-(j_2+1)}(\mathcal{X})}$$
$$\leq M\|g\|_{C^{s-1}}.$$

Summing over all $|\alpha_1| = j_1$ and $|\alpha_2| = j_2$, we obtain that l_{j_1,j_2} satisfies (13).

Next, as for $L_2(x)$, note that if $s = 1$, $L_2(x) = 0$. We want to show that for $s \geq 2$,

$$\|L_2\|_{L^2(\mathcal{X}_\epsilon, g^{p+r})} \leq C\epsilon^{s-2} M \|g\|_{C^{s-1}(\mathcal{X})}.$$

Clearly, if $s = 1$, $L_2(x) = 0$. Now, for $s \geq 2$, we have $S = s - 1$ and since $|R_{s-1}(x, x + \epsilon x^*)| \leq C\epsilon^{s-1}\|g\|_{C^{s-1}(\mathcal{X})}$ for any $x^* \in B_{\mathbf{0}}(1)$ and $d_x^j(\cdot)$ is a j-homogeneous function, we have

$$|L_2(x)| \leq \sum_{j=1}^{s-1} \frac{\epsilon^{j-2}}{j!} \int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p} \eta(\|y\|) \, |(d_x^{j_1} f_g)(y)| \cdot |R_S(x, \epsilon y + x; g^{q/2})| dy$$
$$\leq C\epsilon^{s-2} \|g\|_{C^{s-1}(\mathcal{X})} \sum_{j=1}^{s-1} \frac{1}{j!} \int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p} \eta(\|y\|) \, |(d_x^{j_1} f_g)(y)| dy.$$

Moreover, we have by Cauchy–Schwarz inequality,

$$\int_{\mathcal{X}_\epsilon} g(x)^{p+r} \left(\int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p} \eta(\|y\|) \, |(d_x^{j_1} f_g)(y)| dy \right)^2 dx$$
$$\leq \int_{\mathcal{X}_\epsilon} g(x)^{q-p+r} \left(\int_{B_{\mathbf{0}}(1)} \eta(\|y\|) \, |(d_x^{j_1} f_g)(y)|^2 dy \right) \left(\int_{B_{\mathbf{0}}(1)} \eta(\|y\|) dy \right) dx$$
$$\leq \sigma_0 \int_{B_{\mathbf{0}}(1)} \int_{\mathcal{X}_\epsilon} g(x)^{q-p+r} \eta(\|y\|) ((d^j f_g)(x))^2 dx dy$$
$$\lesssim \sigma_0^2 \int_{\mathcal{X}_\epsilon} g(x)^{p+r} ((d^j f_g)(x))^2 dx$$
$$= \sigma_0^2 \|d^j f_g\|^2_{L^2(\mathcal{X}_\epsilon, g^{p+r})},$$

where in the last step, we use the fact that $|d_x^j f(y)| \leq |d^j f(x)|$ for all $y \in B_{\mathbf{0}}(1)$. Therefore, it yields that

$$\int_{\mathcal{X}_\epsilon} g(x)^{p+r} |L_2(x)|^2 dx$$
$$\leq C \left(\epsilon^{s-2} \|g\|_{C^{s-1}(\mathcal{X})} \right)^2 \sum_{j=1}^{s-1} \int_{\mathcal{X}_\epsilon} g(x)^{p+r} \left(\frac{1}{j!} \int_{B_{\mathbf{0}}(1)} g(x)^{q/2-p} \eta(\|y\|) \, |(d_x^{j_1} f_g)(y)| dy \right)^2 dx$$
$$\leq C \left(\epsilon^{s-2} \|g\|_{C^{s-1}(\mathcal{X})} \right)^2 \sum_{j=1}^{s-1} \|d^j f_g\|^2_{L^2(\mathcal{X}_\epsilon, g^{p+r})}.$$

We obtain the desired bound.

Finally, similar to $L_2(x)$, we obtain the same bound for $L_3(x)$. Combining the obtained bounds for $L_1(x) - L_3(x)$, we obtain (12) for $t = 1$.

Now, we perform the induction step. Assuming the bound (12) holds up to some $t < s/2$, we want to show it also holds for $t + 1$, with $t + 1 < s/2$. For convenience, we introduce the following notation: for any $1 \leq j \leq l \leq s$, denote by $r_{j,l}(x) = r_{(s-l)+j}(x)$. Note again that the functions $r_{j,l}$ implicitly depend on t at the induction step thus they may vary in the below arguments from line to line. For a function $r \in H^l(\mathcal{X}_{t\epsilon}, g; C\|g\|^t_{C^{s-1}(\mathcal{X})} M)$ for some $l \leq s$, if $l \leq 2$, we have by the inductive hypothesis that for any $x \in \mathcal{X}_{(t+1)\epsilon}$,

$$L_{w,\epsilon} r(x) = r_l^l(x) \epsilon^{l-2}.$$

On the other hand, if $2 < l \leq s$, then by the inductive hypothesis, it holds that for any $x \in \mathcal{X}_{(t+1)\epsilon}$,

$$L_{w,\epsilon} r(x) = \sigma_1 \mathcal{L}_w r(x) + \sum_{j=1}^{\lfloor (l-1)/2 \rfloor - 1} r_{2j+2,l}(x) \epsilon^{2j} + r_{l,l}(x) \epsilon^{l-2}. \tag{14}$$

Then, we have

$$L_{w,\epsilon}^{t+1} f(x) = (L_{w,\epsilon} \circ L_{w,\epsilon}^{t} f)(x)$$
$$= \sigma_1^t L_{w,\epsilon} \mathcal{L}_w^t f(x) + \sum_{j=1}^{\lfloor (s-1)/2 \rfloor - k} L_{w,\epsilon} r_{2(j+t)}(x) \epsilon^{2j} + L_{w,\epsilon} r_s(x) \epsilon^{s-2t}. \tag{15}$$

In the following, we will bound these terms on the right-hand side individually. First of all, since $\mathcal{L}_w^t f \in H^{s-2t}(\mathcal{X}, g; C\|g\|_{C^{s-1}(\mathcal{X})}^t M)$, applying (14) yields

$$L_{w,\epsilon} \mathcal{L}_w^t f(x) = \sigma_1 \mathcal{L}_w^{t+1} f(x) + \sum_{j=1}^{\lfloor (s-2t-1)/2 \rfloor - 1} r_{2j+2,s-2t}(x) \epsilon^{2j} + r_{s-2t,s-2t}(x) \epsilon^{s-2t-2}$$
$$= \sigma_1 \mathcal{L}_w^{t+1} f(x) + \sum_{j=1}^{\lfloor (s-1)/2 \rfloor - (t+1)} r_{2(t+1+j)}(x) \epsilon^{2j} + r_s(x) \epsilon^{s-2(t+1)}, \tag{16}$$

where we apply the fact mentioned before that $r_{j,l}(x) = r_{(s-l)+j}(x)$.

Next, suppose $j < \lfloor (s-1)/2 \rfloor - t$. We apply (14) and obtain

$$L_{w,\epsilon} r_{2(j+t)}(x) = \sigma_1 \mathcal{L}_w r_{2(j+t)}(x) + \sum_{i=1}^{\lfloor (s-2j-2t-1)/2 \rfloor - 1} r_{2i+2,s-2(j+t)}(x) \epsilon^{2i}$$
$$+ r_{s-2(j+t),s-2(j+t)}(x) \epsilon^{s-2(j+t)-2}$$
$$= r_{2(j+t+1)}(x) + \sum_{i=1}^{\lfloor (s-1)/2 \rfloor - (j+t+1)} r_{2(i+j+t+1)}(x) \epsilon^{2i} + r_s(x) \epsilon^{s-2(j+t+1)},$$

where we use the fact that $r_{j,l}(x) = r_{(s-l)+j}(x)$ and $\sigma_1 \mathcal{L}_w r_{2(j+t)}(x) = r_{2,s-2(j+t)}(x) = r_{2(j+t+1)}(x)$. Therefore, we have

$$L_{w,\epsilon} r_{2(j+t)}(x) \epsilon^{2j} = r_{2(j+t+1)}(x) \epsilon^{2j} + \sum_{m=1}^{\lfloor (s-1)/2 \rfloor - (k+1)} r_{2(m+t+1)}(x) \epsilon^{2m} + r_s(x) \epsilon^{s-2(k+1)}, \tag{17}$$

where the last equality is by changing the variable $m = i + j$. Moreover, when $j = \lfloor (s-1)/2 \rfloor - t$, we have $2(j+t) = 2\lfloor (s-1)/2 \rfloor$ and we simply calculate that

$$L_{w,\epsilon} r_{2(j+t)}(x) \epsilon^{2j} = r_{s-2(j+t)}^{s-2(j+t)}(x) \epsilon^{s-2(j+k)} \epsilon^{2j} = r_s(x) \epsilon^{s-2(k+1)}. \tag{18}$$

Finally, according to (14), we have

$$L_{w,\epsilon} r_s(x) \epsilon^{s-2t} = r_s(x) \epsilon^{s-2(t+1)}. \tag{19}$$

Combining (16)-(19) with (15), we obtain the proof for $t+1$.

□

Recall that we write $\mathcal{X} = \mathcal{X}_{t\epsilon} \sqcup \partial\mathcal{X}_{t\epsilon}$, where for any $x \in \mathcal{X}_{t\epsilon}$, $B_x(t\epsilon) \subset \mathcal{X}$ and $\partial_{t\epsilon}\mathcal{X}$ as its complement within $\mathcal{X}$ consisting of points 'close' to the boundary.

Lemma 5.5. *For $f \in H_0^s(\mathcal{X}, g; M)$ and $t > 0$ such that $2t < s$, there exists a constant $c > 0$ not depending on M or f such that for all $\epsilon < c$,*

$$\|L_{w,\epsilon}^t f\|_{L^2(\partial_{t\epsilon}\mathcal{X}, g^{p+r})}^2 \lesssim \epsilon^{2(s-2t)} M^2.$$

Proof. Note that according to Lemma 5.2, we have

$$\|L_{w,\epsilon}^t f\|_{L^2(\partial_{t\epsilon}\mathcal{X}, g^{p+r})}^2 \lesssim \frac{1}{\epsilon^4} \|L_{w,\epsilon}^{t-1} f\|_{L^2(\partial_{t\epsilon}\mathcal{X}, g^{p+r})}^2 \lesssim \cdots \lesssim \frac{1}{\epsilon^{4t}} \|f_g\|_{L^2(\partial_{t\epsilon}\mathcal{X}, g^{p+r})}^2.$$

Therefore, it suffices to show for all $\epsilon < c$,

$$\|f_g\|^2_{L^2(\partial_{t\epsilon}\mathcal{X}, g^{p+r})} \lesssim \epsilon^{2s}\|f\|^2_{H^s(\mathcal{X},g)}. \tag{20}$$

In order to deal with f_g near the boundary, we will take a similar procedure used in Green et al. (2023, Proof of Lemma 5) and Leoni (2017, Theorem 18.1) as follows. With loss of generality, we take $t = 1$ as one can view $\epsilon < c/t$ for proving for the general case.

Step I: Local patch. We assume that for some $c_0 > 0$ and a Lipschitz mapping $\phi : \mathbb{R}^{d-1} \to [-c_0, c_0]$ and since $f \in H^s_0(\mathcal{X}, g; M)$, without loss of generality, we can assume that $f_g \in C^\infty_c(U_\psi(c_0))$ with

$$U_\psi(c_0) := \{y \in Q(0, c_0) : \psi(y^{(-d)}) \le y^{(d)}\},$$

where $Q(0, c_0)$ is the d-dimensional hypercube of side length c_0 centered at $\mathbf{0}$. Now, following step 1 in Green et al. (2023, Proof of Lemma 5) by replacing f as f_g, we have

$$\begin{aligned}|f_g(y)|^2 &\lesssim \epsilon^{2(s-1)}\left(\int_{\psi(y^{(-d)})}^{y^{(d)}} |(D^s f_g(y^{(-d)}, z))^{(d)}|dz\right)^2 \\ &\lesssim \epsilon^{2s-1}\int_{\psi(y^{(-d)})}^{y^{(d)}} |(D^s f_g(y^{(-d)}, z))^{(d)}|^2 dz.\end{aligned}$$

Then, we obtain:

$$\begin{aligned}\int_{V_\psi(\epsilon)} g(y)^{p+r}|f_g(y)|^2 dy &\lesssim \int_{Q_{d-1}(c_0)}\int_{\psi(y^{(-d)})}^{\psi(y^{(-d)})+\epsilon} |f_g(y^{(-d)}, y^{(d)})|^2 dy^{(d)} dy^{(-d)} \\ &\lesssim \epsilon^{2s-1}\int_{Q_{d-1}(c_0)}\int_{\psi(y^{(-d)})}^{\psi(y^{(-d)})+\epsilon}\int_{\psi(y^{(-d)})}^{y^{(d)}} |(D^s f_g(y^{(-d)}, z))^{(d)}|^2 dz dy^{(d)} dy^{(-d)},\end{aligned} \tag{21}$$

where $Q_{d-1}(0, c_0)$ is the d-1 dimensional hypercube of side length c_0 centered at $\mathbf{0}$. Also, by changing the integration order, it yields that

$$\begin{aligned}\int_{\psi(y^{(-d)})}^{\psi(y^{(-d)})+\epsilon}\int_{\psi(y^{(-d)})}^{y^{(d)}} |(D^s f_g(y^{(-d)}, z))^{(d)}|^2 dz dy^{(d)} &\lesssim \epsilon\int_{\psi(y^{(-d)})}^{\psi(y^{(-d)})+\epsilon} |(D^s f_g(y^{(-d)}, z))^{(d)}|^2 dz \\ &\lesssim \epsilon\int_{\psi(y^{(-d)})}^{c_0} |(D^s f_g(y^{(-d)}, z))^{(d)}|^2 dz.\end{aligned} \tag{22}$$

Combining (21) and (22), we obtain:

$$\begin{aligned}\int_{V_\psi(\epsilon)} g(y)^{p+r}|f_g(y)|^2 dy &\lesssim \epsilon^{2s}\int_{Q_{d-1}(c_0)}\int_{\psi(y^{(-d)})}^{c_0} g(y^{(-d)}, z)^q |(D^s f_g(y^{(-d)}, z))^{(d)}|^2 dz dy^{(-d)} \\ &\lesssim \epsilon^{2s}\|f\|^2_{H^s(U_\psi(c_0),g)}.\end{aligned}$$

Step 2: Rigid motion of local patch Now suppose at a point $x_0 \in \partial\mathcal{X}$, there exits a rigid motion $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T(x_0) = 0$, and a number C_0 such that we have all $C_0\epsilon \le c_0$,

$$T(Q_T(x_0, c_0) \cap \partial_\epsilon\mathcal{X}) \subseteq V_\psi(C_0\epsilon) \quad \text{and} \quad T(Q_T(x_0, c_0) \cap \mathcal{X}) = U_\psi(c_0), \tag{23}$$

where $Q_T(x_0, c_0)$ is a hypercube in $\mathbb{R}^d$ of side length c_0 centered at x_0 (not necessarily coordinate-axis-aligned). Let $v_g(y) := f_g(T^{-1}(y))$ and $v(y) := f(T^{-1}(y))$ for all $y \in U_\psi(c_0)$. Then, if $f_g \in C^\infty_c(\mathcal{X})$, we have $v_g \in C^\infty_c(U_\psi(c_0))$ such that $\|v_g\|^2_{H^s(U_\psi(c_0))} = \|f_g\|^2_{H^s(Q_T(x_0,c_0))\cap\mathcal{X}}$. Therefore, according to Step 1, we have

$$\int_{V_\psi(C_0\epsilon)} g(x)^{p+r}|v_g(y)|^2 dy \lesssim \epsilon^{2s}\|v\|^2_{H^s(U_\psi(c_0)),g}.$$

Then, it yields that

$$\begin{aligned}
&\int_{Q_T(x_0,c_0)\cap\partial_\epsilon\mathcal{X}} g^{p+r}(x)|f_g(x)|^2dx \\
&= \int_{T(Q_T(x_0,c_0)\cap\partial_\epsilon\mathcal{X})} g^{p+r}(y)|v_g(y)|^2dx \\
&\lesssim \int_{V_\psi(C_0\epsilon)} g(x)^{p+r}|v_g(y)|^2dy \\
&\lesssim \epsilon^{2s}\|v\|^2_{H^s(U_\psi(c_0),g)} \\
&\lesssim \epsilon^{2s}\|f\|^2_{H^s(Q_T(x_0,c_0)\cap\partial_\epsilon\mathcal{X},g)} \lesssim \epsilon^{2s}\|f\|^2_{H^s(\mathcal{X},g)}.
\end{aligned}$$

Step 3: Lipschitz domain. Now we arrive at the last step where we shall deal with the case: $\mathcal{X}$ is assumed to be an open, bounded subset of $\mathbb{R}^d$ with Lipschitz boundary. Again, following the procedure in Green et al. (2023, Proof of Lemma 5). In this case, for every $x_0 \in \partial\mathcal{X}$, there exists a rigid motion $T_{x_0} : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_{x_0}(x_0) = 0$, a number $c_0(x_0)$, a Lipshitz mapping $\psi_{x_0} : \mathbb{R}^{d-1} \to [-c_0(x_0), c_0(x_0)]$ and a number $C_0(x_0)$ satisfying for all $C_0(x_0)\epsilon \le c_0(x_0)$, (23) holds for replacing c_0, C_0, T, ψ by $c_0(x_0), C_0(x_0), T_{x_0}, \psi_{x_0}$ respectively. Therefore, by Step 2, we have

$$\int_{Q_{T_{x_0}}(x_0,c_0(x_0))\cap\partial_\epsilon\mathcal{X}} g^{p+r}(x)|f_g(x)|^2dx \lesssim_{x_0} \epsilon^{2s}\|f\|^2_{H^s(\mathcal{X},g)}.$$

Although the constant in the last bound depends on x_0, by compactness assumption, there exists a finite subset (denoted by $x_{0,1}, \ldots, x_{0,N}$) of the collection of hypercubes $\{Q_{T_{x_0}}(x_0, c_0(x_0)/2) : x_0 \in \partial\mathcal{X}\}$ which covers $\partial\mathcal{X}$. Then, by taking the minimum of all constants with respect to $x_{0,1}, \ldots, x_{0,N}$, we can conclude that

$$\partial_\epsilon(\mathcal{X}) \subseteq \bigcup_{i=1}^N Q_{T_{x_{0,i}}}(x_{0,i}, c_0(x_{0,i})).$$

Consequently, we have

$$\int_{\partial\mathcal{X}} g^{p+r}(x)|f_g(x)|^2dx \lesssim \sum_{i=1}^N \int_{Q_{T_{x_{0,i}}}(x_{0,i},c_0(x_{0,i}))\cap\partial_\epsilon\mathcal{X}} g^{p+r}(x)|f_g(x)|^2dx \lesssim \epsilon^{2s}\|f\|^2_{H^s(\mathcal{X},g)}.$$

Therefore, we proved the desired result (20). □

The following result presents a higher order version of Lemma 5.1 for $s > 1$ and the non-local weighted Sobolev seminorm, $\langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}}$.

Lemma 5.6. *For $f \in H^1(\mathcal{X}, g; M) \cap H^s_0(\mathcal{X}, g; M)$ with $s \in \mathbb{N}_+\backslash\{1\}$, we have*

$$\langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}} \lesssim M^2.$$

Proof of Lemma 5.6. Note here that we fix the assumption that $f \in H^1(\mathcal{X}, g; M)$ besides $f \in H^s_0(\mathcal{X}, g; M)$, which is missing in the statement of Green et al. (2023, Theorem 3). In general, it is not true for $\mathcal{X} \neq \mathbb{R}^d$ that $H^1_0(\mathcal{X}, g; M) = H^1(\mathcal{X}, g; M)$. Based on Lemma 5.1, we will prove Lemma 5.6 in a recursive way for $s > 1$. Recall that $L_{w,\epsilon}$ is self-adjoint with respect to the weighted inner product, meaning $\langle L_{w,\epsilon}f_1, f_2\rangle_{g^{p-r}} = \langle f_1, L_{w,\epsilon}f_2\rangle_{g^{p-r}}$ for any $f_1/g^r, f_2/g^r \in L^2(\mathcal{X}, g^{p+r})$. Also recall the definition of the Dirichlet energy given in Lemma 5.3, which can be stated as $E_{w,\epsilon}(f, \mathcal{X}) = 2\langle L_{w,\epsilon}f, f\rangle_{g^{p-r}}$.

Following the procedure in Green et al. (2023) [3], when $s = 2t + 1$ for $t \ge 1$, by using self-adjointness, we have

$$\langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}} = \langle L^{t+1}_{w,\epsilon}f, L^t_{w,\epsilon}f\rangle_{g^{p-r}} = \frac{1}{2}E_{w,\epsilon}(L^t_{w,\epsilon}f, \mathcal{X}).$$

[3] We remark here that the factor 2 is missing in (Green et al., 2023, Section D.4).

We divide the Dirichlet energy into two parts:

$$\begin{aligned}
&E_{w,\epsilon}(L^t_{w,\epsilon}f, \mathcal{X}) \\
&= \frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}_{t\epsilon}} \int_{\mathcal{X}_{t\epsilon}} (g(x)^{-r} f(x) - g(z)^{-r} f(z))^2 \frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2} g(z)^{1-q/2}} g(x)g(z)dxdz \\
&\quad + \frac{1}{\epsilon^{d+2}} \int_{\partial_{t\epsilon}\mathcal{X}} \int_{\partial_{t\epsilon}\mathcal{X}} (g(x)^{-r} f(x) - g(z)^{-r} f(z))^2 \frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2} g(z)^{1-q/2}} g(x)g(z)dxdz \\
&=: E_{w,\epsilon}(L^t_{w,\epsilon}f, \mathcal{X}_{t\epsilon}) + E_{w,\epsilon}(L^t_{w,\epsilon}f, \partial_{t\epsilon}\mathcal{X}),
\end{aligned}$$

where $\mathcal{X}_{t\epsilon}$ and $\partial\mathcal{X}_{t\epsilon}$ have been introduced right before Lemma 5.4 ($\partial\mathcal{X}_{t\epsilon} \subset \mathcal{X}$ consists of points $t\epsilon$-close to the boundary of $\mathcal{X}$, and $\mathcal{X}_{t\epsilon} = \mathcal{X} \setminus \partial\mathcal{X}_{t\epsilon}$).

By Jensen's inequality, we have

$$\begin{aligned}
E_{w,\epsilon}(L^t_{w,\epsilon}f, \mathcal{X}_{t\epsilon}) \le 3\sigma_1^{2t} E_{w,\epsilon}\left(\sigma_1^t L^t_w f, \mathcal{X}_{t\epsilon}\right) \\
+ \frac{6}{\epsilon^{d+2}} \int_{\mathcal{X}_{t\epsilon}} \int_{\mathcal{X}_{t\epsilon}} \left(g(x)^{-r} L^t_{w,\epsilon} f(x) - g(z)^{-r} \sigma_1^t L^t_w f(z)\right)^2 \\
\frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2} g(z)^{1-q/2}} g(x)g(z)dxdz.
\end{aligned}$$

By definition (5), we have $\mathcal{L}^t_w f \in H^1(\mathcal{X}, g; C\|g\|^t_{C^{s-1}(\mathcal{X})} M)$ for some constant $C > 0$, an application of Lemma 5.1 shows $E_{w,\epsilon}\left(\sigma_1^t L^t_w f, \mathcal{X}_{t\epsilon}\right) \lesssim M^2$. We then focus on the second term on the right-hand side of the above inequality. According to Lemma 5.4, we obtain:

$$\begin{aligned}
&\frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}_{t\epsilon}} \int_{\mathcal{X}_{t\epsilon}} \left(g(x)^{-r} L^t_{w,\epsilon} f(x) - g(z)^{-r} \sigma_1^t \mathcal{L}^t_w f(z)\right)^2 \frac{\eta\left(\frac{\|x-z\|}{\epsilon}\right)}{g(x)^{1-q/2} g(z)^{1-q/2}} g(x)g(z)dxdz \\
&\lesssim \frac{1}{\epsilon^{d+2}} \int_{\mathcal{X}_{t\epsilon}} \int_{\mathcal{X}_{t\epsilon}} g(x)^{p+r} \left(g(x)^{-r} L^t_{w,\epsilon} f(x) - g(x)^{-r} \sigma_1^t \mathcal{L}^t_w f(x)\right)^2 \eta\left(\frac{\|x-z\|}{\epsilon}\right) dxdz \\
&\lesssim \frac{1}{\epsilon^2} \int_{\mathcal{X}_{t\epsilon}} g(x)^{p-r} \left(L^t_{w,\epsilon} f(x) - \sigma_1^t \mathcal{L}^t_w f(x)\right)^2 dx \\
&\lesssim M^2.
\end{aligned}$$

Furthermore, near the boundary, according to Lemma 5.3 and Lemma 5.5, it yields that

$$E_{w,\epsilon}(L^t_{w,\epsilon}f, \partial_{t\epsilon}\mathcal{X}) \lesssim \frac{1}{\epsilon^2} \|L^t_{w,\epsilon} f\|_{L^2(\partial_{t\epsilon}\mathcal{X}, g^{p+r})} \lesssim M^2.$$

Putting all pieces above together, we obtain the proof for the case when s is odd and $t := (s-1)/2$. Similar arguments can be applied to the case when s is even and $t := (s-2)/2$. Therefore, combining all above together, we obtain for all integer $s > 1$:

$$\langle L^s_{w,\epsilon} f, f\rangle_{g^{p-r}} \lesssim M^2.$$

□

We are now in the position to prove the main results of Section 3.2.

Proof of Theorem 3.1. By Cauchy-Schwarz inequality, we have: for all $s \in \mathbb{N}_+$:

$$\|\hat{f} - f\|^2_{w,n} \le 2(\|\mathbb{E}\hat{f} - f|^2_{w,n} + \|\hat{f} - \mathbb{E}\hat{f}\|^2_{w,n}).$$

Then, according to PCR-WLE algorithm in Section 2.2, we obtain

$$\|\mathbb{E}\hat{f} - f\|^2_{w,n} = \sum_{k=K+1}^{n} \langle v_k, f\rangle^2_{w,n} \le \frac{\langle L^s_{w,n,\epsilon} f, f\rangle_{w,n}}{\lambda^s_{K+1}}, \tag{24}$$

and

$$\|\hat{f} - \mathbb{E}\hat{f}\|_{w,n}^2 = \sum_{k=1}^{K} \langle v_k, \varepsilon \rangle_{w,n}^2.$$

Since $\langle v_k, \varepsilon \rangle_{w,n}$ is normally distributed with 0 mean and variance:

$$\mathrm{Var}\langle v_k, \varepsilon \rangle_{w,n} = \frac{1}{n^2} v_k^T D^{\frac{2(p-1-r)}{q-1}} v_k, \tag{25}$$

where $\langle v_k, v_k \rangle_{w,n} = \frac{1}{n} v_k^T D^{\frac{p-1-r}{q-1}} v_k = 1$. Note that $\langle v_k/\sqrt{n}, v_k/\sqrt{n} \rangle_{g^{p-r}} = 1$, then we have

$$\min_{v_k/\sqrt{n} \in \mathbb{R}^n} \frac{1}{n} v_k^T D^{\frac{p-1-r}{q-1}} D^{\frac{p-1-r}{q-1}} v_k \tag{26}$$

is the smallest eigenvalue of the matrix $D^{\frac{p-1-r}{q-1}}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{g^{p-r}}$. As D is a diagonal matrix with the (i,i)-element as d_i, according to Section 6.1, it is bounded from below, say by a constant $C > 0$, almost surly for n large enough. Then, combining (25) and (26), we have:

$$\|\hat{f} - \mathbb{E}\hat{f}\|_{w,n}^2 = \frac{1}{n} \sum_{k=1}^{K} (\sqrt{n} \langle v_k, \epsilon \rangle_{w,n})^2,$$

with $\sqrt{n}\langle v_k, \epsilon \rangle_{w,n}$ being normal with mean 0 and variance

$$\mathrm{Var}(\sqrt{n}\langle v_k, \epsilon \rangle_{w,n}) \geq C > 0.$$

According to an exponential inequality for chi-square distributions from Laurent and Massart (2000), we obtain:

$$\mathbb{P}\left(\|\hat{f} - \mathbb{E}\hat{f}\|_{w,n}^2 \geq \frac{CK}{n} + 2\frac{\sqrt{K}}{n}\sqrt{t} + 2\frac{t}{n} \right) \leq e^{-t}. \tag{27}$$

With (24) and (27), it yields

$$\|\hat{f} - f\|_{w,n}^2 \leq \frac{\langle L_{w,n,\epsilon}^s f, f \rangle_{w,n}}{\lambda_{K+1}^s} + \frac{CK}{n}, \tag{28}$$

with probability at least $1 - e^{-K}$ if $1 \leq K \leq n$. Then, it remains to bound the empirical weighted Sobolev seminorm $\langle L_{w,n,\epsilon}^s f, f \rangle_{w,n}$ and the graph weighted Laplacian eigenvalue λ_{K+1}^s.

We will first focus on $\langle L_{w,n,\epsilon}^s f, f \rangle_{w,n}$ for $s = 1$. By definition (4), we have by symmetry:

$$\mathbb{E}\langle L_{w,n,\epsilon} f, f \rangle_{w,n} = \frac{1}{2}\mathbb{E}\left(\frac{1}{\epsilon^{d+2}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_j^{-\frac{r}{q-1}} f(X_j)|^2 d_i^{\frac{1-p}{q-1}} \frac{\eta\left(\frac{\|X_i - X_j\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_j^{1-q/2}} \right). \tag{29}$$

We would like to point out here that the normalization factor $\epsilon^{-(d+2)}$ is motivated by the fact that a factor of ϵ^{-d} is needed to scale $\eta\left(\frac{\|X_i - X_j\|}{\epsilon}\right)$ and the remaining factor, ϵ^{-2}, stabilized the squared differences of $d_i^{-\frac{r}{q-1}}$ under the expectation.

According to Section 6.1 and by conditioning on X_i and the law of iterated expectation, we have for n large enough,

$$\left| \mathbb{E}\left(\frac{1}{\epsilon^{d+2}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_j^{-\frac{r}{q-1}} f(X_j)|^2 d_i^{\frac{1-p}{q-1}} \frac{\eta\left(\frac{\|X_i - X_j\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_j^{1-q/2}} \right) - 2\langle L_{w,\epsilon} f, f \rangle_{g^{p-r}} \right| \lesssim \Delta(n, \epsilon, \eta, g) + \epsilon, \tag{30}$$

where

$$\Delta(n,\epsilon,\eta,g) := \frac{1}{n}g_{\max} + \frac{\eta(0)}{n\epsilon^d} + \frac{n-1}{n}\Big(\sqrt{\frac{|\log\epsilon|}{n\epsilon^d}} + \epsilon\Big) \to 0 \quad \text{as} \quad n\to\infty.$$

Combining (29), (30) and Lemma 5.1, we obtain:

$$\mathbb{E}\langle L_{w,n,\epsilon}f, f\rangle_{w,n} \lesssim M^2 + \Delta(n,\epsilon,\eta,g) + \epsilon.$$

Consequently, by Markov's inequality, we have: for any $\delta\in(0,1)$,

$$\langle L_{w,n,\epsilon}f, f\rangle_{w,n} \lesssim \frac{1}{\delta}\left(M^2 + \Delta(n,\epsilon,\eta,g) + \epsilon\right), \tag{31}$$

with probability at least $1-\delta$. Note that the above bound on the expected weighted Sobolev seminorm generalizes the results in Green et al. (2023) to the weighted Laplacians by some properties of KDE.

Next, we proceed to the higher order case when $s>1$ for $\langle L^s_{w,n,\epsilon}f, f\rangle_{w,n}$. We define the following difference operator:

$$D_j f(x) = (d_\cdot^{-\frac{r}{q-1}} f(x) - d_j^{-\frac{r}{q-1}} f(X_j)) d_\cdot^{\frac{1-p}{q-1}} w^\epsilon_{\cdot,j},$$

where $d_\cdot$ and $w^\epsilon_{\cdot,j}$ are defined by replacing X_i by x in both d_i and $w^\epsilon_{i,j}$. Furthermore, let $D_{\mathbf{j}}f(x) := (D_{j_1}f\circ\ldots\circ D_{j_s}f)(x)$, where $\mathbf{j}=(j_1,\ldots,j_s)\in[n]^s := \{1,\ldots,n\}^s$. Denote by $(n)^s$ the sub-collection of vectors in $[n]^s$ with no repeated indices and let by $i\mathbf{j} := (i,j_1,\ldots,j_s)$.

Following the idea of Green et al. (2023, Proof of Lemma 3), we decompose the weighted Sobolev seminorm into a U-statistic, which is an unbiased estimator of the non-local Sobolev seminorm $\langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}}$, and a pure bias term:

$$\begin{aligned}
\langle L^s_{w,n,\epsilon}f, f\rangle_{w,n} &= \frac{1}{n}\sum_{i=1}^n d_i^{\frac{p-1-r}{q-1}} L^s_{w,n,\epsilon}f(X_i)\cdot f(X_i)\\
&= \frac{1}{n\epsilon^{2s}}\sum_{i\mathbf{j}\in(n)^{s+1}} d_i^{\frac{p-1-r}{q-1}} D_{\mathbf{j}}f(X_i)\cdot f(X_i)\\
&\quad + \frac{1}{n\epsilon^{2s}}\sum_{i\mathbf{j}\in[n]^{s+1}\setminus(n)^{s+1}} d_i^{\frac{p-1-r}{q-1}} D_{\mathbf{j}}f(X_i)\cdot f(X_i)\\
&=: I_1 + I_2.
\end{aligned} \tag{32}$$

Note that there are errors in Green et al. (2023, Proof of Lemma 3) when bounding both $\mathbb{E}I_1$ and $\mathbb{E}I_2$. Specifically, in Green et al. (2023, Lemma D.3), there should not be a δ appearing in Equation D.4 by Markov's inequality and the power of ϵ should be $2s+d$. Although their final result is correct, we will fix these errors in the following proof. Now, determined by whether all $i\mathbf{j}$ are distinct, the empirical weighted Sobolev seminorm can be divided into two parts, I_1 and I_2. The first one involves all distinct indices where we make approximation by the so-called non-local weighted sobolev norm $\langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}}$; the second part focuses on the case where not all $i\mathbf{j}$ are distinct and use the fact that it is related to a connected subgraph.

As for I_1 from (32), we have

$$\begin{aligned}
\mathbb{E}I_1 &= \frac{1}{n\epsilon^{2s}}\frac{n!}{(n-s-1)!}\mathbb{E}\left(d_i^{\frac{p-1-r}{q-1}} D_{\mathbf{j}}f(X_i)\cdot f(X_i)\right)\\
&= \frac{1}{n\epsilon^{2s}}\frac{n!}{(n-s-1)!}\mathbb{E}\langle D_{\mathbf{j}}f(X_i), f(X_i)\rangle_{g^{p-r}},
\end{aligned}$$

where the operator $D_{\mathbf{j}}$ is iterated for s different times due to the fact that $i\mathbf{j}$ are all distinct. For each iteration, say $s=1$, we have

$$\mathbb{E}\langle D_j f(X_i), f(X_i)\rangle_{g^{p-r}}$$

$$= \frac{\epsilon^2}{2n}\mathbb{E}\left(\frac{1}{\epsilon^{d+2}}|d_i^{-\frac{r}{q-1}}f(X_i) - d_j^{-\frac{r}{q-1}}f(X_j)|^2 d_i^{\frac{1-p}{q-1}}\frac{\eta\left(\frac{\|X_i-X_j\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2}\tilde{d}_j^{1-q/2}}\right). \tag{33}$$

Then, plugging (30) in (33), we obtain

$$\left|\mathbb{E}\langle D_j f(X_i), f(X_i)\rangle_{g^{p-r}} - \frac{\epsilon^2}{n}\langle L_{w,\epsilon}f, f\rangle_{g^{p-r}}\right| \lesssim \frac{\epsilon^2}{2n}(\Delta(n,\epsilon,\eta,g)+\epsilon).$$

After s times iteration, it yields that

$$\left|\mathbb{E}\langle D_{\mathbf{j}} f(X_i), f(X_i)\rangle_{g^{p-r}} - \frac{\epsilon^{2s}}{n^s}\langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}}\right| \lesssim \frac{\epsilon^{2s}}{2^s n^s}(\Delta(n,\epsilon,\eta,g)+\epsilon).$$

Putting all above results back in $\mathbb{E}I_1$, we conclude that for n large enough,

$$\left|\mathbb{E}I_1 - \frac{n!}{n^{s+1}(n-s-1)!}\langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}}\right| \lesssim \frac{n!}{n^{s+1}(n-s-1)!}(\Delta(n,\epsilon,\eta,g)+\epsilon). \tag{34}$$

The Stirling's formula shows

$$\lim_{n\to\infty}\frac{n!}{n^{s+1}(n-s-1)!} = 1.$$

Therefore, by (34), we have for n large enough,

$$\mathbb{E}I_1 \lesssim \langle L^s_{w,\epsilon}f, f\rangle_{g^{p-r}} + (\Delta(n,\epsilon,\eta,g)+\epsilon).$$

According to Lemma 5.6, it yields that

$$\mathbb{E}I_1 \lesssim M^2 + (\Delta(n,\epsilon,\eta,g)+\epsilon). \tag{35}$$

We next shift our attention to I_2 in (32):

$$\frac{1}{n\epsilon^{2s}}\sum_{i\mathbf{j}\in[n]^{s+1}\setminus(n)^{s+1}} d_i^{\frac{p-1-r}{q-1}} D_{\mathbf{j}} f(X_i)\cdot(f(X_i)-f(X_{j_1})).$$

For $i\mathbf{j}$ not all distinctive, if they contains a total of $(k+1)$ distinct indices for example for $1\le k\le s-1$, we have by symmetry:

$$\sum_{i\mathbf{j}\in[n]^{s+1}\setminus(n)^{s+1}} d_i^{\frac{p-1-r}{q-1}} D_{\mathbf{j}} f(X_i)\cdot f(X_i) = \frac{1}{2}\cdot\sum_{i\mathbf{j}\in[n]^{s+1}\setminus(n)^{s+1}} d_i^{\frac{p-1-r}{q-1}} D_{\mathbf{j}} f(X_i)\cdot(f(X_i)-f(X_{j_1})).$$

Observe that in order for

$$d_i^{\frac{p-1-r}{q-1}}|D_{\mathbf{j}} f(X_i)|\cdot|f(X_i)-f(X_{j_1})|$$

to be non-zero, it must be the case that the graph $G_{n,\epsilon}(X_{i\mathbf{j}})$ which is the subgraph induced by the vertices $X_i, X_{j_1},\dots,X_{j_s}$ is complete. Since we have:

$$\begin{aligned} D_{ij}f(x) &= D_i(D_j f(x)) \\ &= D_i\left((d_\cdot^{-\frac{r}{q-1}}f(x) - d_j^{-\frac{r}{q-1}}f(X_j))d_\cdot^{\frac{1-p}{q-1}}w^\epsilon_{\cdot,j}\right) \\ &= (d_\cdot^{-\frac{r}{q-1}}D_j f(x) - d_i^{-\frac{r}{q-1}}D_j f(X_i))d_\cdot^{\frac{1-p}{q-1}}w^\epsilon_{\cdot,i}, \end{aligned}$$

then

$$|D_{j_1j_2}f(X_i)| \le \left(d_i^{-\frac{r}{q-1}}|D_{j_2}f(X_i)| + d_{j_1}^{-\frac{r}{q-1}}|D_{j_2}f(X_{j_1})|\right)d_i^{\frac{1-p}{q-1}}w^\epsilon_{i,j_1}.$$

Repeating the above computation and by induction, it yields that for $s \geq 2$,

$$|D_{\mathbf{j}} f(X_i)| \leq (s-1) d_{\max / \min}^{-\frac{(s-1)r}{q-1}} d_{\max / \min}^{\frac{(s-1)(1-p)}{q-1}} \left(w_{\max}^{\epsilon}\right)^{s-1} \sum_{j \in i\mathbf{j} \setminus \{j_s\}} |D_{j_s} f(X_j)|,$$

where $d_{\max} := \max_{i=1,\ldots,n} d_i$, $d_{\min} := \min_{i=1,\ldots,n} d_i$, $w_{\max} := \max_{i,j=1,\ldots,n} w_{i,j}$ and $d_{\max / \min}$ means it is $d_{\max}$ if $-(s-1)r/(d-1)$ (respectively $(s-1)(1-p)/(q-1)$) are positive and it is $d_{\min}$ otherwise.

According to Section 6.1, we have for n large enough, $d_{\max}$ is bounded from above and $d_{\min}$ is bounded from below a.s. and

$$w_{\max} \lesssim \frac{1}{n\epsilon^d},$$

almost surely.

Consequently, it yields that

$$\begin{aligned}
& d_i^{\frac{p-1-r}{q-1}} |D_{\mathbf{j}} f(X_i)| \cdot |f(X_i) - f(X_{j_1})| \\
&= d_i^{\frac{p-1-r}{q-1}} |D_{\mathbf{j}} f(X_i)| \cdot |f(X_i) - f(X_{j_1})| \cdot \mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}} \\
&\lesssim \frac{1}{(n\epsilon^d)^{s-1}} \sum_{j \in i\mathbf{j} \setminus \{j_s\}} \left(d_i^{\frac{p-1-r}{q-1}} |D_{j_s} f(X_j)| \cdot |f(X_i) - f(X_{j_1})| \cdot \mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}} \right) \\
&= \frac{\epsilon^2}{n^s \epsilon^{d(s-1)}} \sum_{j \in i\mathbf{j} \setminus \{j_s\}} \left(\frac{1}{\epsilon^{d+2}} d_i^{\frac{p-1-r}{q-1}} |d_j^{-\frac{r}{q-1}} f(X_j) - d_{j_s}^{-\frac{r}{q-1}} f(X_{j_s})| d_j^{\frac{1-p}{q-1}} \frac{\eta\left(\frac{\|X_j - X_{j_s}\|}{\epsilon}\right)}{\tilde{d}_j^{1-q/2} \tilde{d}_{j_s}^{1-q/2}} \right. \\
&\qquad \left. |f(X_i) - f(X_{j_1})| \mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}} \right),
\end{aligned} \tag{36}$$

where we again assign ϵ^{d+2} as a normalization factor into the expectation as (29).

Now, note that for $j = i$ in the summand on the right-hand side of (36), we have according to Section 6.1:

$$\begin{aligned}
& \mathbb{E}\left(\frac{1}{\epsilon^{d+2}} d_i^{-\frac{r}{q-1}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_{j_s}^{-\frac{r}{q-1}} f(X_{j_s})| \frac{\eta\left(\frac{\|X_i - X_{j_s}\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_{j_s}^{1-q/2}} |f(X_i) - f(X_{j_1})| \right. \\
&\qquad \left. \mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}} \right) \\
&\lesssim \mathbb{E}\left(\left(\frac{1}{\epsilon^{d+2}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_{j_s}^{-\frac{r}{q-1}} f(X_{j_s})| \frac{\eta\left(\frac{\|X_i - X_{j_s}\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_{j_s}^{1-q/2}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_{j_1}^{-\frac{r}{q-1}} f(X_{j_1})| \right. \right. \\
&\qquad \left. \left. + (\Delta(n, \epsilon, \eta, g) + \epsilon) \right) \mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}} \right) \\
&\lesssim \mathbb{E}\left(\left(\frac{1}{\epsilon^{d+2}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_{j_s}^{-\frac{r}{q-1}} f(X_{j_s})|^2 \frac{\eta\left(\frac{\|X_i - X_{j_s}\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_{j_s}^{1-q/2}} + (\Delta(n, \epsilon, \eta, g) + \epsilon) \right) \right. \\
&\qquad \left. \mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}} \right),
\end{aligned} \tag{37}$$

where the last inequality is by Cauchy–Schwarz inequality and $X_1, \ldots, X_n$ being i.i.d. data. Then, by integrating out all indices in $\mathbf{j}$ not equal to i or j_s, it yields that

$$\mathbb{E}\left(\left(\frac{1}{\epsilon^{d+2}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_{j_s}^{-\frac{r}{q-1}} f(X_{j_s})|^2 \frac{\eta\left(\frac{\|X_i - X_{j_s}\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_{j_s}^{1-q/2}} + (\Delta(n, \epsilon, \eta, g) + \epsilon) \right) \right.$$

$$\mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}}\Bigg)$$

$$\lesssim \left(C\epsilon^d g_{\max} V_d\right)^{k-1} \mathbb{E}\left(\left(\frac{1}{\epsilon^{d+2}}|d_i^{-\frac{r}{q-1}} f(X_i) - d_{j_s}^{-\frac{r}{q-1}} f(X_{j_s})|^2 \frac{\eta\left(\frac{\|X_i - X_{j_s}\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_{j_s}^{1-q/2}}\right.\right.$$

$$\left.\left. + \left(\Delta(n,\epsilon,\eta,g) + \epsilon\right)\right)\right). \tag{38}$$

Therefore, according to (37), (38), (30) and Lemma 5.1, we obtain

$$\mathbb{E}\Bigg(\frac{1}{\epsilon^{d+2}} d_i^{-\frac{r}{q-1}} |d_i^{-\frac{r}{q-1}} f(X_i) - d_{j_s}^{-\frac{r}{q-1}} f(X_{j_s})| \frac{\eta\left(\frac{\|X_i - X_{j_s}\|}{\epsilon}\right)}{\tilde{d}_i^{1-q/2} \tilde{d}_{j_s}^{1-q/2}} |f(X_i) - f(X_{j_1})|$$

$$\mathbf{1}_{\{G_{n,\epsilon}(X_{i\mathbf{j}}) \text{ is connected}\}}\Bigg)$$

$$\lesssim \epsilon^{d(k-1)} \left(M^2 + \Delta(n,\epsilon,\eta,g) + \epsilon\right). \tag{39}$$

Applying a similar approach to all $j \neq j_s$ and plugging (39) in (36) and (32), we have

$$\mathbb{E}I_2 \lesssim \frac{1}{n\epsilon^{2s}} \frac{1}{n^s \epsilon^{d(s-1)}} \sum_{k=1}^{s-1} \epsilon^{d(k-1)} \left(M^2 + \Delta(n,\epsilon,\eta,g)\right) n^{k+1}$$

$$\lesssim \frac{\epsilon^2}{n\epsilon^{2s}} \left(M^2 + \Delta(n,\epsilon,\eta,g) + \epsilon\right) \sum_{k=1}^{s-1} \frac{(n\epsilon^d)^k}{(n\epsilon^d)^s} n.$$

Note that the above sum is bounded from above when $k = s-1$ by the assumption $n\epsilon^d \geq 1$. Finally, we conclude that

$$\mathbb{E}I_2 \lesssim \frac{\epsilon^2}{n\epsilon^{2s+d}} \left(M^2 + \Delta(n,\epsilon,\eta,g) + \epsilon\right). \tag{40}$$

Finally, combining (32), (35) and (40), we obtain:

$$\mathbb{E}\langle L^s_{w,n,\epsilon} f, f\rangle_{w,n} \lesssim M^2 + \left(\Delta(n,\epsilon,\eta,g) + \epsilon\right) + \frac{\epsilon^2}{n\epsilon^{2s+d}} \left(M^2 + \Delta(n,\epsilon,\eta,g) + \epsilon\right)$$

$$\lesssim M^2 + \left(\Delta(n,\epsilon,\eta,g) + \epsilon\right),$$

where the last step is by the assumption that $\epsilon \gtrsim n^{-1/(2(s-1)+d)}$. By Markov's inequality, we have for any $\delta \in (0,1)$,

$$\langle L^s_{w,n,\epsilon} f, f\rangle_{w,n} \lesssim \frac{1}{\delta} \left(M^2 + \left(\Delta(n,\epsilon,\eta,g) + \epsilon\right)\right), \tag{41}$$

with probability at least $1 - 2\delta$. This bound can be considered as a higher order variant of (31) for $s > 1$.

Now, recall the bound (28). We have bounded the empirical weighted Sobolev seminorm by (31) and (41). It remains to bound the eigenvalues λ_{K+1}.

According to Lemma 6.1, we have:

$$\lambda_k = \lambda_k(L_{w,n,\epsilon}) \gtrsim \lambda_k(\mathcal{L}_w) \wedge \epsilon^2, \text{ for all } 2 \leq k \leq n, \tag{42}$$

with probability at least $1 - Cne^{-cn\epsilon^d}$ for some constants $C, c > 0$.

For $s = 1$, combining (28), (31) and (42), we have with probability at least $1 - \delta - Cne^{-cn\epsilon^d} - e^{-K}$ and n large enough:

$$\|\hat{f} - f\|^2_{w,n} \lesssim \frac{M^2}{\delta\left(\lambda_{K+1}(\mathcal{L}_w) \wedge \epsilon^2\right)} + \frac{K}{n}.$$

Furthermore, based on the assumption $\epsilon \lesssim K^{-1/d}$ and Proposition 6.6, the above inequality becomes:

$$\|\hat{f} - f\|_{w,n}^2 \lesssim \frac{M^2}{\delta}(K+1)^{-2/d} + \frac{K}{n}. \tag{43}$$

By balancing the two terms on the right-hand side, we pick $K = \lfloor M^2 n \rfloor^{d/(2+d)}$. Then, it yields that

$$\|\hat{f} - f\|_{w,n}^2 \lesssim \frac{1}{\delta} M^2 (M^2 n)^{-2/(2+d)}. \tag{44}$$

If $M^2 < n^{-1}$, we can take $K = 1$ and obtain from (43) that:

$$\|\hat{f} - f\|_{w,n}^2 \lesssim \frac{1}{n\delta}.$$

If $M > n^{1/d}$, we take $K = n$ and in this case, we actually have $\hat{f}(X_i) = Y_i$ for $i = 1, \ldots, n$ and

$$\|\hat{f} - f\|_{w,n}^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \lesssim C,$$

with probability at least $1 - e^{-n}$ for some constant C. Combining all above cases depending on choices of K, it yields that bound in Theorem 3.1 .

For $s > 1$, the proof follows in a similar way by considering (41) instead of (31). □

5.2 Proof Of Theorem 3.2

Proof of Theorem 3.2. Recall the construction of the estimator based on Lepski's procedure: $\hat{f}_{\mathsf{adapt}} = \hat{f}_{\hat{s},\hat{M}}$ with $\hat{s}, \hat{M}$ given in Section 3.3. Let the event $\mathcal{E}_j$ be that $\hat{s} = s_j$ and suppose $s = s_i$ for the true smooth parameter.

First of all, it suffices to consider $M \in \mathcal{D}$ by realizing that if $M \in (M_{j-1}, M_j)$, then $f \in H^s(\mathcal{X}, g; M)$ with $H^s(\mathcal{X}, g; M_{j-1}) \subset H^s(\mathcal{X}, g; M) \subset H^s(\mathcal{X}, g; M_j)$. Now, we also suppose $M = M_i$ correspondingly and consider bounding the sum:

$$\sum_{j=1}^{N_l} \left(\|\hat{f}_{s_j} - f\|_{w,n}^2 M_i^{-2} (M_i^2 n / \log n)^{2s_i/(2s_i+d)} \mathbf{1}_{\mathcal{E}_j} \right),$$

conditional on the event that the sample points $X_1, \ldots, X_n$ satisfy (28) and (42) with $K = \lfloor M_i^2 n \rfloor^{d/(2s_i+d)}$. These two statements hold with probability at least $1 - Cne^{-Cn\epsilon^d} - e^{-\lfloor M_i^2 n \rfloor^{d/(2s_i+d)}}$. As we will see, the fact that this sum does not explode, relies on the fact that the probabilities of the sets $\mathcal{E}_j$ get small as $n \to \infty$.

First, note that by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\sum_{j=i}^{N_l} \left(\|\hat{f}_{s_j} - f\|_{w,n}^2 M_i^{-2} (M_i^2 n / \log n)^{2s_i/(2s_i+d)} \mathbf{1}_{\mathcal{E}_j} \right) \\
&\le \sum_{j=i}^{N_l} \left(\|\hat{f}_{s_j} - \hat{f}_{s_i} + \hat{f}_{s_j} - f\|_{w,n}^2 M_i^{-2} (M_i^2 n / \log n)^{2s_i/(2s_i+d)} \mathbf{1}_{\mathcal{E}_j} \right) \\
&\le \sum_{j=i}^{N_l} \left(2c_0^2 \mathbf{1}_{\mathcal{E}_j} + 2 \left(\|\hat{f}_{s_i} - f\|_{w,n}^2 M_i^{-2} (M_i^2 n / \log n)^{2s_i/(2s_i+d)} \mathbf{1}_{\mathcal{E}_j} \right) \right) \\
&\le 2c_0^2 + 2 \left(\|\hat{f}_{s_i} - f\|_{w,n}^2 M_i^{-2} (M_i^2 n / \log n)^{2s_i/(2s_i+d)} \right).
\end{aligned}$$

Therefore, according to Theorem 3.1, we have: for any $\delta \in (0,1)$,

$$\sum_{j=i}^{N_l} \left(\|\hat{f}_{s_j} - f\|_{w,n}^2 M_i^{-2} (M_i^2 n / \log n)^{2s_i/(2s_i+d)} \mathbf{1}_{\mathcal{E}_j} \right) \lesssim \frac{1}{\delta},$$

with probability at least $1-\delta\log^{-2s_i/(2s_i+d)} n - Cne^{-Cn\epsilon^d} - e^{-\lfloor M_i^2 n\rfloor^{d/(2s_i+d)}}$.

Next, we consider the other part when $j<i$:

$$\sum_{j=1}^{i-1}\left(\|\hat{f}_{s_j}-f\|_{w,n}^2 M_i^{-2}(M_i^2 n/\log n)^{2s_i/(2s_i+d)}\mathbf{1}_{\mathcal{E}_j}\right). \tag{45}$$

By the definition, on the event $\mathcal{E}_j$, there exists $s'\in\mathcal{B}$ with $s'<s_i$ such that $\|\hat{f}_{s_i}-\hat{f}_{s'}\|_{w,n} > c_0 M'^{-2}(M'^2 n/\log n)^{-s'/(2s'+d)}$. This means $\|\hat{f}_{s_i}-\hat{f}_{s'}\|_{w,n}^2 M'^{-2}(M'^2 n/\log n)^{2s'/(2s'+d)} > c_0^2$. By triangle inequality, this implies we have either $\|\hat{f}_{s_i}-f\|_{w,n}^2 M'^{-2}(M'^2 n/\log n)^{2s'/(2s'+d)} > c_0^2/4$ or $\|\hat{f}_{s'}-f\|_{w,n}^2 M'^{-2}(M'^2 n/\log n)^{2s'/(2s'+d)} > c_0^2/4$. Then, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_j)\le\sum_{l=1}^{i-1}\Big(&\mathbb{P}\left(\|\hat{f}_{s_i}-f\|_{w,n}^2 M_l^{-2}(M_l^2 n/\log n)^{2s_l/(2s_l+d)}>c_0^2/4\right)\\ &+\mathbb{P}\left(\|\hat{f}_{s_l}-f\|_{w,n}^2 M_l^{-2}(M_l^2 n/\log n)^{2s_l/(2s_l+d)}>c_0^2/4\right)\Big).\end{aligned} \tag{46}$$

Since $l<i$, we have $f\in H^{s_i}(\mathcal{X},g;M_l)\subset H^{s_l}(\mathcal{X},g;M_l)$ for all $l<i$. Therefore, it suffices to focus on the concentration inequality of $\hat{f}_{s_l}$ to f, i.e., bounding

$$\mathbb{P}\left(\|\hat{f}_{s_l}-f\|_{w,n}^2 M_l^{-2}(M_l^2 n/\log n)^{2s_l/(2s_l+d)}>c_0^2/4\right). \tag{47}$$

Note that the key problem here is the rate of convergence of $\|\hat{f}_{s_j}-f\|_{w,n}^2$ in (45) does not match the rate $(n/\log n)^{2s_i/(2s_i+d)}$ given there. However, this can be dealt with by controlling the probability of the event $\mathcal{E}_j$. The strategy here is we need a better concentration inequality than what has been proven previously as (41) otherwise the probability of the event $\mathcal{E}_j$ will not decay to 0. Observe that the concentration (41): for n large enough and with probability smaller than $1-2\delta$,

$$\langle L_{w,n,\epsilon}^s f,f\rangle_{w,n}\lesssim\delta^{-1}M^2,$$

is from the application of Markov's inequality with

$$\mathbb{E}\langle L_{w,n,\epsilon}^s f,f\rangle_{w,n}\lesssim M^2,$$

for n large enough. While bounding the first moment gives a concentration inequality with probability $1-2\delta$, establishing a higher moment bound, e.g. the second moment, would result in a better concentration inequality with higher probability similar to Green et al. (2021, Proposition 1), which fits in our proof technique.

Starting with $s=1$ and similar to (30), we have: for n large enough,

$$\begin{aligned}&\mathrm{Var}\langle L_{w,n,\epsilon}f,f\rangle_{w,n}\\ &\lesssim\mathrm{Var}\left(\frac{1}{2}\frac{1}{n^2\epsilon^{d+2}}\sum_{i,j=1}^{n}(g(X_i)^{-r}f(X_i)-g(X_j)^{-r}f(X_j))^2 g(X_i)^{1-p}\frac{\eta\left(\frac{\|X_i-X_j\|}{\epsilon}\right)}{g(X_i)^{1-q/2}g(X_j)^{1-q/2}}\right).\end{aligned} \tag{48}$$

For $i,j\in 1,\dots,n$, let

$$V_{ij}:=(g(X_i)^{-r}f(X_i)-g(X_j)^{-r}f(X_j))^2 g(X_i)^{1-p}\frac{\eta\left(\frac{\|X_i-X_j\|}{\epsilon}\right)}{g(X_i)^{1-q/2}g(X_j)^{1-q/2}}.$$

We have:

$$\mathrm{Var}\left(\sum_{i,j=1}^{n}(g(X_i)^{-r}f(X_i)-g(X_j)^{-r}f(X_j))^2 g(X_i)^{1-p}\frac{\eta\left(\frac{\|X_i-X_j\|}{\epsilon}\right)}{g(X_i)^{1-q/2}g(X_j)^{1-q/2}}\right)$$

$$= \sum_{i,j=1}^{n} \sum_{l,m=1}^{n} \mathrm{Cov}(V_{ij}, V_{lm}).$$

Now, consider the following four scenarios depending on the cardinality of $\{i, j, l, m\}$.

- If $|\{i, j, l, m\}| = 4$, since V_{ij} and V_{lm} are independent, we have $\mathrm{Cov}(V_{ij}, V_{lm}) = 0$.
- If $|\{i, j, l, m\}| = 3$, without loss of generality, say $i = l$, we have by Lipschitz condition,

$$\begin{aligned}\mathrm{Cov}(V_{ij}, V_{im}) &\leq \mathbb{E}[V_{ij}V_{im}] \\ &\lesssim \epsilon^{2d+4}M^4.\end{aligned}$$

- If $|\{i, j, l, m\}| = 2$, without loss of generality, say $i = l$ and $j = m$, similarly, we obtain

$$\begin{aligned}\mathrm{Cov}(V_{ij}, V_{ij}) &\leq \mathbb{E}V_{ij}^2 \\ &\lesssim \epsilon^{d+4}M^4.\end{aligned}$$

- If $|\{i, j, l, m\}| = 1$, we have $V_{ij} = V_{lm} = 0$.

Plugging the above results in (48), it yields that for n large enough,

$$\mathrm{Var}\langle L_{w,n,\epsilon}f, f\rangle_{w,n} \lesssim \frac{1}{4n^4\epsilon^{2d+4}}\left(n^3\epsilon^{2d+4}M^4 + n^2\epsilon^{d+4}M^4\right) \lesssim n^{-1}M^4,$$

where the last step follows by the assumption that $n\epsilon^d \geq 1$. Then, by Markov's inequality, we obtain: for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(|\langle L_{w,n,\epsilon}f, f\rangle_{w,n} - \mathbb{E}\langle L_{w,n,\epsilon}f, f\rangle_{w,n}| \geq \frac{1}{\delta}M^2\right) \lesssim \frac{\delta^2}{n}. \tag{49}$$

Combining (49) and (31), we conclude that for n large enough,

$$\langle L_{w,n,\epsilon}f, f\rangle_{w,n} \lesssim \frac{1}{\delta}M^2$$

holds with probability not less than $1 - \frac{\delta}{n^2}$. Furthermore, following a similar argument in Lemma 5.6, one can show the above high-probability bound also holds for the case $s > 1$. Thus, under the additional Lipschitz assumption that $|f_g(x) - f_g(x')| \leq M\|x - x'\|$, we establish a better bound for the empirical weighted Sobolev seminorm: for all $s \in \mathbb{N}_+$ and n large enough,

$$\langle L^s_{w,n,\epsilon}f, f\rangle_{w,n} \lesssim \frac{1}{\delta}M^2,$$

with probability at least $1 - C\frac{\delta^2}{n}$.

Conditional on the event that the sample points $X_1, \ldots, X_n$ satisfy (28) and (42) with $K = \lfloor M^2 n\rfloor^{d/(2s+d)}$, following the proof of Theorem 3.1 to obtain (44) by using the better concentration inequality we derived above instead, we have for n large enough,

$$\|\hat{f} - f\|^2_{w,n} \lesssim \frac{1}{\delta}M^2(M^2n)^{-2s/(2s+d)},$$

with probability at least $1 - C\delta^2 n^{-1} - Cne^{-Cn\epsilon^d} - e^{-\lfloor M^2 n\rfloor^{d/(2s+d)}}$ under the minimax optimal setting for M.

Now, returning to our mission (47), by setting $\delta^{-1} = c_0^2/4 \cdot \log^{2s_l/(2s_l+d)} n$, we have:

$$\begin{aligned}&\mathbb{P}\left(\|\hat{f}_{s_l} - f\|^2_{w,n}M_l^{-2}(M_l^2 n/\log n)^{2s_l/(2s_l+d)} > c_0^2/4\right) \\ &\leq 16Cc_0^{-4}n^{-1}\log^{-2s_l/(2s_l+d)} n + Cne^{-Cn\epsilon^d} + e^{-\lfloor M^2_{\min} n\rfloor^{d/(2s+d)}}.\end{aligned}$$

With (46), we obtain:

$$\mathbb{P}(\mathcal{E}_j) \le 16Cc_0^{-4}n^{-1}\log^{1-2s_{\min}/(2s_{\min}+d)} n + Cne^{-Cn\epsilon^d}\log n + e^{-\lfloor M_{\min}^2 n\rfloor^{d/(2s+d)}}\log n.$$

Combining the above result with (45) and noting that on $\mathcal{E}_j^c$, $\mathbf{1}_{\mathcal{E}_j} = 0$, it yields that

$$\sum_{j=1}^{i-1}\left(\|\hat{f}_{s_j} - f\|_{w,n}^2 M_i^{-2}(M_i^2 n/\log n)^{2s_i/(2s_i+d)}\mathbf{1}_{\mathcal{E}_j}\right) \lesssim \frac{1}{\delta},$$

with probability at least

$$1 - \delta\log^{-2s_i/(2s_i+d)} n - 16Cc_0^{-4}n^{-1}\log^{2-2s_{\min}/(2s_{\min}+d)} n - Cne^{-Cn\epsilon^d}\log^2 n - e^{-\lfloor M_{\min}^2 n\rfloor^{d/(2s+d)}}\log^2 n.$$

□

6 AUXILIARY RESULTS

In the subsequent two sections, we introduce some important properties of KDE and eigenvalues of the weighted Laplacian matrices $L_{w,n,\epsilon}$ and the weighted Laplacian operators $\mathcal{L}_w$ used in the previous proof respectively.

6.1 Property Of Kernel Density Estimation

Consider a Kernel density estimator (KDE) on $\mathcal{X}$:

$$g_n(x) := \frac{1}{n\epsilon^d}\sum_{j=1}^{n}\eta\left(\frac{\|x - X_j\|}{\epsilon}\right),$$

where η is a kernel function.

In Giné and Guillou (2002), it has been proven that the above KDE satisfies the following almost sure convergence:

$$\|g_n(x) - \mathbb{E}g_n(x)\|_\infty = O_{a.s.}\left(\sqrt{\frac{|\log\epsilon|}{n\epsilon^d}}\right),$$

given the assumption that the kernel η satisfies the kernel VC-type condition (A4) and see Remark 3.2 for more details.

As for the bias, it is well-known that there exists a boundary effect on KDE due to the fact that (with probability 1) all the samples lie in the support of the density. However, when we are far enough away from the boundary such that $B_x(\epsilon) \subset \mathcal{X}$, we have

$$\begin{aligned}|\mathbb{E}g_n(x) - g(x)| &= \left|\int_{\mathcal{X}}\frac{1}{\epsilon^d}\eta\left(\frac{\|x-y\|}{\epsilon}\right)g(y)dy - g(x)\right| \\ &\le \int_{\|z\|\le 1}\eta(\|z\|)|g(x+\epsilon z) - g(x)|dz \\ &\lesssim \epsilon\int_{\mathbb{R}^d}\|z\|\eta(\|z\|)dz \lesssim \epsilon,\end{aligned}$$

where the last step is by the assumption that g is Lipschitz. As a result, for such values of x,

$$\|g_n(x) - g(x)\|_\infty = O_{a.s.}\left(\sqrt{\frac{|\log\epsilon|}{n\epsilon^d}} + \epsilon\right).$$

When x is near the boundary, i.e., $B_x(\epsilon) \not\subset \mathcal{X}$, we have $X_i \in B_x(\epsilon)$ with probability less than $C\epsilon$ for some constant $C > 0$. Then:

$$\mathbb{E}g_n(x) \le g_{\max}\int_{\|z\|\le 1}\eta(\|z\|)dz < \infty,$$

and

$$\mathbb{E}g_n(x) = \int_{\{\|z\|\le 1\}\cap\{x+\epsilon z\in\mathcal{X}\}} \eta(\|z\|)g(x+\epsilon z)dz \ge g_{\min}\int_{\{\|z\|\le 1\}\cap\{x+\epsilon z\in\mathcal{X}\}} \eta(\|z\|)dz > 0,$$

under the assumption (A1) on $\mathcal{X}$. Therefore, we have for all $x \in \mathcal{X}$, $g_n(x)$ is bounded from above and below a.s. for n large enough.

By conditioning on X_i and the law of total probability, we have for all $i \in [n]$ and $B_\epsilon(X_i) \in \mathcal{X}$,

$$\Delta^-(n,\epsilon,\eta,g) \le g_n(X_i) - g(X_i) \le \Delta^+(n,\epsilon,\eta,g),$$

almost surely with

$$\Delta^-(n,\epsilon,\eta,g) := -\frac{1}{n}g_{\max} + \frac{\eta(0)}{n\epsilon^d} - \frac{n-1}{n}\Delta(n,\epsilon),$$
$$\Delta^+(n,\epsilon,\eta,g) := -\frac{1}{n}g_{\min} + \frac{\eta(0)}{n\epsilon^d} + \frac{n-1}{n}\Delta(n,\epsilon),$$

and

$$\Delta(n,\epsilon) := \sqrt{\frac{|\log\epsilon|}{n\epsilon^d}} + \epsilon.$$

Since we are seeking a high-probability bound in Theorems 3.1, it is not necessarily required to have an exact estimation near the boundary, which happens with probability of the order ϵ. However, various approaches including data reflection, transformations, boundary kernels and local likelihood, have been proposed for boundary correction.

6.2 Property Of Eigenvalues

In this section, we focus on introducing some results on the eigenvalues of the weighted Laplacian $L_{n,w,\epsilon}$ and the weighted Laplacian operator $\mathcal{L}_w$ based on analysis in Calder and Trillos (2022), Green et al. (2021).

6.2.1 Transportation Distance Between Measures

For a probability measure G defined on $\mathcal{X}$ and a map $T : \mathcal{X} \to \mathcal{X}$, denote by $T_\sharp G$ the push-forward of G by T, i.e., the measure such that for any Borel subset $U \subseteq \mathcal{X}$, it holds that

$$T_\sharp G(U) := G(T^{-1}(U)).$$

When $T_\sharp G$ is taken as the empirical measure of G denoted by G_n, T is called the transportation map between G and G_n and we define the ∞-transportation distance between G and G_n as

$$d_\infty(G,G_n) := \inf_{T:T_\sharp G = G_n} \|T - \mathrm{Id}\|_{L^\infty(G)}, \tag{50}$$

where Id is the identity mapping. We denote by $\tilde{T}$ the optimal ∞-optimal transport map (∞-OT map) between G and G_n, i.e,, the map that achieves the infimum (50).

Now, following Green et al. (2021), let

$$\tilde{\delta} := \max\{n^{-1/d}, C\epsilon\},$$

where $C > 0$ is some constant not depending on n and we also let $\theta > 0$ be some constant not depending on n. We present the following result from Green et al. (2021).

Proposition 6.1 (cf. Proposition 3 of Green et al. (2021)). *Under the assumptions (A1) and (A2), with probability greater than $1 - Cne^{-Cn\theta^2\tilde{\delta}^d}$, there exists a probability measure $\tilde{G}_n$ with density $\tilde{g}_n$ such that*

$$d_\infty(G_n,\tilde{G}_n) \le C\tilde{\delta},$$

and such that

$$\|g - \tilde{g}_n\|_\infty \le C(\theta + \tilde{\delta}),$$

where $C > 0$ is some constant not depending on n.

6.2.2 Discretization And Interpolation Maps

The key procedure adopted in Calder and Trillos (2022) is to construct two maps: a discretization map $\tilde{\mathcal{P}} : L^2(G) \to L^2(\tilde{G}_n)$ and an interpolation map $\tilde{\mathcal{I}} : L^2(\tilde{G}_n) \to L^2(G)$, that are "almost" isometries.

For $X_i, i = 1, \ldots, n$, define

$$\tilde{U}_i := \tilde{T}^{-1}(\{X_i\}).$$

Then, we define the contractive discretization map $\tilde{\mathcal{P}} : L^2(G) \to L^2(\tilde{G}_n)$ by

$$(\tilde{\mathcal{P}} f)(X_i) := n \cdot \int_{\tilde{U}_i} f(x)\tilde{g}_n(x)dx.$$

Moreover, the interpolation map $\tilde{\mathcal{I}} : L^2(\tilde{G}_n) \to L^2(G)$ is given by

$$\tilde{\mathcal{I}} u := \Lambda_{\epsilon - 2\tilde{\delta}}(\tilde{\mathcal{P}}^* u).$$

Here, $\tilde{\mathcal{P}}^* = u \circ \tilde{T}$ is the adjoint of $\tilde{\mathcal{P}}_n$, i.e.,

$$(\tilde{\mathcal{P}}^* u)(x) = \sum_{j=1}^{n} u(x_i)\mathbf{1}_{x \in U_i},$$

and $\Lambda_{\epsilon - 2\tilde{\delta}}$ is a kernel smoothing operator with respect to a kernel K (defined below) with the bandwidth $\epsilon - 2\tilde{\delta}$. The kernel K is defined by

$$K(x, y) := \frac{1}{\epsilon^d} \zeta\left(\frac{\|x - y\|}{\epsilon}\right),$$

where

$$\zeta(t) := \frac{1}{\sigma_1} \int_t^\infty \eta(s) s ds.$$

Then, define the operator Λ_h, for $h > 0$, by

$$\Lambda_h f(x) := \frac{1}{\tau(x)} \int_{\mathcal{X}} K(x, y) f(y) g(y) dy,$$

where $\tau(x) := \int_{\mathcal{X}} K(x, y) g(y) dy$ is a normalization factor.

Furthermore, we define the Dirichlet energies:

$$b_{w,\epsilon}(u) := \langle L_{w,n,\epsilon} u, u \rangle_{g^{p-r}},$$

and

$$D_w(f) := \begin{cases} \int_{\mathcal{X}} \|\nabla f_g(x)\|^2 g(x)^q dx & \text{if } f \in H^1(\mathcal{X}, g), \\ \infty & \text{o.w.} \end{cases}$$

Clearly, when $w = (p, q, r) = (1, 2, 0)$, the above Dirichlet energies become the ones associated with the unnormalized Laplacian, i.e., $w = (p, q, r) = (1, 2, 0)$:

$$b_\epsilon(u) := \langle (\tilde{D} - \tilde{W}) u, u \rangle,$$

and

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

The following two propositions from Green et al. (2021), whose proof is based on Proposition 6.1, shows the fact that discretization map $\tilde{\mathcal{P}}$ and interpolation map $\tilde{\mathcal{I}}$ are almost isometries.

Proposition 6.2 (cf. Proposition 4 of Green et al. (2021)). *With probability at least* $1 - Cne^{-Cn\theta^2\tilde{\delta}^d}$, *we have for any* $f \in L^2(\mathcal{X})$,

$$b_\epsilon(\tilde{\mathcal{P}}f) \le C(1 + C(\theta + \tilde{\delta}))\left(1 + C\frac{\tilde{\delta}}{\epsilon}\right)\sigma_1 \cdot D_2(f),$$

and for any $u \in L^2(G_n)$,

$$\sigma_1 D_2(\tilde{\mathcal{I}}u) \le C(1 + C(\theta + \tilde{\delta}))\left(1 + C\frac{\tilde{\delta}}{\epsilon}\right) \cdot b_\epsilon(u).$$

Proposition 6.3 (cf. Proposition 5 of Green et al. (2021)). *With probability at least* $1 - Cne^{-Cn\theta^2\tilde{\delta}^d}$, *we have for any* $f \in L^2(\mathcal{X})$,

$$\left|\|f\|^2_{L^2(G)} - \|\tilde{\mathcal{P}}f\|^2_{L^2(G_n)}\right| \le C\tilde{\delta}\|f\|_{L^2(G)}\sqrt{D_2(f)} + C(\theta + \tilde{\delta})\|f\|^2_{L^2(G)},$$

and for any $u \in L^2(G_n)$,

$$\left|\|u\|^2_{L^2(G_n)} - \|\tilde{\mathcal{I}}u\|^2_{L^2(G)}\right| \le C\epsilon\|u\|_{L^2(G_n)}\sqrt{b_\epsilon(u)} + C(\theta + \tilde{\delta})\|u\|^2_{L^2(G_n)}.$$

Now, as we consider the Dirichlet energies $b_{w,\epsilon}(u)$ and $D_w(f)$ for the weighted Laplacian. Note that by the boundedness assumption of the density g, we have there exist constants $C > 0$ and $C' > 0$ such that

$$C'\int_{\mathcal{X}} \|\nabla f_g(x)\|^2 g(x)^q dx \le \int_{\mathcal{X}} \|\nabla f_g(x)\|^2 g(x)^2 dx \le C\int_{\mathcal{X}} \|\nabla f_g(x)\|^2 g(x)^q dx.$$

Also, with transformation $v := D^{-r/(q-1)}u$ for $q \neq 1$, we have

$$\langle L_{w,n,\epsilon}u, u\rangle_{g^{p-r}} = \langle (D - W)v, v\rangle.$$

This also holds for $q = 1$ by definition (3). According to Section 6.1, we obtain that there exist constants $C > 0$ and $C' > 0$ such that for large n, almost surely,

$$C' b_{w,\epsilon}(u) \le b_\epsilon(u) \le C b_{w,\epsilon}(u).$$

Consequently, following the proof in Green et al. (2021), we present the following propositions parallelling Proposition 6.2 and 6.3 associated with the weighted case.

Proposition 6.4. *With probability at least* $1 - Cne^{-Cn\theta^2\tilde{\delta}^d}$, *we have for any* $f \in L^2(\mathcal{X}, g^{p-r})$,

$$b_\epsilon(\tilde{\mathcal{P}}f) \le C(1 + C(\theta + \tilde{\delta}))\left(1 + C\frac{\tilde{\delta}}{\epsilon}\right)\sigma_1 \cdot D_2(f),$$

and for any $u \in L^2(G_n)$,

$$\sigma_1 D_2(\tilde{\mathcal{I}}u) \le C(1 + C(\theta + \tilde{\delta}))\left(1 + C\frac{\tilde{\delta}}{\epsilon}\right) \cdot b_\epsilon(u),$$

where $C > 0$ *is some constant not depending on* n *or* f.

Proposition 6.5. *With probability at least* $1 - Cne^{-Cn\theta^2\tilde{\delta}^d}$, *we have for any* $f \in L^2(\mathcal{X}, g^{p-r})$,

$$\left|\|f\|^2_{L^2(\mathcal{X},g^{p-r})} - \|\tilde{\mathcal{P}}f\|^2_{w,n}\right| \le C\tilde{\delta}\|f\|_{L^2(\mathcal{X},g^{p-r})}\sqrt{D_w(f)} + C(\theta + \tilde{\delta})\|f\|^2_{L^2(\mathcal{X},g^{p-r})} + \Delta(n, \epsilon, \eta, g) + \epsilon,$$

and for any $u \in L^2(G_n)$,

$$\left| \|u\|_{w,n}^2 - \|\tilde{\mathcal{I}}u\|_{L^2(\mathcal{X},g^{p-r})}^2 \right| \leq C\epsilon \|u\|_{w,n} \sqrt{b_{w,\epsilon}(u)} + C(\theta + \tilde{\delta})\|u\|_{w,n}^2 + \Delta(n,\epsilon,\eta,g) + \epsilon,$$

where $C > 0$ *is some constant not depending on* n *or* f *and*

$$\Delta(n,\epsilon,\eta,g) = \frac{1}{n} g_{\max} + \frac{\eta(0)}{n\epsilon^d} + \frac{n-1}{n}\Delta(n,\epsilon),$$
$$\Delta(n,\epsilon) := \sqrt{\frac{|\log \epsilon|}{n\epsilon^d}} + \epsilon.$$

Also, we state the following Weyl's law whose proof follows Dunlop et al. (2020, Lemma 7.10).

Proposition 6.6 (Weyl's law)**.** *There exist constants* $C, C' > 0$ *such that*

$$C' l^{2/d} \leq \lambda_l(\mathcal{L}_w) \leq C l^{2/d},$$

for all $l \geq 2$.

Therefore, by following Green et al. (2021, Proof of Lemma 2) except that we replace Propositions 6.2 and 6.3 by Propositions 6.4 and 6.5, we obtain the following bound for the eigenvalues.

Lemma 6.1. *Under the assumptions (A1) and (A2), there exist constant* $C, C' > 0$ *and* $N > 0$ *such that for* $n \geq N$ *and* $C(\log n/n)^{1/d} \leq \epsilon \leq C$*, with probability larger than* $1 - Cne^{-Cn\epsilon^d}$*, it holds that*

$$C' \min\{l^{2/d}, \epsilon^{-2}\} \leq \lambda_l(L_{w,n,\epsilon}) \leq C \min\{l^{2/d}, \epsilon^{-2}\},$$

for all $2 \leq l \leq n$.