Information and Inference: A Journal of the IMA (2021) **00**, 1–69
https://doi.org/10.1093/imaiai/iaab019

Convergence of graph Laplacian with kNN self-tuned kernels

Xiuyuan Cheng†
Department of Mathematics, Duke University, Durham, NC 27708, USA
†Corresponding author. Email: xiuyuan.cheng@duke.edu

and

Hau-Tieng Wu
Department of Mathematics and Department of Statistical Science, Duke University

[Received on 8 November 2020; revised on 1 February 2021; accepted on 29 June 2021]

Kernelized Gram matrix W constructed from data points $\{x_i\}_{i=1}^N$ as $W_{ij} = k_0(\frac{\|x_i - x_j\|^2}{\sigma^2})$ is widely used in graph-based geometric data analysis and unsupervised learning. An important question is how to choose the kernel bandwidth σ, and a common practice called self-tuned kernel adaptively sets a σ_i at each point x_i by the k-nearest neighbor (kNN) distance. When x_is are sampled from a d-dimensional manifold embedded in a possibly high-dimensional space, unlike with fixed-bandwidth kernels, theoretical results of graph Laplacian convergence with self-tuned kernels have been incomplete. This paper proves the convergence of graph Laplacian operator L_N to manifold (weighted-)Laplacian for a new family of kNN self-tuned kernels $W_{ij}^{(\alpha)} = k_0(\frac{\|x_i - x_j\|^2}{\epsilon \hat{\rho}(x_i)\hat{\rho}(x_j)})/\hat{\rho}(x_i)^\alpha \hat{\rho}(x_j)^\alpha$, where $\hat{\rho}$ is the estimated bandwidth function by kNN and the limiting operator is also parametrized by α. When $\alpha = 1$, the limiting operator is the weighted manifold Laplacian Δ_p. Specifically, we prove the point-wise convergence of $L_N f$ and convergence of the graph Dirichlet form with rates. Our analysis is based on first establishing a C^0 consistency for $\hat{\rho}$ which bounds the relative estimation error $|\hat{\rho} - \bar{\rho}|/\bar{\rho}$ uniformly with high probability, where $\bar{\rho} = p^{-1/d}$ and p is the data density function. Our theoretical results reveal the advantage of the self-tuned kernel over the fixed-bandwidth kernel via smaller variance error in low-density regions. In the algorithm, no prior knowledge of d or data density is needed. The theoretical results are supported by numerical experiments on simulated data and hand-written digit image data.

Keywords: graph Laplacian; manifold learning; k-nearest neighbor density estimator.

1. Introduction

Kernelized Gram matrix computed from data vectors $\{x_i\}_{i=1}^N$ in $\mathbb{R}^D$ has been a pivotal tool in kernel methods [39], graph-based manifold learning and geometric data analysis [1,2,11,12,46,52] and semi-supervised learning [24,25,36,45], among others. Applications range broadly from model reduction of chemical systems [13,38,43,47] to general data visualization [7,23,29,55]. The graph *affinity matrix* W, which is real symmetric and has non-negative entries, can be viewed as weights on the edges of a weighted undirected graph, denoted as $\mathcal{G} = (V, E)$, $V = \{1, \cdots, N\}$ and $E = \{(i,j), W_{ij} > 0\}$. The kernelized affinity matrix W takes the form of a kernelized Gram matrix, that is, $W_{ij} = K(x_i, x_j)$ for some real symmetric kernel function $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. In particular, a widely used setting is

$$W_{ij} = k_0\left(\frac{\|x_i - x_j\|^2}{\epsilon}\right), \tag{1.1}$$

for some univariate kernel function k_0 and some $\epsilon > 0$. For example, $k_0(r) = e^{-r}$ gives the Gaussian affinity W.

Given an affinity matrix W, the un-normalized and normalized (random-walk) graph Laplacian matrices are usually constructed as $(D - W)$ and $I - D^{-1}W$, respectively, where D, called the *degree matrix*, is the diagonal matrix with $D_{ii} = \sum_{j=1}^{N} W_{ij}$. The row-stochastic matrix $D^{-1}W$ gives the transition law of a random walk on the graph $\mathcal{G}$, which is a discrete diffusion process on the graph [32]. The eigenvalues and eigenvectors of the graph Laplacian matrix provide a dimension-reduced representation of the data samples and are used for downstream tasks like clustering and dimension-reduced embedding. Variants of the basic form (1.1) include adaptive kernel bandwidth [37,58], anisotropic kernel [6,10,43,47], adoption of landmark sets [4,28,41], kernel normalization schemes [31,57] and neural network approaches [18,34,40]. The current paper focuses on the adaptive bandwidth problem and the analysis of the kernelized graph Laplacian.

The convergence of the graph Laplacian to a certain limiting operator as the graph size (number of data samples) $N \to \infty$ is a classical theoretical problem. When x_i's are sampled from a low-dimensional manifold $\mathcal{M}$ embedded in the ambient space $\mathbb{R}^D$, the convergence of the graph Laplacian to a differential operator on the manifold $\mathcal{M}$ has been proved in several places, when $N \to \infty$ and $\epsilon \to 0$ under a joint limit [2,3,11,22,42,50], and more recently [9,15,51]. Of particular importance is when the limiting operator recovers the manifold Laplace–Beltrami operator $\Delta_{\mathcal{M}}$ or the weighted Laplace operator

$$\Delta_p := \Delta_{\mathcal{M}} + \frac{\nabla_{\mathcal{M}} p}{p} \cdot \nabla_{\mathcal{M}}, \tag{1.2}$$

where $\nabla_{\mathcal{M}}$ denote the manifold derivative and p is the density of a positive measure (not necessarily integrated to 1, i.e., a probability density). Both $\Delta_{\mathcal{M}}$ and Δ_p are intrinsic operators associated with different measures independent of the particular embedding of $\mathcal{M}$ in $\mathbb{R}^D$. The weighted Laplacian Δ_p is the Fokker–Plank operator of the diffusion process on the manifold $\mathcal{M}$, and its spectral decomposition in $(\mathcal{M}, pdV)$ reveals key physics quantities of the stochastic process, e.g., the low-lying eigenfunctions of Δ_p indicate the meta-stable states of the diffusion process [16,33]. Thus, when the discrete diffusion process on the finite-sample graph has a continuous limit, the graph Laplacian that approximates the limiting operator Δ_p can be applied to data-driven analysis of dynamical systems and clustering analysis of data clouds [35]. The asymptotic analysis of kernelized graph Laplacian thus lays the theoretical foundation for applications of graph Laplacian methods in high-dimensional data analysis.

A problem in the affinity matrix construction (1.1) is the choice of the scalar parameter ϵ, or $\sigma := \sqrt{\epsilon}$, which is called the *kernel bandwidth*. In practice, especially when data vectors are in high-dimensional space or the sampling density is not even, the choice of ϵ can be challenging and the performance of kernel Laplacian methods may also become sensitive to the choice, see, e.g., [37]. A common practice to overcome the issue of choosing ϵ is called 'self-tuning', that is, setting a personalized bandwidth for each point x_i and then using these adaptive bandwidths to compute the kernelized affinity. Specifically, the original self-tune spectral clustering method [58] considered the affinity matrix as

$$W_{ij} = k_0\left(\frac{\|x_i - x_j\|^2}{\hat{R}_i \hat{R}_j}\right), \tag{1.3}$$

where $\hat{R}_i$ equals the distance from x_i to its k-nearest neighbor (kNN) in the dataset X itself, where k is a parameter chosen by the user. While several works in the literature have addressed such kernel construction, to the knowledge of the authors, the same type of results of graph Laplacian convergence as for the fixed-bandwidth kernel has only been partially established, particularly when the kernel bandwidth is unknown and needs to be estimated from data. A more detailed review of related works is given in Section 1.2.

In this paper, we introduce the following family of graph affinity matrices, defined for $\alpha \in \mathbb{R}$, $\epsilon > 0$,

$$W_{ij}^{(\alpha)} := k_0\left(\frac{\|x_i - x_j\|^2}{\epsilon \hat{\rho}(x_i)\hat{\rho}(x_j)}\right)\frac{1}{\hat{\rho}(x_i)^\alpha \hat{\rho}(x_j)^\alpha}, \tag{1.4}$$

where the kernel function k_0 is non-negative and satisfies certain regularity and decay conditions (cf. Assumption 3.1). We also split the dataset into X and Y, and the stand-alone Y is used to estimate $\hat{\rho}(x_i)$ for each x_i in X, where $\hat{\rho}$ is a function mapping from $\mathbb{R}^D$ to $\mathbb{R}$, called the (estimated) *bandwidth function*. Specifically, $\hat{\rho}$ is normalized from the kNN distance $\hat{R}_i$ by $\hat{\rho}(x_i) = \hat{R}_i(\frac{1}{m_0}\frac{k}{N_y})^{-1/d}$, k is the parameter in kNN, $N_y = |Y|$ and $m_0 > 0$ is a constant with analytical expression. The normalization in $\hat{\rho}(x_i)$ by $(\frac{1}{m_0}\frac{k}{N_y})^{-1/d}$ is to guarantee that $\hat{\rho}$ has an $O(1)$ limit. In view of (1.4), (1.3) is a special case where $\alpha = 0$ and $\epsilon = \left(\frac{1}{m_0}\frac{k}{N_y}\right)^{2/d}$. The usage of a stand-alone Y to estimate $\hat{\rho}$ reduces dependence and in practice can reduce variance error (cf. Section 4.3). Note that while the above definition of $\hat{\rho}$ involves the manifold dimensionality d, the practical algorithm (cf. Algorithm 1) does not require knowledge or estimation of d. We summarize the algorithm in Section 1.1, where we introduce the construction of $\hat{\rho}$ and choice of parameters in more detail.

The theoretical results of our work are twofold. To summarize, suppose $N = N_x$ and $N_y = \Theta(N)$.

• We prove that when data are sampled according to a smooth density p on a smooth compact d-dimensional manifold $\mathcal{M}$, $\hat{\rho}$ uniformly converges to $\bar{\rho} = p^{-1/d}$ on $\mathcal{M}$, where the point-wise relative error $|\hat{\rho} - \bar{\rho}|/\bar{\rho}$ is uniformly bounded with high probability (w.h.p.) by (cf. Theorem 2.3)

$$\varepsilon_\rho = O((k/N)^{2/d}, \sqrt{\log N/k}).$$

The choice of k that balances the two errors is $k \sim N^{1/(1+d/4)}$, which leads to $\varepsilon_\rho = O(N^{-1/(2+d/2)})$ (cf. Remark 2.1). In particular, the constant in front of the second term (the 'variance error') is independent from the sampling density p. This result augments previous analysis of kNN estimation in literature, and the bound of the relative error (rather than absolute error) illustrates a different bias-variance error trade-off between the kNN estimator $\hat{\rho}$ and the standard fixed-bandwidth kernel density estimation (KDE) (cf. Remark 2.2). The relative error bound is useful for our analysis of the self-tuned kernel Laplacian.

• Conditioning on the good event of an accurate $\hat{\rho}$ estimation, we establish a series of convergence results of kernelized graph Laplacians with self-tuned kernels. Specifically, we prove

- The convergence of the graph Dirichlet form (cf. Theorem 3.3), where the error is

$$O(\epsilon, \varepsilon_\rho, N^{-1/2}\epsilon^{-d/4}).$$

TABLE 1 *List of default notations*

$\mathcal{M}$	d-dimensional manifold in $\mathbb{R}^D$
p	data sampling density on $\mathcal{M}$, uniformly bounded between p_{min} and p_{max}
$\Delta_{\mathcal{M}}$	Laplace–Beltrami operator, also as Δ
Δ_p	weighted Laplace operator
$\nabla_{\mathcal{M}}$	manifold gradient, also as ∇
$\bar{\nabla}$	gradient in ambient space $\mathbb{R}^D$
$\bar{\rho}$	$p^{-1/d}$, population bandwidth function, uniformly bounded between $\rho_{min} := p_{max}^{-1/d}$ and $\rho_{max} := p_{min}^{-1/d}$
$\hat{\rho}$	estimated bandwidth function
ε_ρ	error bound of $\sup_{\mathcal{M}} \lvert\hat{\rho} - \bar{\rho}\rvert/\bar{\rho}$
k	k-nearest neighbor in kNN
W	general kernelized affinity matrix
D	degree matrix, $D_{ii} = \sum_j W_{ij}$
ϵ	kernel bandwidth parameter used in theoretical analysis
σ_0	kernel bandwidth parameter used in Algorithm 1
α	the power in $\hat{\rho}^\alpha$ used to normalized the self-tuned kernel
k_0, h	function $\mathbb{R} \to \mathbb{R}$, h used in kNN estimation, k_0 used to construct affinity kernel
X	dataset used for computing W
Y	dataset used for estimating $\hat{\rho}$
$\hat{R}_i$	distance to k-nearest neighbor of x_i in X
$\hat{R}$	$\hat{R}(x)$ is the distance to k-nearest neighbor of x in Y
N_x	number of samples in X
N_y	number of samples in Y
N	N_x or N_y depending on context
m_0	$m_0[h] := \int_{\mathbb{R}^d} h(\lvert u\rvert^2)\, \mathrm{d}u$
m_2	$m_2[h] := \frac{1}{d}\int_{\mathbb{R}^d} \lvert u\rvert^2 h(\lvert u\rvert^2)\, \mathrm{d}u$
$\hat{K}$	normalized self-tuned kernel function $\mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ parametrized by α
$W^{(\alpha)}$	W computed using $\hat{K}$ parametrized by α
$\mathcal{L}^{(\alpha)}$	limiting differential operator parametrized by α
L_N	graph Laplacian operator of W
f	function on $\mathcal{M}$, also denote the vector $\{f(x_i)\}_i \in \mathbb{R}^{N_x}$
$\mathcal{E}_p$	differential Dirichlet form of Δ_p
$\mathcal{E}^{(\alpha)}$	kernelized Dirichlet form of kernel $\hat{K}$
E_N	graph Dirichlet form of W
$\hat{p}$	estimated density
ε_p	error bound of $\sup_{\mathcal{M}} \lvert\hat{p} - p\rvert/p$
β	the power in $\hat{p}^\beta$ used to normalized the fixed-bandwidth kernel
ρ	general kernel bandwidth function
$G_\epsilon^{(\rho)}$	kernel integral operator with variable bandwidth ρ
$\tilde{\rho}$	perturbed bandwidth function from ρ
	Asymptotic notations
$O(\cdot)$	$f = O(g)$: $\lvert f\rvert \leqslant C\lvert g\rvert$ in the limit, $C > 0$, $O^{[a]}(\cdot)$ declaring the constant dependence on a

Continued

TABLE 1 *Continued*

$\Theta(\cdot)$	$f = \Theta(g)$: for $f, g \geqslant 0$, $C_1 g \leqslant f \leqslant C_2 g$ in the limit, $C_1, C_2 > 0$, $\Theta^{[a]}(\cdot)$ declaring the constant dependence on a
$\sim$	$f \sim g$ same as $f = \Theta(g)$
$o(\cdot)$	$f = o(g)$: for $g > 0$, $\lvert f\rvert/g \to 0$ in the limit, $o^{[a]}(\cdot)$ declaring the constant dependence on a
$\Omega(\cdot)$	$f = \Omega(g)$: for $f, g > 0, f/g \to \infty$ in the limit

When the superscript $^{[a]}$ is $^{[1]}$, it declares that the constants are absolute ones. By definition, for finite m, $O(g_1) + \cdots + O(g_m) = O(|g_1| + \cdots + |g_m|)$, which is also denoted as $O(g_1, ..., g_m)$.

The choice of ϵ to balance bias and variance errors is $\epsilon \sim N^{-1/(2+d/2)}$, which makes $O(\epsilon, N^{-1/2}\epsilon^{-d/4}) = O(N^{-1/(2+d/2)})$, same as the order of ε_ρ with the optimal scaling of k (cf. Remark 3.1).

- The convergence of $L_N f(x)$ for the (re-normalized form of) random-walk and un-normalized graph Laplacian operators L_N, respectively (cf. Theorems 3.5 and 3.6), where the error is

$$O(\epsilon, \varepsilon_\rho/\epsilon, N^{-1/2}\epsilon^{-1/2-d/4}).$$

A weak convergence result in Theorem 3.7 shows that the error is $O(\epsilon, \varepsilon_\rho, N^{-1/2}\epsilon^{-1/2})$.

These results are compared with the counterparts for fixed bandwidth kernel, in terms of the Dirichlet form convergence (cf. Theorem 3.4) and the point-wise operator convergence (cf. Theorem 3.8).

As for how N_y should scale with N_x, if we set N_x and N_y independently, for the graph Dirichlet form convergence result, the overall error is optimized when $N_y \sim N_x$ (cf. Remark 3.1). The other rates, e.g., the point-wise convergence rate, lead to different theoretical optimal choices of N_y with respect to N_x. We empirically study the splitting of stand-alone Y in Section 4.3 with further discussion.

A key difference comparing the self-tuned kernel and the fixed-bandwidth kernel lies in the constant dependence on density p in front of the variance term. For example, in Theorem 3.8, the fixed-bandwidth kernel leads to a variance bound proportional to $p^{-1/2}$, while the factor is $p^{1/d}$ in Theorem 3.6 for the self-tuned kernel. The negative factor $-1/2$ suggests a large variance error at places where $p(x)$ is small, reflecting the difficulty of the fixed bandwidth kernel in such cases. The positive factor $1/d$ for the self-tuned kernel suggests an improvement, which can be intuitively expected, from a theoretical perspective.

Table 2 gives a roadmap of analysis to facilitate the reading of Sections 2 and 3. Our theoretical results are supported by numerical experiments in Section 4. Besides model (1.4), we also propose another affinity kernel using mixed normalization by $\hat{\rho}$ and a density estimator $\hat{p}$ that recovers the Laplace–Beltrami operator and does not require knowledge of d, and we give empirical results in Section 4.4. Apart from simulated manifold data, we apply the self-tuned diffusion kernels to the MNIST dataset of hand-written digit images [**49**], where the self-tuned kernel shows a better stability than the fixed-bandwidth kernel at data points sampled at low-density places, as is consistent with the theory.

Notations. A list of default notations is provided in Table 1. We use superscripts in big-O notations to declare dependence of the implied constants. In this work, we mainly track the dependence on the data density function p, and we treat constants depending on $\mathcal{M}$, h, k_0, including d, as absolute ones so as to simplify presentation. For $O^{[\cdot,\mathcal{M},h,\mathrm{x}]}(\cdot)$, we may omit $(\mathcal{M}, h)$ in the superscript and write as $O^{[\mathrm{x}]}(\cdot)$ so as to track dependence on x only. Specific constant dependence can be recovered from proof.

ALGORITHM 1. kNN Self-tuned kernel graph Laplacian (with stand-alone Y)

Input: datasets X and Y, and algorithm parameters k, $1 < k < N_y$, $\sigma_0 > 0$, $\alpha \in \mathbb{R}$.

Output: kNN values $\hat{R}$, graph affinity matrix W, degree matrix D, the eigenpairs (Ψ, Λ).

External: subroutine EIG (EIG-GEN) that solves eigen-problem (generalized eigen-problem) for real symmetric matrices.

1: **function** SELFTUNEDKERNEL(X, Y, k, σ_0, α)

2: $N_x \leftarrow \text{size}(X)$, $N_y \leftarrow \text{size}(Y)$

3: Compute kNN distances $\hat{R}_i \leftarrow \|x - y^{(x,k)}\|$, where $y^{(x,k)}$ is the k-nearest neighbor of x in Y.

4: Compute the matrix W by

$$W_{ij} = k_0\left(\frac{\|x_i - x_j\|^2}{\sigma_0^2 \hat{R}_i \hat{R}_j}\right)\frac{1}{\hat{R}_i^\alpha \hat{R}_j^\alpha} \tag{1.5}$$

$\triangleright W$ can be constructed as a sparse matrix

5: Compute the degree matrix D, $D_{ii} \leftarrow \sum_j W_{ij}$

6: Compute the eigenvalue/eigenvectors (Ψ, Λ) of either (1) L_{un} by EIG$(D - W)$ or (2) $L_{rw'}$ by EIG-GEN$(D - W, DD^2_{\hat{R}})$, where the diagonal matrix $D_{\hat{R}} = \text{diag}\{\hat{R}_i\}_{i=1}^{N_x}$

7: **return** $\hat{R}$, W, D, (Ψ, Λ)

8: **end function**

1.1 *Summary of algorithm*

The algorithm to compute the self-tuned kernel graph Laplacian on dataset X with a stand-alone dataset Y to estimate the kNN bandwidth function is summarized in Algorithm 1. It has the following parameters to be specified by the user,

- k: the k in kNN bandwidth estimation. By Theorem 2.3, the choice of k that balances the bias and variance errors is theoretically $k \sim N_y^{1/(1+d/4)}$.
- α: the kernel is normalized by $\hat{R}^\alpha$, where α is a real number. Different choice of α leads to a family of different limiting operators depending on α, as is analyzed in Section 3. In particular, choosing $\alpha = 1$ recovers the weighted Laplacian Δ_p.
- σ_0: the kernel bandwidth parameter. To compare with the notation in Section 3, $\sigma_0 \hat{R}_i = \sqrt{\epsilon}\hat{\rho}(x_i)$, and that is $\epsilon = \sigma_0^2(\frac{1}{m_0[h_{knn}]}\frac{k}{N_y})^{2/d}$.

Choosing $\sigma_0 = \Theta(1)$ corresponds to $\epsilon \sim (k/N_y)^{2/d}$. Under the optimal scaling of k which is $k \sim N_y^{1/(1+d/4)}$ (cf. Remark 2.1), the above scaling of ϵ is also the optimal one to balance errors of estimating the Dirichlet form (cf. Theorem 3.3, Remark 3.1). The algorithm does not require any prior

knowledge of the intrinsic dimensionality d, which can be applied to general manifold data. We also introduce a variant form of kernel matrix to recover $\Delta_{\mathcal{M}}$ in Section 4.4.

A common usage, as in the original self-tune kernel in [58], is to estimate $\hat{R}_i$ from the dataset X itself rather than a stand-alone Y. The analysis of the former case will be more complicated, as estimating kNN from X introduces more dependence across the kernel matrix entries, while a stand-alone Y allows a two-step analysis: showing the uniform point-wise convergence of the kNN bandwidth function first and then analyzing the kernel matrix W conditioning on a good event of Y, as we do in the current paper. Empirically, we find that using a stand-alone Y gives a comparable result and can improve the performance by reducing the variance error when N_y is large (Section 4.3). This suggests the usage of extra data samples in the bandwidth estimation when they are not used in the kernel matrix construction due to memory or computational constraints.

1.2 *Related works*

kNN estimated kernel bandwidth was used in the original self-tune-based algorithm, particularly for the spectral clustering purpose [58]. To fully understand the role of the kNN estimated kernel bandwidth, an understanding of its relationship with the nearest-neighbor density estimator (NNDE), an approach closely related to but different from the well-known KDE, is necessary. NNDE was studied in the classical statistical literature dating back to 60s [19,27,30], and more general variable bandwidth KDE was studied in 90s [20,48]. The uniform convergence with probability (w.p.) 1 was proved in [14]. Our result bounds the relative error $|\hat{\rho} - \bar{\rho}|/\bar{\rho}$ uniformly w.h.p., which implies uniform convergence w.p. 1 and is under the manifold data density setting.

With this relationship in mind, to our knowledge, the first paper dealing with the asymptotical behavior of self-tune-based algorithm in the manifold setup is [50]. The limiting operator of self-tune kernel has been identified in following a general framework of kernel construction of the graph Laplacian. In particular, the uniform convergence of NNDE in [14] was used to derive the limiting operator of kNN constructed graph Laplacian. However, the proof is without error rate, and the impact of NNDE has not been fully analyzed. A recent paper addressing this issue is [5], where the convergence rate has been proved. However, the formulation in [5] assumes knowledge of the desired bandwidth function to use, or that of the density function, and the impact of NNDE is not discussed. Also, the algorithm in [5] needs to estimate the intrinsic dimension d if not given, which may be difficult in practice.

Self-tune-based algorithms are natural generalizations of those with fixed bandwidths. When the bandwidth is fixed, its asymptotical analysis has been widely discussed, particularly under the manifold setup. For example, the point-wise convergence of the graph Laplacian to the Laplace–Beltrami operator, or more general weighted Laplace–Beltrami operator, has been extensively discussed in [2,11,21,42]. The spectral convergence of the graph Laplacian to the (weighted-) Laplace–Beltrami operator is more challenging and has attracted several attentions. See, for example, [3,17,44,51,54,56]. Recently, the spectral convergence in the L^2 sense has been provided with rates in [9]; the spectral convergence in the L^∞ sense, as well as the uniform heat kernel reconstruction, has been provided with rates in [15]. We establish a point-wise convergence of the graph Laplacian operator and convergence of the graph Dirichlet form for the kNN self-tuned kernel graph Laplacian with rates.

Other approaches of adaptive kernel bandwidth choice include multiscale singular value decomposition (SVD) [13,26,38]. A bandwidth selection method based on preserving data geometric information was proposed in [37]. These methods can be computationally more expensive than kNN self-tuned bandwidth.

2. kNN estimation of kernel bandwidth

In this section, we prove the uniform convergence of the kNN constructed bandwidth function $\hat{\rho}$, which is computed from a stand-alone dataset Y, to $\bar{\rho} = p^{-1/d}$ w.h.p. and in terms of relative error (cf. Theorem 2.3). We simplify notation by setting $N = N_y$ in this section. All proofs are in Appendix.

Let $\mathcal{M}$ denote the low-dimensional manifold and dV the volume element of $\mathcal{M}$. When $\mathcal{M}$ is orientable, dV is the Riemann volume form; otherwise, dV is the measure associated with the local volume form. In both cases, $(\mathcal{M}, dV)$ is a measure space. More differential geometry set-ups are provided in Appendix A.2.

ASSUMPTION 2.1 (Assumption on manifold $\mathcal{M}$ and data density p). (A1) $\mathcal{M}$ is a d-dimensional C^∞ and compact manifold without boundary, isometrically embedded in $\mathbb{R}^D$ via ι. When there is no danger of confusion, we use the same notation x to denote $x \in \mathcal{M}$ and $\iota(x) \in \mathbb{R}^D$.

(A2) $p \in C^\infty(\mathcal{M})$ and is uniformly bounded both from below and above, that is, $\exists p_{min}, p_{max} > 0$ s.t.

$$0 < p_{min} \leqslant p(x) \leqslant p_{max} < \infty, \quad \forall x \in \mathcal{M}.$$

Smoothness of $\mathcal{M}$ and p suffices most application scenarios and theoretically can be relaxed by standard functional approximation techniques. For simplicity, we consider smooth $\mathcal{M}$ and p only.

2.1 *kNN construction of $\hat{\rho}$*

Given $Y = \{y_j\}_{j=1}^N$ (recall that $N = N_y$ here), the kNN-estimated bandwidth function $\hat{\rho}(x)$ is a scalar field on $x \in \mathcal{M}$ computed from Y, defined as

$$\hat{\rho}(x) := \hat{R}(x) \left(\frac{1}{m_0[h]} \frac{k}{N} \right)^{-1/d}, \quad \hat{R}(x) := \inf_r \left\{ r > 0, \text{ s.t. } \sum_{j=1}^N \mathbf{1}_{\{\|y_j - x\| < r\}} \geqslant k \right\}, \tag{2.1}$$

where m_0 is a functional defined for function h on $[0, \infty)$ sufficiently decayed as $m_0[h] := \int_{\mathbb{R}^d} h(|u|^2)\, \mathrm{d}u$. We use m_0 to denote the scalar when not to emphasize the dependence on the function h. Note that for $h = \mathbf{1}_{[0,1)}$, $m_0[h]$ equals the volume of unit d-ball. The definition (2.1) is equivalent to that $\hat{R}(x) = \|x - y^{(k,x)}\|$, where $y^{(k,x)}$ is the k-nearest neighbor of x in Y. The following lemma gives a direct proof of the piecewise differentiability and Lipschitz continuity of the knn-constructed $\hat{R}$. The Lipschitz constant of $\hat{R}$ is important in our proof of the uniform convergence of $\hat{\rho}$.

LEMMA 2.1 Suppose Y has distinct data points y_j and $1 < k < N$. Then $\hat{R}$ defined in (2.1) is Lipschitz continuous on $\mathbb{R}^D$ with $\mathrm{Lip}_{\mathbb{R}^D}(\hat{R}) \leqslant 1$. Moreover, $\hat{R}$ is C^∞ on $\mathbb{R}^D \backslash E$, where E is a finite union of (D-1)-hyperplanes (finitely many points when $D = 1$).

One may consider variants of kNN-estimator. Specifically, a generalization of (2.1) can be $\hat{R}(x) = \inf_r \left\{ r > 0, \text{ s.t. } \sum_{j=1}^N h\left(\frac{\|x - y_j\|^2}{r^2} \right) \geqslant k \right\}$, where one can introduce weights proportional to the distance $\|x - y_j\|$ by considering a more general h. The definition (2.1) is equivalent to taking $h = \mathbf{1}_{[0,1)}$. One advantage of the classical kNN-estimator (2.1) is its efficient computation by the fast kNN algorithm. We are not aware of any other widely used weighted version of kNN-estimator; thus, we postpone the possible extension to larger class of h to future work.

TABLE 2 *Roadmap of analysis*

Parameters	Estimation of bandwidth $\hat{\rho}$ (k, N_y)	Convergence of graph Laplacian (ϵ, N_x)		
			kNN self-tuned	fixed-bandwidth
Results	Proposition 2.2 + Lemma 2.1 ⇒ Theorem 2.3	Dirichlet form	Theorem 3.3	Theorem 3.4
		Point-wise	L'_{rw}: Theorem 3.5 L_{un}: Theorem 3.6	L_{rw}: Theorem 3.8
		Weak form	Theorem 3.7	-
		Needed lemmas: Lemma 3.1 ⇒ Proposition 3.2 ⇒ Theorem 3.3 & 3.7 Lemma 3.2 ⇒ Theorem 3.5 & 3.6		

2.2 C^0 *consistency of* $\hat{\rho}$

The concentration of $\hat{\rho}$ at $\bar{\rho}$ at a point x_0 is a result of the concentration of the independent sum in (2.1), which we prove in the following proposition:

PROPOSITION 2.2 Under Assumption 2.1, if as $N \to \infty$, $k = o(N)$ and $k = \Omega(\log N)$, then, for any $s > 0$, when N is sufficiently large, for any $x \in \mathcal{M}$, w.p. $> 1 - 2N^{-s}$,

$$\frac{|\hat{\rho}(x) - \bar{\rho}(x)|}{\bar{\rho}(x)} = O^{[p]}\left(\left(\frac{k}{N}\right)^{2/d}\right) + \frac{3}{d}\sqrt{\frac{s \log N}{k}}, \tag{2.2}$$

where the constant in $O^{[p]}(\cdot)$ is determined by p, the threshold for large N depends on p and s and both are uniform for all x.

Combined with the global Lipschitz continuity of $\hat{\rho}$ (Lemma 2.1) and a bound of the covering number of $\mathcal{M}$ (Lemma A.4), we are ready to prove the main result of this section:

THEOREM 2.3 Under Assumption 2.1, $\hat{\rho}$ defined as in (2.1). If as $N \to \infty$, $k = o(N)$ and $k = \Omega(\log N)$, then when N is sufficiently large, w.p. higher than $1 - N^{-10}$,

$$\sup_{x \in \mathcal{M}} \frac{|\hat{\rho}(x) - \bar{\rho}(x)|}{\bar{\rho}(x)} = O^{[p]}\left(\left(\frac{k}{N}\right)^{2/d}\right) + \frac{3\sqrt{13}}{d}\sqrt{\frac{\log N}{k}},$$

and the right-hand side (r.h.s.) is $o^{[p]}(1)$.

REMARK 2.1 In the error bound, the $O((k/N)^{2/d})$ term is the 'bias' error, and the $O(\sqrt{\log N/k})$ term is the 'variance' error. To balance the two errors, k should be chosen according to $k^{-1/2} \sim (k/N)^{2/d}$ (where we omit the $\sqrt{\log N}$ factor), and that is $k \sim N^{1/(1+d/4)}$. In this scaling, the constant in front theoretically depends on p and generally is impractical to estimate.

REMARK 2.2 One can compare Theorem 2.3 to the estimation error bound of a fixed bandwidth KDE estimator of p, e.g., for $\epsilon > 0$,

$$\hat{p}(x) := \frac{\epsilon^{-d/2}}{m_0[h_{kde}]} \frac{1}{N_y} \sum_{j=1}^{N_y} h_{kde}\left(\frac{\|x - y_j\|^2}{\epsilon}\right), \tag{2.3}$$

where $h_{kde} : \mathbb{R}_+ \to \mathbb{R}$ is usually a non-negative regular function. When $h_{kde} \geqslant 0$ and satisfies Assumption A.3, by analyzing the bias and variance errors of the independent sum in (2.3) and using Lemma A.5, one can verify that $\hat{p}(x) = p(x) + O^{[p]}(\epsilon) + O^{[1]}\left(p(x)^{1/2}\sqrt{\frac{\log N}{N\epsilon^{d/2}}}\right)$. This gives that

$$\frac{|\hat{p}(x) - p(x)|}{p(x)} = O^{[p]}(\epsilon) + O^{[1]}\left(p(x)^{-1/2}\sqrt{\frac{\log N}{N\epsilon^{d/2}}}\right). \tag{2.4}$$

Note that the variance error in (2.4) has a factor of $p(x)^{-1/2}$, while for $\hat{\rho}$ by kNN the variance error term is uniformly bounded for all x by $O^{[1]}(\sqrt{\log N/k})$ and the constant is independent of $p(x)$. This difference between kNN $\hat{\rho}$ and fixed-bandwidth KDE $\hat{p}$ is numerically verified in Section 4.1 (Fig. 1). The comparison shows that the fixed-bandwidth KDE estimator $\hat{p}$ and the kNN estimator $\hat{\rho}$ conduct a different trade-off between the bias and variance errors: at place where $p(x)$ is small, intuitively, the kNN estimator trades the bias error for a smaller variance error and may have an advantage. For both estimators, the overall estimation error (bias + variance) depends on the choice of the parameters ϵ and k, respectively. The specific comparison would involve a more explicit bias error analysis, where the constant in $O^{[p]}(\cdot)$ involves higher order derivatives of p and is not further pursued here.

2.3 C^1 *divergence of* $\hat{\rho}$

As shown in the proof of Lemma 2.1, for any $x \in \mathbb{R}^D \backslash E$, $|\bar{\nabla}\hat{R}(x)| = 1$, where $\bar{\nabla}$ denotes the gradient in the ambient space $\mathbb{R}^D$. Thus,

$$|\bar{\nabla}\hat{\rho}(x)| = \left(\frac{1}{m_0[h]}\frac{k}{N_y}\right)^{-1/d}, \quad \forall x \in \mathbb{R}^D \backslash E,$$

which is $\Omega(1)$ as $\frac{k}{N_y} \to 0$. This means that $\nabla_{\mathcal{M}}\hat{\rho}(x)$ point-wise diverges almost everywhere and cannot have point-wise consistency to $\nabla_{\mathcal{M}}\bar{\rho}(x)$, which is $O(1)$.

While the l-th $\mathbb{R}^D$-derivative of $\hat{\rho}$ can be bounded to be $O((\frac{k}{N})^{-l/d})$ (Lemma A.7), this C^1 inconsistency of $\hat{\rho}$ by kNN estimation poses challenge to the graph Laplacian convergence because the limiting operator (see Section 3.1) involves $\nabla_{\mathcal{M}}\rho$ when ρ is a deterministic bandwidth function [5]. On the other hand, the wide usage of self-tuned diffusion kernel in spectral clustering and spectral embedding suggests that the kNN-estimated $\hat{\rho}$ can lead to a consistent estimator of certain limiting manifold differential operators though the C^0 consistency of $\hat{\rho}$ alone may not be able to directly prove that.

In Section 3, we will show theoretically that the point-wise consistency of the graph Laplacian operator has a different and worse error rate than the consistency of the graph Dirichlet form, where consistency in both cases is obtained but under different conditions on k, N_y related to the bandwidth

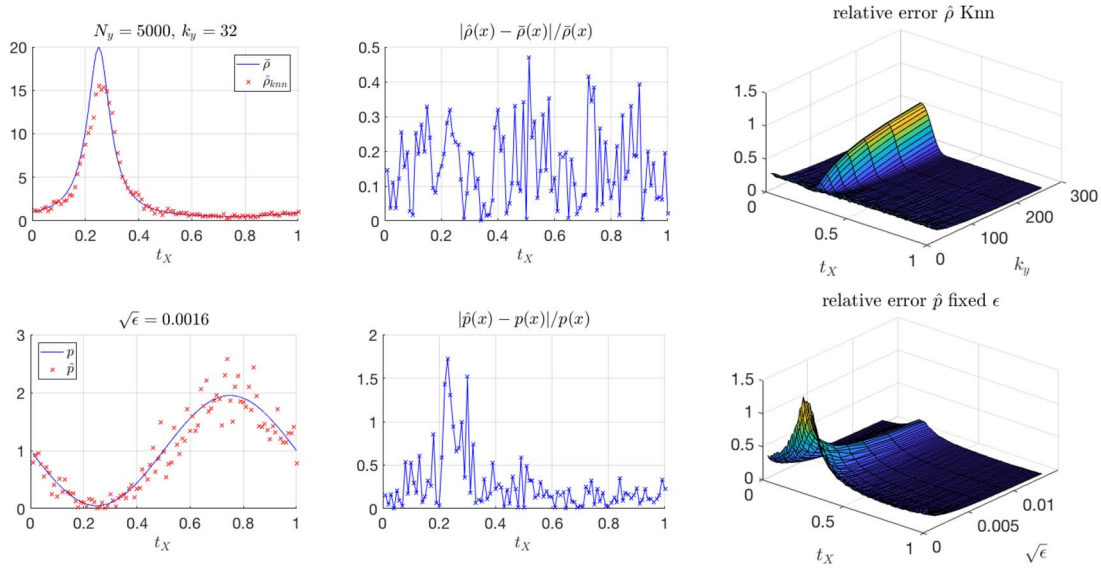

FIG. 1. kNN estimation of $\bar{\rho} = p^{-1/d}$ and KDE estimation of p for $x \in S^1$ embedded in $\mathbb{R}^2$, and $0 \leqslant t_X \leqslant 1$ is the intrinsic coordinate (arclength). (Top) The left two plots show a typical realization of $\hat{\rho}$ by kNN defined in (2.1) compared with $\bar{\rho}$, and the right plot shows the relative error for varying values of k_y, $N_y = 5000$, averaged over 500 runs. (Bottom) Same plots for $\hat{p}$ defined in (2.3) compared with p, and relative error for varying values of ϵ.

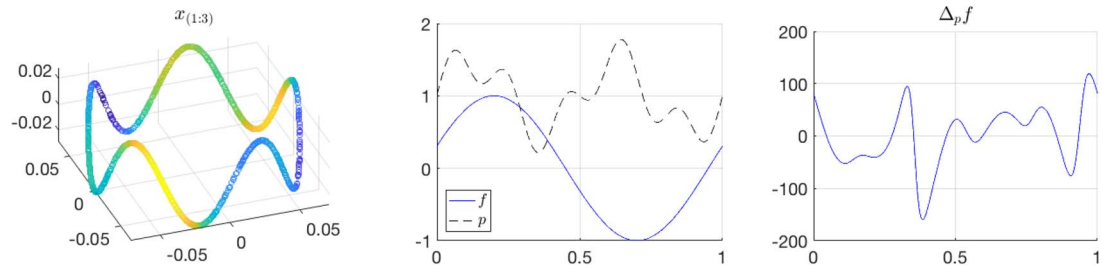

FIG. 2. Data in $\mathbb{R}^4$ lying on a 1D closed curve of length 1. (Left) First 3 coordinates of 2000 samples with color indicating the density function p. (Middle) The density function p and a function f, and (Right) $\Delta_p f$, all plotted v.s. the intrinsic coordinate (the arclength) on $[0, 1]$.

parameter ϵ. The distinction is also revealed in experiments in Section 4: while point-wisely $L_N f(x)$ can be oscillating and deviating from the $\mathcal{L}f(x)$, where L_N is the graph Laplacian and $\mathcal{L}$ is the limiting differential operator, the Dirichlet form has much smaller error especially when ϵ is small (Fig. 3 and Fig. 4).

3. Analysis of graph Laplacian

In this section, we analyze the convergence of self-tuned graph Laplacian computed from dataset $X = \{x_i\}_{i=1}^{N_x}$, $x_i \sim p$, sampled on $\mathcal{M}$, where $\hat{\rho}(x_i)$ has been computed from a stand-alone dataset Y,

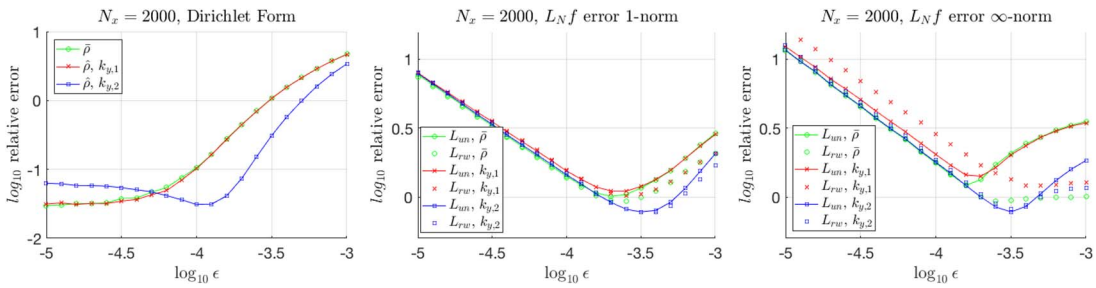

FIG. 3. Relative error of (left) Dirichlet form computed using L_{un}, and (middle) the Err_1 error in (4.1) of $L_N f$ computed by various L_N plotted v.s. a range of values of ϵ, $N_y = 4000$, $k_y = 32, 256$, averaged over 500 runs. (Right) Same plot for the Err_∞ error.

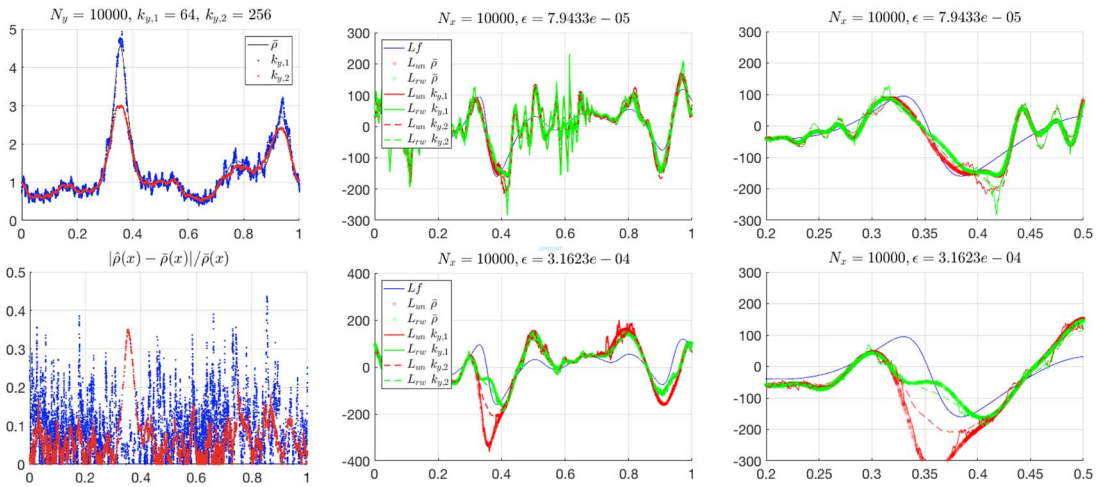

FIG. 4. $L_N f$ by estimated $\hat{\rho}$ compared with true $\Delta_p f$, denoted as Lf (blue curve), with two values of k_y. The data are as in Fig. 2. (Left) kNN-estimated $\hat{\rho}$ and relative errors. (Right upper) Estimated $L_N f$ where L_N equals L_{un} and $L_{rw'}$, and using kNN-estimated $\hat{\rho}$, compared with using population $\bar{\rho}$. The right plot is the zoom in of the left plot on interval $[0.2, 0.5]$. (Right bottom) Same plot as the right upper panel at another value of ϵ. See more explanation in Section 4.2.

and we assume that Theorem 2.3 holds. We first introduce the notations of limiting operators and the Dirichlet forms in Section 3.1 and then prove

- The convergence of the kernelized Dirichlet form in Section 3.2, as a middle-step result;
- The convergence of the graph Dirichlet form in Section 3.3;
- The convergence of $L_N f(x)$ for un-normalized and random-walk graph Laplacian operators L_N in Section 3.4.

We simplify notation $N = N_x$ in the section. All proofs are in Appendix. The following regularity and decay condition is needed for the function k_0 in (1.4). The condition on k_0 as in [**11**] is introduced in Assumption A.3, and here we further assume non-negativity and C^4 regularity of k_0 for simplicity.

ASSUMPTION 3.1 (Assumption on k_0). k_0 satisfies Assumption A.3 and, in addition,

(C1) Regularity. k_0 is continuous on $[0,\infty)$ and C^4 on $(0,\infty)$.

(C2) Decay condition. $\exists a, a_l > 0$, s.t., $|k_0^{(l)}(\xi)| \le a_l e^{-a\xi}$ for all $\xi > 0$, $l = 0, 1, \cdots, 4$.

(C3) Non-negativity. $k_0 \geqslant 0$ on $[0,\infty)$.

We use $m_0 = m_0[k_0]$ and $m_2 = m_2[k_0]$ if the kernel function dependence is not clarified.

3.1 *Notation of manifold Laplacian operators and Dirichlet forms*

Recall the weighted Laplacian Δ_p defined as in (1.2) on $(\mathcal{M}, pdV)$, where p is the density of a positive measure on $\mathcal{M}$. Below, we write $\Delta_{\mathcal{M}}$ as Δ, $\nabla_{\mathcal{M}}$ as ∇, when there is no danger of confusion.

Take a positive C^1 function ρ on $\mathcal{M}$. As will appear in the analysis, we introduce $\mathcal{L}_\rho^{(\alpha)}$ as

$$\mathcal{L}_\rho^{(\alpha)} := \Delta + 2\frac{\nabla p}{p}\cdot\nabla + (d - 2\alpha + 2)\frac{\nabla\rho}{\rho}\cdot\nabla. \tag{3.1}$$

When $\rho = \bar{\rho} = p^{-1/d}$, one can verify that $\mathcal{L}^{(\alpha)} := \mathcal{L}_{\bar{\rho}}^{(\alpha)}$ satisfies

$$\mathcal{L}^{(\alpha)} = \Delta + \left(1 + \frac{2(\alpha-1)}{d}\right)\frac{\nabla p}{p}\cdot\nabla. \tag{3.2}$$

We will show that the operators $\mathcal{L}^{(\alpha)}$ and $p^{\frac{2(\alpha-1)}{d}}\mathcal{L}^{(\alpha)}$ are the limiting operators of the (modified) random-walk and un-normalized graph Laplacians, respectively.

The *differential Dirichlet form* associated with Δ_p is defined as

$$\mathcal{E}_p(f,f) := -\langle f, \Delta_p f\rangle_p = \int_{\mathcal{M}} p|\nabla f|^2\,\mathrm{d}V,$$

where $\langle f, g\rangle := \int_{\mathcal{M}} fg dV$ for $f, g \in L^\infty(\mathcal{M})$, and $\langle f, g\rangle_q = \int_{\mathcal{M}} fgq dV$ for q a density of a positive measure on $\mathcal{M}$. In below, we may omit $\mathrm{d}V$ in the notation of integral over $\mathcal{M}$, that is, $\int_{\mathcal{M}} f$ means $\int_{\mathcal{M}} f\,\mathrm{d}V$. Given a graph affinity matrix W and a vector $f : V \to \mathbb{R}, f \in \mathbb{R}^N$, we consider the (normalized) *graph Dirichlet form* defined as

$$\begin{aligned} E_N(f,f) &:= \frac{2}{\epsilon m_2}\frac{1}{N^2}\epsilon^{-d/2} f^T(D - W)f \\ &= \frac{2}{\epsilon m_2}\frac{1}{N^2}\sum_{i,j=1}^N \epsilon^{-d/2} W_{ij} f_i(f_i - f_j) = \frac{1}{\epsilon m_2}\frac{1}{N^2}\sum_{i,j=1}^N \epsilon^{-d/2} W_{ij}(f_i - f_j)^2. \end{aligned} \tag{3.3}$$

We will prove that the graph Dirichlet form converges to the differential Dirichlet form of a density $p_\alpha := p^{1+\frac{2(\alpha-1)}{d}}$ on $\mathcal{M}$. This is consistent with the above limiting operator, as one can verify that

$$-\langle f, p^{\frac{2(\alpha-1)}{d}}\mathcal{L}^{(\alpha)} f\rangle_p = -\langle f, \Delta_{p_\alpha} f\rangle_{p_\alpha} = \mathcal{E}_{p_\alpha}(f,f).$$

3.2 *Convergence of the kernelized Dirichlet form*

Consider $W = W^{(\alpha)}$ as in (1.4), and define

$$\hat{K}(x,y) := \epsilon^{-d/2} k_0\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{1}{\hat{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}. \tag{3.4}$$

Then, by definition, $\epsilon^{-d/2}W_{ij}^{(\alpha)} = \hat{K}(x_i,x_j)$. We use 'hat' to emphasize the dependence on the estimated bandwidth $\hat{\rho}$. When $f_i = f(x_i)$ for $f : \mathcal{M} \to \mathbb{R}$ (here we use the notation f for both the function and the vector), $E_N(f,f)$ has the following population counterpart which is an integral form on $\mathcal{M}$,

$$\mathcal{E}^{(\alpha)}(f,f) := \frac{1}{\epsilon m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2\hat{K}(x,y)p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y). \tag{3.5}$$

We call $\mathcal{E}^{(\alpha)}(f,f)$ the *kernelized Dirichlet form*. The following proposition proves the convergence of $\mathcal{E}^{(\alpha)}(f,f)$ to the differential Dirichlet form $\mathcal{E}_{p_\alpha}(f,f)$:

PROPOSITION 3.2 Suppose $\hat{\rho}$ satisfies that $\sup_{x\in\mathcal{M}}\frac{|\hat{\rho}(x)-\bar{\rho}(x)|}{|\bar{\rho}(x)|} < \varepsilon_\rho < 0.1$. Then for any $f \in C^\infty(\mathcal{M})$,

$$\mathcal{E}^{(\alpha)}(f,f) = \mathcal{E}_{p_\alpha}(f,f)(1+O^{[\alpha]}(\varepsilon_\rho)) + O^{[f,p]}(\epsilon), \quad p_\alpha = p^{1+2(\alpha-1)/d}.$$

In the proposition, we omit the dependence on α in the notation of the $O^{[f,p]}(\epsilon)$ term. Here and in below, we omit the dependence on α and track that on p and f in the big-O notation, unless we want to stress the former. The proposition leads to the convergence of $\mathbb{E}E_N(f,f)$ and is used in proving the convergence of $E_N(f,f)$ (Theorem 3.3) and the weak convergence of L_Nf (Theorem 3.7). An important technical object used in the analysis of $\mathcal{E}^{(\alpha)}(f,f)$ and later analysis is the following integral operator $G_\epsilon^{(\rho)}$ defined for $f \in C^\infty(\mathcal{M})$ and any $\epsilon > 0$,

$$G_\epsilon^{(\rho)}f(x) := \epsilon^{-d/2}\int_{\mathcal{M}} k_0\left(\frac{\|x-y\|^2}{\epsilon\rho(y)}\right)f(y)\,\mathrm{d}V(y), \tag{3.6}$$

which is well defined when ρ is positive and has some regularity so that the integral exists, e.g., C^0 regularity and bounded from below. The following lemma is a reproduce of a similar step used in [5] where we derive point-wise error bound (see remark A.1).

LEMMA 3.1 Under Assumption 2.1, suppose k_0 satisfies Assumption 3.1, f and ρ are in $C^4(\mathcal{M})$, and $0 < \rho_{min} < \rho < \rho_{max}$ uniformly on $\mathcal{M}$, then

$$\begin{aligned} G_\epsilon^{(\rho)}f &= m_0 f\rho^{\frac{d}{2}} + \epsilon\frac{m_2}{2}(\omega f\rho^{1+\frac{d}{2}} + \Delta(f\rho^{1+\frac{d}{2}})) + r_\epsilon^{(2)}, \\ \sup_{x\in\mathcal{M}}|r_\epsilon^{(2)}(x)| &\leqslant c_\rho(1+\sum_{l=0}^4\|D^{(l)}f\|_\infty)(1+\sum_{l=0}^4\|D^{(l)}\rho^{-1}\|_\infty)\epsilon^2 = O^{[f,\rho]}(\epsilon^2), \end{aligned} \tag{3.7}$$

where $c_\rho > 0$ is a constant depending on $(\mathcal{M}, k_0, \rho_{max}, \rho_{min})$ (a rational function of (ρ_{max}, ρ_{min}) where the coefficients depend on $(\mathcal{M}, k_0)$), $D^{(l)}$ is l-th manifold intrinsic derivatives, and $\omega(x)$ depends on local derivatives of the extrinsic manifold coordinates at x.

However, we cannot directly apply the lemma to (3.5) because $\hat{\rho}$ does not have C^4 regularity. The proof of Proposition 3.2 is via substituting $\hat{\rho}$ by $\bar{\rho}$ and control the error, and then applying Lemma 3.1 where $\rho = \bar{\rho}$ which is C^∞. As a postponed discussion, modifying $\hat{\rho}$ to be C^∞ is considered in Appendix A.4 under another limiting setting.

3.3 *Convergence of the graph Dirichlet form*

We will show that when $N \to \infty$ and $\epsilon \to 0$ under a proper joint limit, $E_N(f,f)$ converges to $\mathcal{E}_{p_\alpha}(f,f)$. This means that with the self-tuned kernel the graph Dirichlet form asymptotically recovers the differential Dirichlet form of the weighted Laplacian Δ_{p_α} on $(\mathcal{M}, p_\alpha dV)$. In particular,

- When $\alpha = 1$, $p_1 = p$, and the graph Laplacian recovers Δ_p on $(\mathcal{M}, pdV)$.
- When $\alpha = 0$, $p_0 = p^{1-2/d}$, thus the original self-tune graph Laplacian recovers weighted Laplacian with a modified density.
- When $\alpha = 1 - \frac{d}{2}$, p_α is a constant, then the graph Laplacian recovers $\Delta_{\mathcal{M}}$ and the Dirichlet form with uniform density. We provide an approach to obtain $\Delta_{\mathcal{M}}$ when d is not known in Section 4.4.

For the estimated $\hat{\rho}$ from Y, suppose Theorem 2.3 holds, we thus consider the randomness over X conditioning on a realization of Y under the good event, which already guarantees the uniform smallness of $|\hat{\rho}(x) - \bar{\rho}(x)|/\bar{\rho}(x)$.

THEOREM 3.3 Suppose $\hat{\rho}$ satisfies that $\sup_{x\in\mathcal{M}} \frac{|\hat{\rho}(x)-\bar{\rho}(x)|}{|\bar{\rho}(x)|} < \varepsilon_\rho < 0.1$, and as $N \to \infty$,

$$\epsilon = o(1), \quad \epsilon^{d/2} N = \Omega(\log N), \quad \varepsilon_\rho = o(1),$$

then for any $f \in C^\infty(\mathcal{M})$, when N is sufficiently large, w.p. $> 1 - 2N^{-10}$, and $p_\alpha = p^{1+2(\alpha-1)/d}$,

$$E_N(f,f) = \mathcal{E}_{p_\alpha}(f,f)(1 + O^{[\alpha]}(\varepsilon_\rho)) + O^{[f,p]}(\epsilon) + O^{[1]}\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}} \int_{\mathcal{M}} |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}}}\right).$$

REMARK 3.1 By Remark 2.1, the optimal choice of k to minimize ε_ρ is when $k \sim N_y^{1/(1+d/4)}$ and this leads to $\varepsilon_\rho \sim N_y^{-1/(2+d/2)}$ up to a $\log N$ factor. The possible $\log N$ factor is no longer declared in all the scalings in this remark. To make $\epsilon \sim \varepsilon_\rho$, it gives $\epsilon \sim (\frac{k}{N_y})^{2/d} \sim N_y^{-1/(2+d/2)}$. The scaling $\epsilon \sim (\frac{k}{N_y})^{2/d}$ is the same one as in the original kNN self-tune kernel (1.3). Meanwhile, in the error bound in Theorem 3.3, leaving the term due to ε_ρ aside, the other two terms of bias and variance errors are balanced when $\epsilon \sim N_x^{-1/(2+d/2)}$, and this gives the overall error of the two terms as $N_x^{-1/(2+d/2)}$. Compared to $\varepsilon_\rho \sim N_y^{-1/(2+d/2)}$ at the optimal scaling of k with N_y, the overall error bound in Theorem 3.3 is balanced when $N_y = \Theta(N_x)$.

To see the effect of self-tuning kernel, we compare Theorem 3.3 with the following theorem for a fixed-bandwidth kernel normalized by density estimators, defined for $\beta \leqslant 1$ as

$$W_{ij}^{(\beta)} = k_0\left(\frac{\|x_i - x_j\|^2}{\epsilon}\right)\frac{1}{\hat{p}(x_i)^\beta \hat{p}(x_j)^\beta}, \tag{3.8}$$

assuming $\hat{p}(x_i) > 0$ for all i. Let $E_{N,\epsilon}(f,f)$ equals (3.3) with $W = W^{(\beta)}$, and that gives

$$E_{N,\epsilon}(f,f) := \frac{1}{\epsilon m_2}\frac{1}{N^2}\sum_{i,j=1}^{N} \epsilon^{-d/2} W_{ij}^{(\beta)}(f_i - f_j)^2 .$$

Below, in Theorems 3.4 and 3.8 about fixed-bandwidth kernel, we track the constant dependence on β more carefully since for the special case where $\beta = 0$ no density estimation is needed.

THEOREM 3.4 Suppose as $N \to \infty$, $\epsilon = o(1)$, $\epsilon^{d/2}N = \Omega(\log N)$, and if $\beta \neq 0$, the estimated density $\hat{p}$ satisfies that $\sup_{x\in\mathcal{M}} \frac{|\hat{p}(x)-p(x)|}{p(x)} < \varepsilon_p < 0.1$, and $\varepsilon_p = o(1)$, then for any $f \in C^\infty(\mathcal{M})$, when N is sufficiently large, w.p. $> 1 - 2N^{-10}$, and $c_\beta = \max\{1.1^{-\beta-1}, 0.9^{-\beta-1}\}$,

$$E_{N,\epsilon}(f,f) = \mathcal{E}_{p^{2-2\beta}}(f,f)(1 + O^{[1]}(\beta c_\beta \varepsilon_p)) + O^{[f,p,\beta]}(\epsilon) + O^{[1]}\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}}\int_{\mathcal{M}} |\nabla f|^4 p^{2-4\beta}}\right).$$

In particular, when $\beta = 0$, $E_{N,\epsilon}(f,f) = \mathcal{E}_{p^2}(f,f) + O^{[f,p]}(\epsilon) + O^{[1]}\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}}\int_{\mathcal{M}} |\nabla f|^4 p^2}\right)$.

Note that $\Delta_{p^{2-2\beta}} = \Delta + 2(1-\beta)\frac{\nabla p}{p}\cdot\nabla$, which is consistent with the limiting operator of the original Diffusion Map paper [11], and in particular, $\beta = \frac{1}{2}$ recovers Δ_p. Strictly speaking, the setting is different because in [11], $D_{ii}^{1/2}$ is used to normalize the affinity matrix $W_{ij} = k_0\left(\frac{\|x_i-x_j\|^2}{\epsilon}\right)$, and $D_{ii} = \sum_j k_0\left(\frac{\|x_i-x_j\|^2}{\epsilon}\right)$. While D_{ii} can be viewed as a KDE, normalizing by D_{ii} introduces dependence and techniques to analyze normalized graph Lapalcian are needed, e.g., as in Theorem 3.5.

REMARK 3.2 The error bound in Theorem 3.4 has an $O(\varepsilon_p)$ term (when $\beta \neq 0$) which appears to be the counterpart of the $O(\varepsilon_\rho)$ term in the bound in Theorem 3.3. We have shown in Remark 2.2 that the relative error of $\hat{p}$ by a fixed-bandwidth KDE (2.3) behaves differently from that of $\hat{\rho}$. Specifically, when variance error dominates, $|\hat{p}(x) - p(x)|/p(x)$ is proportional to $p(x)^{-1/2}$, while the variance error in $|\hat{\rho}(x) - \bar{\rho}(x)|/\bar{\rho}(x)$ can be made small uniformly for $x \in \mathcal{M}$ independent of $p(x)$. This means that, when p is small at some places, to prevent the variance error in $|\hat{p}(x) - p(x)|/p(x)$ to be too large, the choice of the fixed-bandwidth parameter ϵ may be restricted by the smallness of p. In such cases, the kNN self-tuned kernel can have better robustness due to that the kNN estimator $\hat{\rho}$ adopts a different bias-variance error trade-off which uniformly controls the variance error term.

We postpone further discussion about fixed bandwidth kernel, since the current paper focuses on the estimated variable bandwidth kernel.

3.4 *Convergence of $L_N f$*

We consider two types of graph Laplacian operator $L_N f$, where, using kernel $K^{(\alpha)}_{\epsilon,\hat{\rho}}$ as in (1.4), the unnormalized graph Laplacian operator applied to $f \in C^\infty(\mathcal{M})$ is defined as

$$L^{(\alpha)}_{un} f(x) = \frac{2\epsilon^{-\frac{d}{2}-1}}{m_2} \frac{1}{\hat{\rho}(x)^\alpha} \frac{1}{N} \sum_{j=1}^{N} k_0\left(\frac{\|x-x_j\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(x_j)}\right) \frac{f(x_j)-f(x)}{\hat{\rho}(x_j)^\alpha}, \tag{3.9}$$

and the (modified) random-walk graph Laplacian operator is

$$L^{(\alpha)}_{rw'} f(x) = \frac{1}{\epsilon \frac{m_2}{2m_0}\hat{\rho}(x)^2} \left(\frac{\sum_{j=1}^{N} k_0\left(\frac{\|x-x_j\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(x_j)}\right) \frac{f(x_j)}{\hat{\rho}(x_j)^\alpha}}{\sum_{j=1}^{N} k_0\left(\frac{\|x-x_j\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(x_j)}\right) \frac{1}{\hat{\rho}(x_j)^\alpha}} - f(x) \right). \tag{3.10}$$

In the matrix form, the operator differs from the usual random-walk Laplacian ($I - D^{-1}W$) by multiplying another diagonal matrix $D_{\hat{\rho}}^{-2}$ (up to multiplying a constant and the sign); thus, we call it 'modified' and denote it by 'rw-prime'.

The point-wise convergence of $L_N f(x)$ at a fixed point $x \in \mathcal{M}$ is a more traditional setting under which the convergence to a limiting diffusion operator has been considered in several papers [**5**,**11**,**42**]. The closest one is the result in [**5**], and an extension of the method therein leads to a convergence to $\mathcal{L}^{(\alpha)}_{\hat{\rho}} f$ under the asymptotic that $\epsilon = o((k/N_y)^{4/d})$ (cf. Theorems A.5 and A.6 in Appendix A.4). However, this convergence result does not imply consistency to $\mathcal{L}^{(\alpha)}_{\hat{\rho}} f = \mathcal{L}^{(\alpha)} f$, due to the lack of convergence of $\frac{\nabla \hat{\rho}}{\hat{\rho}}$ to $\frac{\nabla \bar{\rho}}{\bar{\rho}}$, as discussed in Section 2.3. Meanwhile, note that the uniform C^0 consistency of $\hat{\rho}$ to $\bar{\rho}$ does imply the weak convergence of $\mathcal{L}^{(\alpha)}_{\hat{\rho}} f \to \mathcal{L}^{(\alpha)}_{\bar{\rho}} f$ when $\varepsilon_\rho \to 0$, a result of the same type as Theorem 3.7, while the latter shows an improved variance error (with $\epsilon^{-1/2}$ rather than $\epsilon^{-d/4-1/2}$).

Back to the point-wise convergence of $L_N f(x)$. To be able to establish the consistency to $\mathcal{L}^{(\alpha)} f$, we instead consider another limiting regime of ϵ, namely $\epsilon = \Omega(\varepsilon_\rho)$, which is $\Omega((k/N_y)^{2/d})$ up to a factor of $\sqrt{\log N}$ under the optimal scaling of k as in Remark 2.1, and we take a different approach. The following lemma shows that substituting $\bar{\rho}$ with $\hat{\rho}$ in $G^{(\rho)}_\epsilon f(x)$ incurs an extra error of $O(\varepsilon_\rho)$ point-wisely.

Lemma 3.2 Under the same condition of Lemma 3.1, in particular, f and ρ are in $C^4(\mathcal{M})$. Suppose a positive integrable $\tilde{\rho} : \mathcal{M} \to \mathbb{R}^+$ satisfies that $\sup_{x\in\mathcal{M}} \frac{|\tilde{\rho}(x)-\rho(x)|}{\rho(x)} < \varepsilon < 0.1$, then when the ϵ in $G^{(\cdot)}_\epsilon$ is sufficiently small,

$$G^{(\tilde{\rho})}_\epsilon f = G^{(\rho)}_\epsilon f + \tilde{r}, \quad \sup_{x\in\mathcal{M}} |\tilde{r}(x)| \leqslant c'_\rho \|f\|_\infty \varepsilon = O^{[f,\rho]}(\varepsilon), \tag{3.11}$$

where $c'_\rho > 0$ is a constant depending on $(\mathcal{M}, k_0, \rho_{max}, \rho_{min})$.

With the lemma, the following two theorems prove the point-wise convergence to the limiting operators of the two graph Laplacians operators, assuming that $\varepsilon_\rho = o(\epsilon)$.

THEOREM 3.5 Suppose $\hat{\rho}$ satisfies that $\sup_{x\in\mathcal{M}} \frac{|\hat{\rho}(x)-\bar{\rho}(x)|}{|\bar{\rho}(x)|} < \varepsilon_\rho < 0.1$, and as $N \to \infty$,

$$\epsilon = o(1), \quad \epsilon^{d/2+1}N = \Omega(\log N), \quad \varepsilon_\rho = o(\epsilon),$$

then for any $f \in C^\infty(\mathcal{M})$, when N is sufficiently large and the threshold is determined by $(\mathcal{M}, f, p)$ and uniform for all x, w.p. higher than $1 - 4N^{-10}$,

$$L_{rw}^{(\alpha)} f(x) = \mathcal{L}^{(\alpha)} f(x) + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right) + O^{[1]}\left(\|\nabla f\|_\infty p(x)^{1/d}\sqrt{\frac{\log N}{N\epsilon^{d/2+1}}}\right), \tag{3.12}$$

where the constants in big-O are uniform for all $x \in \mathcal{M}$.

REMARK 3.3 As shown in the proof of Theorem 3.5, at x where $\nabla f(x) \neq 0$, the variance error can be bounded by $O^{[1]}\left(|\nabla f(x)| p(x)^{1/d}\sqrt{\frac{\log N}{N\epsilon^{d/2+1}}}\right)$ with the same high probability and the threshold of large N possibly depends on x. One can also verify $O^{[1]}\left((|\nabla f(x)| + 0.1) p(x)^{1/d}\sqrt{\frac{\log N}{N\epsilon^{d/2+1}}}\right)$ as the variance error bound for large N with x-uniform threshold. The addition of 0.1 is to make the factor $(|\nabla f(x)| + 0.1)$ uniformly bounded from below and prevent the bound to vanish when $\nabla f(x) = 0$, and 0.1 can be any other positive constant. If the behavior at a point x is of interest, theoretically the variance error can be improved in rate at x where $\nabla f(x)$ vanishes [42]. As we mainly track the influence of $p(x)$ which may be small at some x, we adopt the $\|\nabla f\|_\infty$ factor in the theorem for simplicity. The same applies to the point-wise convergence results in Theorems 3.6 and 3.8.

THEOREM 3.6 With notation and condition same as those in Theorem 3.5, when N is sufficiently large and the threshold is determined by $(\mathcal{M}, f, p)$ and uniform for all x, w.p. higher than $1 - 2N^{-10}$,

$$L_{un}^{(\alpha)} f(x) = p^{\frac{2(\alpha-1)}{d}} \mathcal{L}^{(\alpha)} f(x) + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right) + O^{[1]}\left(\|\nabla f\|_\infty p(x)^{\frac{2\alpha-1}{d}}\sqrt{\frac{\log N}{N\epsilon^{d/2+1}}}\right).$$

In Theorems 3.5 and 3.6, the error bound has an additional term of $O(\frac{\varepsilon_\rho}{\epsilon})$ compared with that in [5,42]. The technical reason is that we use Lemma 3.2 to substituting $\hat{\rho}$ with $\bar{\rho}$, which gives $O(\varepsilon_\rho)$ error at the '$O(1)$' level but not at the '$O(\epsilon)$' level. In the proof of Theorem 3.3, the $O(\varepsilon_\rho)$ substituting error takes place at the '$O(\epsilon)$' level thanks to the quadratic form.

For the un-normalized graph Laplacian operator, the additional $O(\frac{\varepsilon_\rho}{\epsilon})$ error can be removed if we consider the weak convergence, which can be of interest in certain settings.

THEOREM 3.7 Suppose $\hat{\rho}$ satisfies that $\sup_{x\in\mathcal{M}} \frac{|\hat{\rho}(x)-\bar{\rho}(x)|}{|\bar{\rho}(x)|} < \varepsilon_\rho < 0.1$, and as $N \to \infty$,

$$\epsilon = o(1), \quad \epsilon N = \Omega(\log N), \quad \varepsilon_\rho = o(1),$$

then for any $\varphi \in C^\infty(\mathcal{M})$, when N is large, w.p. $> 1 - 2N^{-10}$, $p_\alpha := p^{1+\frac{2(\alpha-1)}{d}}$,

$$\langle \varphi, L_{un}^{(\alpha)} f \rangle_p = \langle \varphi, \Delta_{p_\alpha} f \rangle_{p_\alpha} + O^{[p,f,\varphi]}\left(\epsilon, \varepsilon_\rho\right) + O^{[1]}\left(\|\varphi\|_\infty \|p^{\alpha/d}\|_\infty \sqrt{\frac{\log N}{N\epsilon} \int p_\alpha |\nabla f|^2} \right).$$

Note that the above weak convergence result is only possible for the un-normalized operator because technically the $D^{-1}W$ normalization in the random-walk operator breaks the linearity.

At last, we compare with the graph Laplacian operator L_N defined by fixed bandwidth kernel matrix (3.8), namely

$$L_{\epsilon,rw}^{(\beta)} f(x) = \frac{1}{\epsilon \frac{m_2}{2m_0}} \left(\frac{\sum_{j=1}^N k_0\left(\frac{\|x-x_j\|^2}{\epsilon}\right) \frac{f(x_j)}{\hat{p}(x_j)^\beta}}{\sum_{j=1}^N k_0\left(\frac{\|x-x_j\|^2}{\epsilon}\right) \frac{1}{\hat{p}(x_j)^\beta}} - f(x) \right).$$

The counterpart of Theorem 3.5 is the following:

THEOREM 3.8 Suppose as $N \to \infty$, $\epsilon = o(1)$, $\epsilon^{d/2+1} N = \Omega(\log N)$, and if $\beta \neq 0$, the estimated density $\hat{p}$ satisfies that $\sup_{x\in\mathcal{M}} \frac{|\hat{p}(x)-p(x)|}{p(x)} < \varepsilon_p < 0.1$, and $\varepsilon_p = o(\epsilon)$. Then for any $f \in C^\infty(\mathcal{M})$, when N is large and the threshold is determined by $(\mathcal{M}, f, p, \beta)$ and uniform for all x,

$$L_{\epsilon,rw}^{(\beta)} f(x) = \Delta_{p^{2-2\beta}} f(x) + O^{[f,p,\beta]}\left(\epsilon, \beta\frac{\varepsilon_p}{\epsilon}\right) + O^{[1]}\left(\|\nabla f\|_\infty p(x)^{-1/2} \sqrt{\frac{\log N}{N\epsilon^{d/2+1}}} \right). \tag{3.13}$$

In particular, when $\beta = 0$, the bias error term is reduced to $O^{[f,p]}(\epsilon)$.

The counterpart for un-normalized graph Laplacian can be derived similarly and omitted. The limiting operator is the same as in Theorem 3.4 and consistent with the result in [**11**]. Compared with Theorem 3.5, apart from the needed condition on the relative error of $\hat{p}$ (cf. Remarks 2.2 and 3.2), the variance error term in (3.13) has a factor of $p(x)^{-1/2}$, while for self-tuned kernel $W^{(\alpha)}$ the factor is $p(x)^{1/d}$ as in (3.12). This can be expected because the self-tuned bandwidth is designed to overcome the difficulty of low data density by enlarging the kernel bandwidth at those places, and our analysis reveals the effect by the reduced variance error at x where $p(x)$ is small. Such advantage is supported by experiments on the hand-written digit image dataset in Section 4.5.

4. Numerical experiments

This section presents numerical experiments to verify the theory. Codes available at the public repository https://github.com/xycheng/selftune-kNN. In this section, we denote the parameter k as k_y when the kNN estimation is conducted on the dataset Y. In Subsection 4.3, we use the notation k_x when computing the kNN estimation on X.

4.1 *kNN estimator $\hat{\rho}$ of $\bar{\rho} = p^{-1/d}$*

We numerically examine the kNN estimation of $\bar{\rho}$, namely $\hat{\rho}$ as defined in (2.1), and compare it with the fixed bandwidth KDE estimator $\hat{p}$ as in (2.3), where $h_{kde}(r) := e^{-r/(4/\pi)}$. The dataset is sampled from

a circle of length 1 isometrically embedded in $\mathbb{R}^2$ i.i.d. according to a density function p, which equals 0.05 at $t_X = 0.25$, $0 \leqslant t_X \leqslant 1$ being the intrinsic coordinate (arclength), as shown in Fig. 1. The plots on the right-hand side show the difference of the relative error at place where p is low. As k_y decreases (ϵ increases), the variance error starts to dominate, and $\hat{\rho}$ gives the relative error uniformly small across locations, as predicted by Theorem 2.3. In contrast, $\hat{p}$ gives a larger relative error near $t_X = 0.25$. The result empirically verifies Theorem 2.3 and Remark 2.2.

4.2 *Estimation of Dirichlet form and $L_N f(x)$*

On the simulated data lying on a one-dimensional smooth manifold embedded in $\mathbb{R}^4$ with a non-uniform density p (Fig. 2), we compute self-tuned graph Laplacians on $N_x = 2000$ data samples using (1) L_{un} (3.9) and (2) $L_{rw'}$ as in (3.10), $\alpha = 1$, where $\hat{\rho}$ is estimated from a stand-alone dataset Y with $N_y = 4000$, and $k_y = 32, 256$, respectively. To evaluate the influence of $\bar{\rho}$ estimation, we also compute L_{un} and $L_{rw'}$ where $\hat{\rho}$ is replaced to be the true $\bar{\rho}$. The relative errors of

- The Dirichlet form $\langle f, \Delta_p f \rangle_p$,
- The point-wise error measure by

$$\mathrm{Err}_1 := \sum_{i=1}^{N_x} |L_N f(x_i) - \Delta_p f(x_i)|, \quad \mathrm{Err}_\infty := \max_{1 \leqslant i \leqslant N_x} |L_N f(x_i) - \Delta_p f(x_i)|, \tag{4.1}$$

are given in Fig. 3. The error of $L_N f$ shows a scale of about $\epsilon^{-d/4-1/2} = \epsilon^{-0.75}$ in Fig. 3 right two plots, when the variance error dominates due to the small value of ϵ. In comparison, the accuracy of Dirichlet form is less sensitive to the small value of ϵ, as shown in Fig. 3(left), which is consistent with the theoretical result in Theorems 3.3 and 3.5.

Note that the relative 1-norm error shown in the plot divides Err_1 by $\|\{\Delta_p f(x_i)\}_{i=1}^{N_x}\|_1$; thus, its magnitude (about or greater than 1 in Fig. 3) depends on the choice of the test function f. Same with the relative ∞-norm error. When N_x is increased to be 10,000, with the same f, the smallest relative error across ϵ is about 0.5 (averaged over 20 runs).

Taking $N_x = 10,000$, $N_y = 10,000$, with $k_y = 64, 256$, respectively, we visualize in Fig. 4 snapshots of single realizations of $L_N f$. With smaller value of ϵ, the estimated $L_N f$ has more oscillation around the true value, and when ϵ is larger, the oscillation is less but the function $L_N f$ is significantly biased at certain places on the manifold. Note that when k_y is larger, the estimated $\hat{\rho}$ is smoother but has a significant bias at places where p is small, and such bias is also reflected in the estimated $L_N f$. The two Laplacians, L_{un} and $L_{rw'}$, give comparable results.

4.3 *The influence of stand-alone Y*

We compare with the results using X to estimate $\hat{\rho}$. The dataset is the same as that in Fig. 2. $N_x = 2000$ and $k_x = 32$ are used to compute $\hat{\rho}_X$. Take $N_y = \{2000, 4000, \cdots, 32000\}$ and $k_y = \{37, 64, \cdots, 338\}$, where k_y is chosen to scale as $N_y^{4/5}$, according to Theorem 2.3 ($d = 1$). The result for one realization with the largest N_y is in Fig. 5, where using a stand-alone Y of a much larger size than X reduces the error in the estimated $\hat{\rho}$ as well as the oscillation in the estimated $L_N f$ (plots for $L_{rw'} f$ are similar and not shown). The relative errors of Dirichlet form and Err_∞ of $L_N f$ across ϵ are shown in Fig. 6, where using $\hat{\rho}_X$ and $\hat{\rho}_Y$ give comparable accuracy. Moreover, the result with $\hat{\rho}_Y$ approaches $L_N f$ computed from $\bar{\rho}$ as

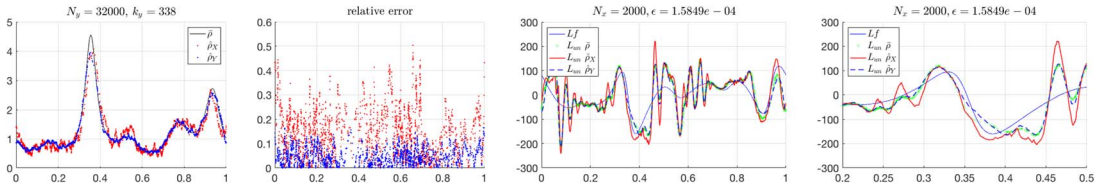

FIG. 5. Same plots as Fig. 4 for self-tune Laplacian computed with $\hat{\rho}_X$. (Left two) kNN-estimated $\hat{\rho}$ and relative errors computed from X and Y. (Right two) Estimated L_{unf}, with the zoom on the interval $[0.2, 0.5]$.

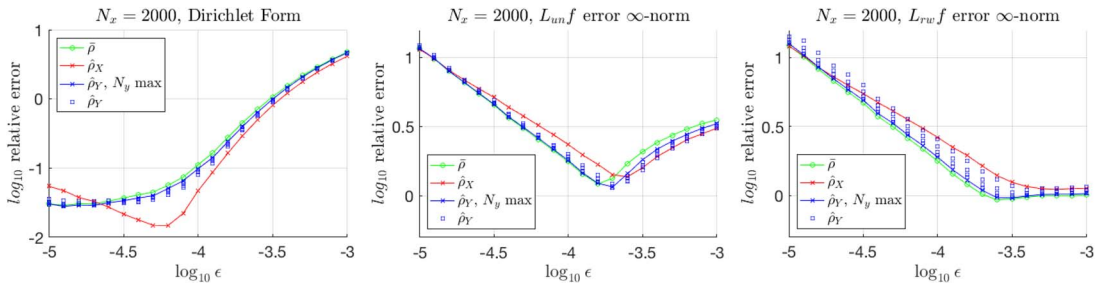

FIG. 6. Relative error of (Left) Dirichlet form computed using L_{un}, and (Middle) the Err_∞ error of L_{un} computed using $\bar{\rho}$, $\hat{\rho}_X$ and $\hat{\rho}_Y$ over a range of ϵ and different $\{N_y, k_y\}$ (blue squares), averaged over 500 runs. (Right) Same plot for $L_{rw'}$.

N_y and k_y increase. This suggests that when significantly more data samples than N_x are available, using the rest as Y to estimate the bandwidth function $\hat{\rho}$ may improve the estimation of the self-tuned graph Laplacian. With limited number of data samples, splitting stand-alone Y may worsen the performance (due to decreasing N_x) than using the whole dataset as X and estimating the bandwidth on itself.

4.4 *Recovery of the Laplace–Beltrami operator*

According to the theory, the limiting operator is $\Delta_{\mathcal{M}}$ when $\alpha = 1 - \frac{d}{2}$. Here we examine two ways to recover $\Delta_{\mathcal{M}}$:

(1) By the self-tuned kernel affinity $W^{(1-\frac{d}{2})}$.

(2) By a normalization with a combination of $\hat{p}$ and $\hat{\rho}$,

$$W_{ij} = \frac{k_0\left(\frac{\|x_i - x_j\|^2}{\epsilon \hat{\rho}(x_i)\hat{\rho}(x_j)}\right)}{(\hat{\rho}\hat{p}^{1/2})(x_i)(\hat{\rho}\hat{p}^{1/2})(x_j)} = \frac{W^{(1)}_{ij}}{\hat{p}^{1/2}(x_i)\hat{p}^{1/2}(x_j)}, \tag{4.2}$$

because $\bar{\rho}^{-d/2} = p^{1/2}$. The second approach does not need prior knowledge or estimation of the intrinsic dimensionality d and thus can be applied in more general scenarios.

Consider the same one-dimensional manifold data used in Fig. 2. Take $N_x = 1000$, $\epsilon = 10^{-4}$, and compute $\hat{\rho}$ from X. The embeddings by the first four (non-trivial) eigenvectors of various graph Laplacians are shown in Fig. 7, where the last column shows the result produced by affinity matrix $W_{ij} = \frac{K_{ij}}{d_i d_j}$, where K_{ij} is the fix-bandwidth kernel affinity $K_{ij} = k_0(\frac{\|x_i - x_j\|^2}{\epsilon})$ and $d_i = \sum_j K_{ij}$, as in [11]. Recall that the eigenfunctions of the Laplace–Beltrami operator are sine and cosine functions with different frequencies. In this example, the random-walk graph Laplacian produces a visually better

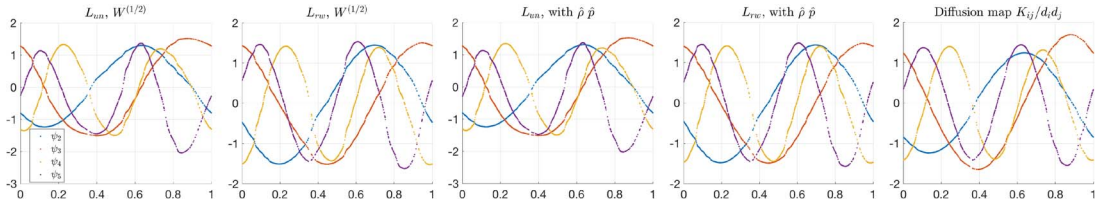

FIG. 7. Plots of the first 4 (non-trivial) eigenvectors of graph Laplacians which approximate eigenfunctions of $\Delta_{\mathcal{M}}$ of S^1. From left to right: L_{un}, $L_{rw'}$ using $W^{(1-\frac{d}{2})}$; L_{un}, $L_{rw'}$ using (4.2); degree $d_i^{-1}d_j^{-1}$normalized Diffusion Map [**11**]. Data as in Fig. 2, $N_x = 1000$, the kNN self-tune bandwidth $\hat{\rho}$ computed from X with $k_x = 21$, $\epsilon = 1e-4$.

eigenfunction approximation compared with the unnormalized graph Laplacian. We postpone the study of the random-walk graph Laplacian with self-tuned kernel to future investigation.

4.5 *Embedding of hand-written digits data*

We implement the spectral embedding on $N_x = 1000$ samples from the MNIST dataset, containing five classes (digits '0', '1', ..., '4') with 200 images in each class. The hand-written images can be viewed as lying near certain low-dimensional sub-manifolds in the $\mathbb{R}^{784}$ ambient space (each sample is a 28×28 gray-scale image). We use $k_x = 7$, and compute $\hat{R}_i$ as the L^2 distance between the i-th image to its k_x-th nearest neighbor. Also compute $\hat{\mu}_i := \frac{1}{N}\sum_{j=1}^{N} h_{kde}(\frac{\|x_i - x_j\|^2}{\epsilon_{kde}})$, where $\epsilon_{kde}^{1/2} = \text{Median}_i\{\hat{R}_i\}$. Here, $\hat{\mu}$ is the (un-normalized) density estimator. Consider two self-tuned kernel affinities:

$$W_{ij}^{(1)} = k_0\left(\frac{\|x_i - x_j\|^2}{\sigma_0^2 \hat{R}_i \hat{R}_j}\right)\frac{1}{\sigma_0^2 \hat{R}_i \hat{R}_j}, \quad W_{ij}^{'} = k_0\left(\frac{\|x_i - x_j\|^2}{\sigma_0^2 \hat{R}_i \hat{R}_j}\right)\frac{1}{\sigma_0^2 \hat{R}_i \hat{R}_j \sqrt{\hat{\mu}_i \hat{\mu}_j}}. \tag{4.3}$$

We use $L_{rw'} = D_{\hat{R}}^{-2}(D^{-1}W - I)$, where $W = W^{(1)}$ or $W^{'}$, $(D_{\hat{R}})_{ii} = \sigma_0 \hat{R}_i$, and D is the degree matrix of W. The parameter σ_0^2 serves as the dimension-less bandwidth 'ϵ'. We also compare with the fixed-bandwidth kernel affinity matrix, where $\epsilon^{1/2} = \sigma_0 \text{Median}_i(\hat{R}_i)$, called the Diffusion Map (DM) embedding. The embeddings by the first three (non-trivial) eigenvectors over a range of values of σ_0 are shown in Fig. 8.

We observe that the DM embedding is disconnected at small value of σ_0 and consists of points which are far away from the bulk (outlier points), due to sensitivity to data points which are relatively farther away from its neighbor samples. As illustrated in Fig. 9, the outlier points in the DM embedding are those whose values of $\hat{R}_i$ are large. In comparison, both the self-tuned kernels provide informative embeddings of the dataset over the range of values of σ_0, showing improved stability at small values of σ_0 to data samples lying at places where the data density is low. The $W^{(1)}$ kernel affinity shows a better stability than the W' kernel at the small value of σ_0 on this dataset, due to that the W' kernel still involves a fixed-bandwidth KDE $\hat{\mu}$ in the normalization.

5. Discussion

Apart from what has been mentioned in the text, the following lists a few possible future directions. First, we use a stand-alone Y to estimate the bandwidth function $\hat{\rho}$ for theoretical convenience. Extending the

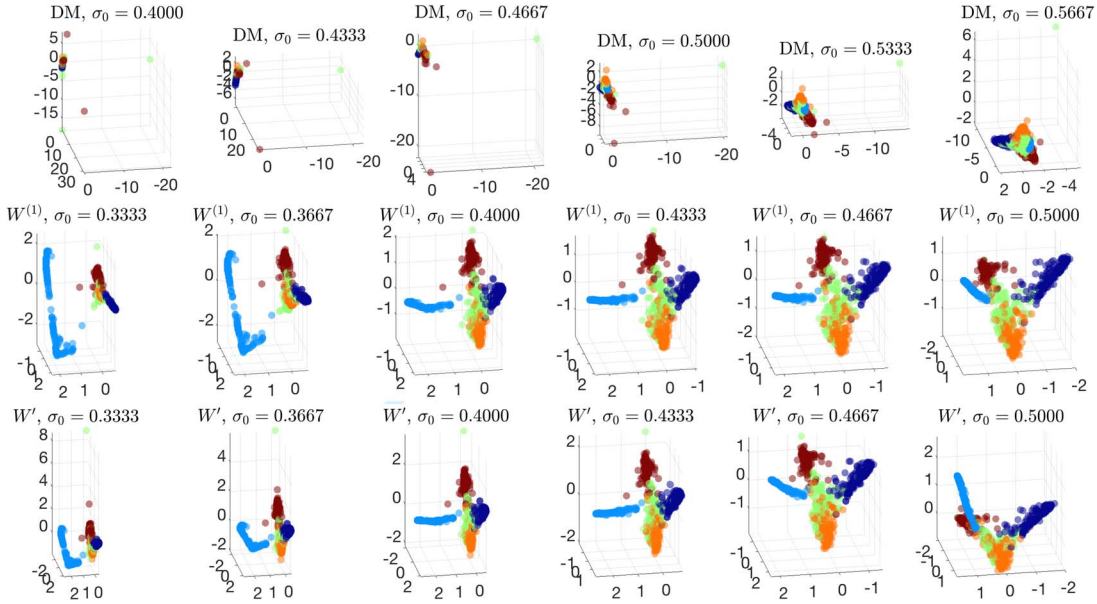

FIG. 8. Eigenvector embedding of $N_x = 1000$ MNIST hand-written digit images of 5 classes, $k_x = 7$, colored by digit class labels. (Top) By $\frac{K_{ij}}{d_i d_j}$ with fixed-bandwidth kernel, (Middle) by self-tuned kernel $W^{(1)}$, (Bottom) by self-tuned kernel W', as defined in (4.3).

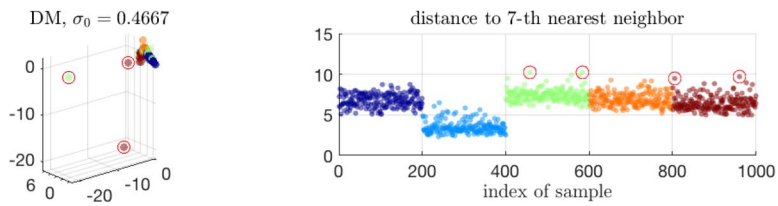

FIG. 9. Outliers in the fixed-bandwidth kernel embedding. (Left) One embedding in the top panel of Fig. 8 with a proper rotation for visualization purpose, where the outlier samples are marked with red circles. (Right) Values of $\hat{R}_i$, i.e., the distance to the seventh nearest neighbor, of the $N_x = 1000$ samples. The outlier samples in the left plot are marked by red circles. The plot is colored by digit class labels.

result to the case where $\hat{\rho}$ is computed from X itself can be of both theoretical and practical interest, especially when the number of data samples is not large. Second, one can continue to derive the spectral convergence, namely the convergence of eigenvalues and eigenvectors of the self-tuned graph Laplacian matrix to the eigenvalues and eigenfunctions of the associated limiting operators, with rates. Our graph Dirichlet form convergence rate is better than the operator point-wise convergence rate by a factor of $\epsilon^{-1/2}$, and since the Dirichlet form convergence largely implies the spectral convergence in the L^2 norm [8], this suggests that the spectral convergence rate may also be better than the pointwise convergence rate for the graph Laplacian operator in a proper sense. This theoretical speculation is supported by our empirical results. A uniform spectral convergence would also be important for various practical applications. At last, the random-walk graph Laplacian in our experiments sometimes shows a

better performance compared with the unnormalized graph Laplacian, especially in terms of eigenvector convergence. A theoretical justification then is needed, which is possibly similar to that in [54], and will be based on the spectral convergence result if can be established.

Acknowledgements

The authors thank the anonymous reviewers for the valuable feedback, which has helped to improve the work. The project was initiated as a *DoMath* Project titled 'Local affinity construction for dimension reduction methods' for undergraduate summer research, and the authors thank the Duke Mathematics Department for organizing and hosting the program. In 2018 summer, Tyler Lian, Inchan Hwang, Joseph Saldutti and Ajay Dheeraj participated in the project and contributed to initial experiments and analysis. The authors thank Didong Li for guiding the undergraduate team work as well as helpful discussion on NNDE and estimated bandwidth function.

Funding

National Science Foundation (DMS-2007040 to X.C.); Alfred P. Sloan Foundation (to X.C.).

References

1. Balasubramanian, M. & Schwartz, E. L. (2002) The isomap algorithm and topological stability. *Science*, **295**, 7–7.
2. Belkin, M. & Niyogi, P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, **15**, 1373–1396.
3. Belkin, M. & Niyogi, P. (2007) Convergence of Laplacian eigenmaps. *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 129–136.
4. Bermanis, A., Salhov, M., Wolf, G. & Averbuch, A. (2016) Measure-based diffusion grid construction and high-dimensional data discretization. *Appl. Comput. Harmon. Anal.*, **40**, 207–228.
5. Berry, T. & Harlim, J. (2016) Variable bandwidth diffusion kernels. *Appl. Comput. Harmon. Anal.*, **40**, 68–96.
6. Berry, T. & Sauer, T. (2016) Local kernels and the geometric structure of data. *Appl. Comput. Harmon. Anal.*, **40**, 439–469.
7. Borg, I. & Groenen, P. (2003) Modern multidimensional scaling: theory and applications. *J. Educ. Meas.*, **40**, 277–280.
8. Burago, D., Ivanov, S. & Kurylev, Y. (2015) A graph discretization of the Laplace–Beltrami operator. *J. Spectr. Theory*, **4**, 675–714.
9. Calder, J. & Trillos, N. G. (2019) Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs (in press). arXiv:1910.13476.
10. Cheng, X., Cloninger, A. & Coifman, R. R. (2020) Two-sample statistics based on anisotropic kernels. *Inf. Inference J. IMA*, **9**, 677–719.
11. Coifman, R. R. & Lafon, S. (2006) Diffusion maps. *Appl. Comput. Harmon. Anal.*, **21**, 5–30.
12. Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. & Zucker, S. W. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci.*, **102**, 7426–7431.
13. Crosskey, M. & Maggioni, M. (2017) ATLAS: a geometric approach to learning high-dimensional stochastic systems near manifolds. *Multiscale Model. Simul.*, **15**, 110–156.
14. Devroye, L. P. & Wagner, T. J. (1977) The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.*, **5**, 536–540.

15. Dunson, D. B., Wu, H.-T. & Wu, N. (2021) Spectral convergence of graph Laplacian and heat kernel reconstruction in L∞ from random samples. *Appl. Comput. Harmon. Anal.*, **55**, 282–336.

16. Eckhoff, M., et al. (2005) Precise asymptotics of small eigenvalues of reversible diffusions in the metastable regime. *Ann. Probab.*, **33**, 244–299.

17. Eldridge, J., Belkin, M. & Wang, Y. (2018) Unperturbed: spectral analysis beyond Davis-Kahan. *Algorithmic Learning Theory*. PMLR, pp. 321–358.

18. Gong, H., Pan, C., Yang, Q., Lu, H. & Ma, S. (2006) Neural network modeling of spectral embedding. *Proceedings of the British Machine Vision Conference*, Durham, England: BMVA Press, pp. 227–236.

19. Hall, P. (1983) On near neighbour estimates of a multivariate density. *J. Multivariate Anal.*, **13**, 24–39.

20. Hall, P., Hu, T. C. & Marron, J. S. (1995) Improved variable window kernel estimates of probability densities. *Ann. Statist.*, **23**, 1–10.

21. Hein, M. (2006) Uniform convergence of adaptive graph-based regularization. *International Conference on Computational Learning Theory*. Berlin, Heidelberg: Springer, pp. 50–64.

22. Hein, M., Audibert, J.-Y. & Von Luxburg, U. (2005) From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. *International Conference on Computational Learning Theory*. Berlin, Heidelberg: Springer, pp. 470–485.

23. Hinton, G. E. & Roweis, S. T. (2002) Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, Cambridge: MIT Press, pp. 857–864.

24. Li, Q., Han, Z. & Wu, X.-M. (2018) Deeper insights into graph convolutional networks for semi-supervised learning. *In Thirty-Second AAAI Conference on Artificial Intelligence.* (in press). arXiv:1801.07606.

25. Li, Q., Wu, X.-M., Liu, H., Zhang, X. & Guan, Z. (2019) Label efficient semi-supervised learning via graph filtering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington D.C.: IEEE Computer Society, pp. 9582–9591.

26. Little, A. V., Jung, Y.-M. & Maggioni, M. (2009) Multiscale estimation of intrinsic dimensionality of data sets. *2009 AAAI Fall Symposium Series*, Palo Alto, CA: AAAI Press.

27. Loftsgaarden, D. O., Quesenberry, C. P., et al. (1965) A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.*, **36**, 1049–1051.

28. Long, A. W. & Ferguson, A. L. (2017) Landmark diffusion maps (L-dMaps): accelerated manifold learning out-of-sample extension. *Appl. Comput. Harmon. Anal.*, **47**, 190–211.

29. Maaten, L. V. D. & Hinton, G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

30. Mack, Y. & Rosenblatt, M. (1979) Multivariate k-nearest neighbor density estimates. *J. Multivariate Anal.*, **9**, 1–15.

31. Marshall, N. F. & Coifman, R. R. (2019) Manifold learning with bi-stochastic kernels. *IMA J. Appl. Math.*, **84**, 455–482.

32. Masuda, N., Porter, M. A. & Lambiotte, R. (2017) Random walks and diffusion on networks. *Phys. Rep.*, **716**, 1–58.

33. Matkowsky, B. & Schuss, Z. (1981) Eigenvalues of the Fokker–Planck operator and the approach to equilibrium for diffusions in potential fields. *SIAM J. Appl. Math.*, **40**, 242–254.

34. Mishne, G., Shaham, U., Cloninger, A. & Cohen, I. (2017) Diffusion nets. *Appl. Comput. Harmon. Anal.*, **47**, 259–285.

35. Nadler, B., Lafon, S., Kevrekidis, I. & Coifman, R. R. (2006) Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators. *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 955–962.

36. Nadler, B., Srebro, N. & Zhou, X. (2009) Semi-supervised learning with the graph laplacian: the limit of infinite unlabelled data. *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, vol. 22, pp. 1330–1338.

37. Perrault-Joncas, D. C., Meila, M. & McQueen, J. (2017) Improved graph laplacian via geometric consistency. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 4460–4469.

38. Rohrdanz, M. A., Zheng, W., Maggioni, M. & Clementi, C. (2011) Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, **134**, 03B624.

39. Scholkopf, B. & Smola, A. J. (2018) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning Series, Cambridge, MA: MIT Press.

40. Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R. & Kluger, Y. (2018) Spectralnet: spectral clustering using deep neural networks. *In International Conference on Learning Representations.* (in press). arXiv:1801.01587.

41. Shen, C. & Wu, H.-T. (2020) Scalability and robustness of spectral embedding: landmark diffusion is all you need (in press). arXiv:2001.00801.

42. Singer, A. (2006) From graph to manifold Laplacian: the convergence rate. *Appl. Comput. Harmon. Anal.*, **21**, 128–134.

43. Singer, A., Erban, R., Kevrekidis, I. G. & Coifman, R. R. (2009) Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 16090–16095.

44. Singer, A. & Wu, H.-T. (2016) Spectral convergence of the connection laplacian from random samples. *Inf. Inference J. IMA*, **6**, 58–123.

45. Slepcev, D. & Thorpe, M. (2019) Analysis of p-Laplacian regularization in semisupervised learning. *SIAM J. Math. Anal.*, **51**, 2085–2120.

46. Talmon, R., Cohen, I., Gannot, S. & Coifman, R. R. (2013) Diffusion maps for signal processing: a deeper look at manifold-learning techniques based on kernels and graphs. *IEEE Signal Process Mag.*, **30**, 75–86.

47. Talmon, R. & Coifman, R. R. (2013) Empirical intrinsic geometry for nonlinear modeling and time series filtering. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 12535–12540.

48. Terrell, G. R. & Scott, D. W. (1992) Variable kernel density estimation. *Ann. Statist.*, **20**, 1236–1265.

49. The MNIST (Modified National Institute of Standards and Technology) database webpage. http://yann.lecun.com/exdb/mnist/.

50. Ting, D., Huang, L. & Jordan, M. (2010) An analysis of the convergence of graph Laplacians. *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 1079–1086.

51. Trillos, N. G., Gerlach, M., Hein, M. & Slepčev, D. (2020) Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Found. Comput. Math.*, **20**, 827–887.

52. Van Der Maaten, L., Postma, E. & Van den Herik, J. (2009) Dimensionality reduction: a comparative review. *J. Mach. Learn. Res.*, **10**, 13.

53. Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47. Cambridge, England: Cambridge University Press.

54. Von Luxburg, U., Belkin, M. & Bousquet, O. (2008) Consistency of spectral clustering. *Ann. Statist.*, **36**, 555–586.

55. Wang, S.-C., Wu, H.-T., Huang, P.-H., Chang, C.-H., Ting, C.-K. & Lin, Y.-T. (2020) Novel imaging revealing inner dynamics for cardiovascular waveform analysis via unsupervised manifold learning. *Anesth. Analg.*, **130**, 1244–1254.

56. Wang, X. (2015) Spectral convergence rate of graph Laplacian (in press). arXiv:1510.08110.

57. Wormell, C. L. & Reich, S. (2021) Spectral convergence of diffusion maps: improved error bounds and an alternative normalisation. *SIAM J. Numer. Anal.*, **59**, 1687–1734.

58. Zelnik-Manor, L. & Perona, P. (2005) Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 1601–1608.

A. Appendix

A.1 *Proofs*

A.1.1 *Proofs in Section 2*

Proof of Lemma 2.1. Given Y and $k > 1$ fixed, define

$$\mathcal{S}_Y := \left\{ x \in \mathbb{R}^D,\ s.t.\ \exists j \neq j',\ \|x - y_j\| = \|x - y_{j'}\| \right\}.$$

Since y_js are distinct points, $\mathcal{S}_Y$ is a collection of finitely many hyperplanes in $\mathbb{R}^D$ (finitely many points when $D = 1$), and $\mathcal{S}_Y \cap Y = \emptyset$. Whenever x lies outside $\mathcal{S}_Y$, the set $\{\|x - y_j\|\}_{j=1}^N$ consists of distinct non-negative values. The set $\mathbb{R}^D \backslash \mathcal{S}_Y$ is open and consists of a finite union of polygons (the polygons can be unbounded), as illustrated in Fig. A.1.

We prove the lemma in three parts as below.

Part 1: to prove that $\hat{R}$ is piece-wise C^∞ on $\mathbb{R}^D \backslash \mathcal{S}_Y$, and on each polygon $\mathbf{p}$ in $\mathbb{R}^D \backslash \mathcal{S}_Y$, $\hat{R}(x) = \|x - y_{\mathbf{p}}\|$ for a point $y_{\mathbf{p}} \in Y$ and outside $\mathbf{p}$.

First, for each (open) polygon $\mathbf{p}$ and any $x \in \mathbf{p}$, the k-nearest neighbor (kNN) of x in Y is uniquely defined due to the fact that the distance list $\{\|x - y_j\|\}_{j=1}^N$ has distinct values. Thus, the function $\hat{R}(x)$ equals $\|x - y^{(k,x)}\|$, where $y^{(k,x)}$ is the kNN of x in Y.

Second, we claim that the point $y^{(k,x)}$ is the same $y \in Y$ for all x inside the polygon $\mathbf{p}$ because the ordered list of nearest neighbors is fixed for all x within $\mathbf{p}$. Indeed, for the ordered list to cross, the distances of $\|x - y_j\|$ and $\|x - y_{j'}\|$ need to be equal at some x, and this x lies on $\mathcal{S}_Y$. We call this point $y_{\mathbf{p}}$, and then $\hat{R}(x) = \|x - y_{\mathbf{p}}\|$ for $x \in \mathbf{p}$.

Third, we claim that $y_{\mathbf{p}} \notin \mathbf{p}$. Note that each polygon $\mathbf{p}$ has at most one point y_j inside it. Because otherwise, suppose $y_j \neq y_{j'}$ are both inside $\mathbf{p}$, then so is the middle point $\frac{y_j + y_{j'}}{2}$ due to that $\mathbf{p}$ is convex, but $\frac{y_j + y_{j'}}{2}$ is in $\mathcal{S}_Y$ and cannot intersect with $\mathbf{p}$. Now if $y_{\mathbf{p}} \in \mathbf{p}$, then by definition $y_{\mathbf{p}}$ is the kNN of itself, which means that $k = 1$. This contradicts with the condition that $k > 1$.

The above gives us that $\hat{R}(x) = \|x - y_{\mathbf{p}}\|$ is C^∞ and hence $\|\bar{\nabla}\hat{R}\| = 1$ inside $\mathbf{p}$, by the fact that the mapping $x \mapsto \|x\|$ is C^∞ on $\mathbb{R}^D \backslash \{0\}$. These properties hold for all polygons $\mathbf{p}$; thus, $\hat{R}(x)$ is C^∞ on $\mathbb{R}^D \backslash \mathcal{S}_Y$, and $\|\bar{\nabla}\hat{R}\| = 1$ at point of differentiability.

Part 2: to prove that $\mathrm{Lip}_{\mathbb{R}^D}(\hat{R}) \leqslant 1$.

We assume that $\hat{R}$ is continuous $\mathbb{R}^D$, which will be proved in Part 3. By Part 1, we have that $\hat{R}$ is Lipschitz 1 on each open polygon $\mathbf{p}$, and combined with the continuity of $\hat{R}$ at points on the boundary of $\mathbf{p}$, we have that $\mathrm{Lip}_{\bar{\mathbf{p}}}(\hat{R}) \leqslant 1$, where $\bar{\mathbf{p}}$ is the closure of $\mathbf{p}$, that is,

$$|\hat{R}(z) - \hat{R}(z')| \leqslant \|z - z'\|, \quad \forall z, z' \in \bar{\mathbf{p}}. \tag{A.1}$$

For two points $x \neq x'$ in $\mathbb{R}^D$, we want to show that $|\hat{R}(x) - \hat{R}(x')| \leqslant \|x - x'\|$. Consider the segment line l connecting the two points. If l is contained in some $\bar{\mathbf{p}}$, then the claim is proved. Otherwise, there is a sub-segment l_0 connecting from x and $z_1 \in \mathcal{S}_Y$ such that l_0 is in some $\bar{\mathbf{p}}$. Continue the process gives finitely many distinct points $\{z_1, \cdots, z_M\} \subset l$ such that the sub-segment l_i connecting from z_i to z_{i+1} is contained in some $\bar{\mathbf{p}}$ for $i = 0$ to M, where $z_0 = x$ and $z_{M+1} = x'$. Note that this decomposition of l into the union of l_is holds even when one or both of x and x' are in $\mathcal{S}_Y$, as illustrated in Fig. A.1.

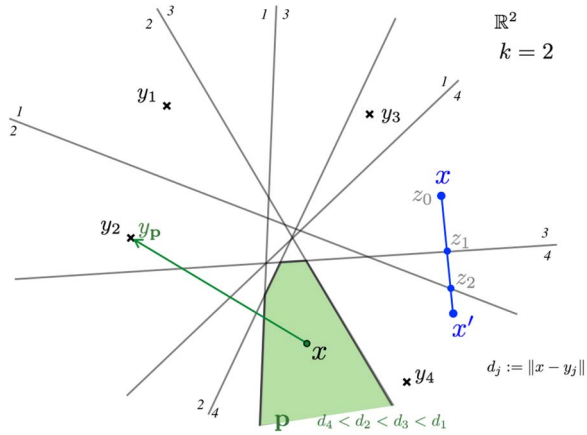

FIG. A.1. Illustration of the set $\mathcal{S}_Y$ and example polygons in the proof of Lemma 2.1, $D = 2$, $Y = \{y_1, \cdots, y_4\}$, $k = 2$. For each polygon $\mathbf{p}$, there is a point $y_{\mathbf{p}} \in Y$ such that $\hat{R}(x) = \|x - y_{\mathbf{p}}\|$ for $x \in \mathbf{p}$.

Now by construction, $\|x - x'\| = \sum_{i=0}^{M} \|z_i - z_{i+1}\|$. Meanwhile, applying (A.1) to each l_i gives that $|\hat{R}(z_i) - \hat{R}(z_{i+1})| \leqslant \|z_i - z_{i+1}\|$. Thus,

$$|\hat{R}(x) - \hat{R}(x')| \leqslant \sum_{i=0}^{M} |\hat{R}(z_i) - \hat{R}(z_{i+1})| \leqslant \sum_{i=0}^{M} \|z_i - z_{i+1}\| = \|x - x'\|.$$

Part 3: to prove that $\hat{R}$ is continuous on $\mathbb{R}^D$.

To finish the proof, it remains to prove the continuity of $\hat{R}$ on $\mathbb{R}^D$. For any $x_0 \in \mathbb{R}^D$, let $\hat{R}(x_0) = r_0$. Since Y has distinct points by assumption, at most one point y_j coincides with x_0. Since $k > 1$, $r_0 > 0$. We prove that when $x \to x_0$, $\hat{R}(x) \to r_0$. Define

$$F(x, r) := \sum_{j=1}^{N} \mathbf{1}_{\{\|x - y_j\| < r\}}.$$

Recall that

$$\hat{R}(x) = \inf\{r > 0, \text{ s.t. } F(x, r) \geqslant k\}.$$

Since $F(x_0, r)$ is monotonically increasing as r increases, for any $r' := r_0 + \varepsilon > r_0$, $F(x_0, r') \geqslant k$. This means that $|Y \cap B_{r'}(x_0)| := k' \geqslant k$. Since $B_{r'}(x_0)$ is an open ball, and there are k' many y_js lying inside it, they also all lie inside $B_{r''}(x_0)$ where $r_0 < r'' < r'$. Thus, when $\|x - x_0\| < (r' - r'')/2 := r'''$, these k' points of y_j also lie inside $B_{r'}(x)$, then $F(x, r') \geqslant k' \geqslant k$. This gives that $\hat{R}(x) \leqslant r_0 + \varepsilon$, whenever $\|x - x_0\| < r'''$.

Meanwhile, for any $0 < r' := r_0 - \varepsilon < r_0$, by definition $F(x_0, r' + \frac{\varepsilon}{2}) < k$, i.e., $|Y \cap B_{r'+\frac{\varepsilon}{2}}(x_0)| := k' < k$. This means that for any $y \in Y \backslash B_{r'+\frac{\varepsilon}{2}}(x_0)$, the distance $\|y - x_0\| \geqslant r' + \frac{\varepsilon}{2}$. Thus, when $\|x - x_0\| < \frac{\varepsilon}{4}$, the smallest distance $\|y - x\|$ for any $y \in Y \backslash B_{r'+\frac{\varepsilon}{2}}(x_0)$ is $\geqslant r' + \frac{\varepsilon}{4}$, and then $F(x, r') \leqslant k' < k$. This shows that $\hat{R}(x) \geqslant r_0 - \varepsilon$, whenever $\|x - x_0\| < \frac{\varepsilon}{4}$. Putting together, this proves the continuity of $\hat{R}(x)$ at x_0. □

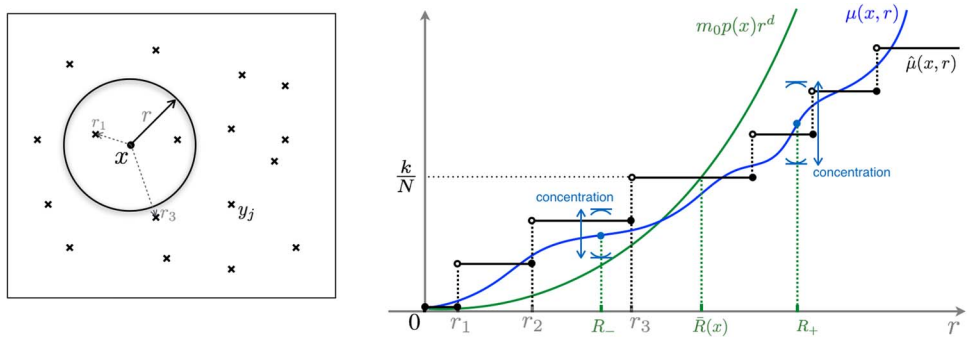

FIG. A.2. Given a dataset Y and a fixed x, plots of $\hat{\mu}(x,r)$, $\mu(x,r)$ and $m_0[h]p(x)r^d$ as functions of r. The values of $\hat{R}(x)$, $\bar{R}(x)$ and $R_\pm$ are marked. These quantities are used in the proof of Proposition 2.2.

Proof of Proposition 2.2. Recall that $\bar{\rho}(x) = p(x)^{-1/d}$. Define

$$\bar{R}(x) := \bar{\rho}(x) \left(\frac{1}{m_0[h]} \frac{k}{N} \right)^{1/d}. \tag{A.2}$$

Then, since we have $\hat{\rho}(x) = \hat{R}(x) \left(\frac{1}{m_0[h]} \frac{k}{N} \right)^{-1/d}$ and $\bar{\rho}(x) = \bar{R}(x) \left(\frac{1}{m_0[h]} \frac{k}{N} \right)^{-1/d}$, the proposition can be equivalently proved by controlling $|\hat{R}(x) - \bar{R}(x)|/\bar{R}(x)$. For the given $s > 0$, define

$$\delta_r := t_1 \left(\frac{k}{N} \right)^{2/d} + \frac{t_2}{d} \sqrt{\frac{s \log N}{k}}, \tag{A.3}$$

where $t_1 = \Theta^{[p]}(1)$, $t_2 = \Theta^{[1]}(1)$, both will be determined later. We will show that, when N exceeds a threshold depending on (p,s), for any $x \in \mathcal{M}$ fixed, w.p. greater than $1 - 2N^{-s/4}$,

$$\bar{R}(x)(1 - \delta_r) \leqslant \hat{R}(x) \leqslant \bar{R}(x)(1 + \delta_r). \tag{A.4}$$

To prove (A.4), we introduce some notations. Denote

$$R_-(x) := \bar{R}(x)(1 - \delta_r), \quad \text{and} \quad R_+(x) = \bar{R}(x)(1 + \delta_r).$$

Let $h = \mathbf{1}_{[0,1)}$, and define, for any $x \in \mathcal{M}$ and $r > 0$,

$$\hat{\mu}(x,r) := \frac{1}{N} \sum_{j=1}^{N} h\left(\frac{\|x - y_j\|^2}{r^2} \right) =: \frac{1}{N} \sum_{j=1}^{N} H_j(x,r),$$

then, by (2.1), $\hat{R}(x) = \inf_r \left\{ r > 0, \text{ s.t. } \hat{\mu}(x,r) \geqslant \frac{k}{N} \right\}$. For fixed x and r, H_j are i.i.d. random variables, and

$$\mathbb{E} H_j(x,r) = \int_{\mathcal{M}} h\left(\frac{\|x - y\|^2}{r^2} \right) p(y)\, \mathrm{d}V(y) =: \mu(x,r). \tag{A.5}$$

Below, to simplify notation, we omit the dependence on x in $\bar{R}$, $R_\pm$ and H_j when there is no confusion. The argument is for a fixed x, and we make sure that the constants t_1 and t_2 in δ_r as well as the large-N threshold are uniform for all x.

We first address the lower bound in (A.4). By definition, $\hat{\mu}(x,r)$ is monotonically increasing on $(0,\infty)$. We claim that

$$\Pr[\hat{R}(x) < R_-] \leqslant \Pr\left[\hat{\mu}(x,R_-) \geqslant \frac{k}{N}\right]. \tag{A.6}$$

Because $\hat{R}(x) = \inf\{r > 0, \hat{\mu}(x,r) \geqslant \frac{k}{N}\}$, if $\hat{R}(x) < R_-$, there is some r', $\hat{R}(x) < r' < R_-$ such that $\hat{\mu}(x,r') \geqslant \frac{k}{N}$, and by monotonicity $\hat{\mu}(x,R_-) \geqslant \hat{\mu}(x,r') \geqslant \frac{k}{N}$.

To bound the probability $\Pr\left[\hat{\mu}(x,R_-) \geqslant \frac{k}{N}\right]$, we use that the expectation $\mu(x,R_-)$ would be smaller than $\frac{k}{N}$ under some conditions for δ_r defined in (A.3).

Note that by definition (A.2), $\bar{R}(x) = \Theta^{[p]}((\frac{k}{N})^{1/d}) = o^{[p]}(1)$, and the implied constant is uniform for all x by the uniform boundedness of $\bar{\rho}$. Also, we have that $\delta_r = o^{[p]}(1)$ under the asymptotic condition on k. As a result, we have that

$$R_- = \Theta^{[p]}(\bar{R}(x)) = \Theta^{[p]}\left((\frac{k}{N})^{1/d}\right) = o^{[p]}(1). \tag{A.7}$$

Then, Lemma A.6 gives that when N is sufficiently large and then R_- is small,

$$\begin{aligned}
\mu(x,R_-) &= m_0[h]p(x)R_-^d + O^{[p]}(R_-^{d+2}) \\
&= m_0[h]p(x)\bar{R}^d(1-\delta_r)^d + O^{[p]}(\bar{R}^{d+2}) \\
&= m_0[h]p(x)\bar{R}^d\left(1 - d\delta_r + O^{[1]}(\delta_r^2) + O^{[p]}(\bar{R}^2)\right) \\
&\leqslant \frac{k}{N}(1 - 0.9d\delta_r) + O^{[p]}(\bar{R}^{d+2}) =: \frac{k}{N} - \delta_{\mu_-}, \text{(by that } m_0[h]p(x)\bar{R}^d = \frac{k}{N})
\end{aligned} \tag{A.8}$$

where the inequality in the last row is obtained by that $\delta_r = o^{[p]}(1)$, and the large-N threshold here only depends on p. Note that the implied constant of $O^{[p]}(\bar{R}^{d+2})$, denoted as c_p, is uniform for all x. Meanwhile, by uniform boundedness of p from below, we have

$$\bar{R}(x) \leqslant \max_{x\in\mathcal{M}} \bar{\rho}(x)\left(\frac{1}{m_0[h]}\frac{k}{N}\right)^{1/d} = \left(\frac{1}{p_{min}m_0[h]}\right)^{1/d}\left(\frac{k}{N}\right)^{1/d}.$$

Denote $c_{p,1} := \left(p_{min}m_0[h]\right)^{-1/d}$, and choose

$$t_1 := \frac{c_p c_{p,1}^{d+2}}{0.8d} = \Theta^{[p]}(1), \tag{A.9}$$

which is uniform for all x, then $t_1 \cdot 0.9d\left(\frac{k}{N}\right)^{1+2/d} > c_p c_{p,1}^{d+2}\left(\frac{k}{N}\right)^{1+2/d} \geqslant c_p\bar{R}^{d+2}$. Thus, when N is sufficiently large and the threshold depends on p, we have

$$\begin{aligned}
\delta_{\mu_-} &= 0.9d\frac{k}{N}\left(t_1\left(\frac{k}{N}\right)^{2/d} + \frac{t_2}{d}\sqrt{\frac{s\log N}{k}}\right) + O^{[p]}(\bar{R}^{d+2}) \\
&> t_2 \cdot 0.9\frac{k}{N}\sqrt{\frac{s\log N}{k}} = t_2 \cdot 0.9\left(\frac{k}{N}\right)^{1/2}\sqrt{\frac{s\log N}{N}} =: \tilde{s}.
\end{aligned} \tag{A.10}$$

To use the concentration of $\hat{\mu}(R_-)$ at $\mu(R_-)$, we compute the boundedness and variance of $H_j(R_-)$. Because $0 \leqslant h \leqslant 1$, so is H_j, and then $|H_j| \leqslant L_H = 1$. The variance

$$\mathrm{Var}(H_j) \leqslant \mathbb{E}H_j^2 = \int_{\mathcal{M}} h^2\left(\frac{\|x-y\|^2}{R_-^2}\right) p(y)\,\mathrm{d}V(y) = \mu(R_-),$$

because the kernel function $h : \mathbb{R} \to \mathbb{R}$ satisfies $h^2 = h$. Thus, by that $\mu(R_-) = m_0[h]p(x)R_-^d + O^{[p]}(R_-^{d+2})$ with (A.7), and that $\delta_r = o^{[p]}(1)$, when N is sufficiently large,

$$\mathrm{Var}(H_j) \leqslant 1.1 m_0[h]p(x)R_-^d \leqslant 1.5 m_0[h]p(x)\bar{R}^d = 1.5\frac{k}{N} =: \bar{\nu}_H,$$

and the two inequalities hold when N exceeds a threshold depending on p only. By the classical Bernstein inequality, as long as $\tilde{s}L_H < 3\bar{\nu}_H$, then

$$\Pr[\hat{\mu}(R_-) - \mu(R_-) > \tilde{s}] < e^{-\frac{1}{4}\tilde{s}^2\frac{N}{\bar{\nu}_H}}.$$

To verify that $\tilde{s}L_H < 3\bar{\nu}_H$: note that it is equivalent to that $t_2 \cdot 0.9 < 3 \cdot 1.5(\frac{k}{s\log N})^{1/2}$, and since we have assumed $k = \Omega(\log N)$, if we have $t_2 = \Theta(1)$, then it holds when N is sufficiently large where the threshold depends on s. This is fulfilled by setting t_2 being an absolute constant such that

$$\frac{(t_2 0.9)^2}{4 \cdot 1.5} = 1, \quad 0 < t_2 < 3. \tag{A.11}$$

Thus, together with (A.8) and (A.10), we have

$$\mu(R_-) \leqslant \frac{k}{N} - \delta_{\mu_-} < \frac{k}{N} - \tilde{s}.$$

As a result, (A.6) continues as

$$\Pr[\hat{R}(x) < R_-] \leqslant \Pr\left[\hat{\mu}(R_-) \geqslant \frac{k}{N}\right] \leqslant \Pr\left[\hat{\mu}(R_-) > \mu(R_-) + \tilde{s}\right] < e^{-\frac{1}{4}\tilde{s}^2\frac{N}{\bar{\nu}_H}} = N^{-s},$$

which proves that w.p. higher than $1 - N^{-s}$, the lower bound $\hat{R}(x) \geqslant R_-$ holds. We call the event $[\hat{R}(x) \geqslant R_-]$ the good event E_1. All the large-N thresholds depend on (p, s) and are uniform for all x.

The upper bound is proved in a similar way. Specifically,

$$\begin{aligned}\mu(R_+) &= \frac{k}{N}(1+\delta r)^d + O^{[p]}(R_+^{d+2}) \geqslant \frac{k}{N}(1 + 0.9d\delta r) + O^{[p]}(\bar{R}^{d+2}) \\ &= \frac{k}{N} + 0.9d\frac{k}{N}\left(t_1(\frac{k}{N})^{2/d} + \frac{t_2}{d}\sqrt{\frac{s\log N}{k}}\right) + O^{[p]}(\bar{R}^{d+2}),\end{aligned}$$

and the implied constant in $O^{[p]}(\bar{R}^{d+2})$, c_p, is same as the above by Lemma A.6. Then, again by the uniform upper bound of $\bar{R}(x)$ by $c_{p,1}(\frac{k}{N})^{1/d}$, by setting t_1 to be that in (A.9), we have

$$\mu(R_+) > \frac{k}{N} + t_2 0.9\left(\frac{k}{N}\right)^{1/2}\sqrt{\frac{s\log N}{N}} = \frac{k}{N} + \tilde{s}.$$

Same as before, $H_j(R_+)$ is bounded by 1 and for a sufficiently large N,

$$\mathrm{Var}(H_j) \leqslant \mathbb{E}H_j^2 = \mu(R_+) \leqslant 1.5\frac{k}{N} = \bar{\nu}_H\,.$$

By letting t_2 as in (A.11), we have

$$\Pr[\hat{R}(x) > R_+] \leqslant \Pr\left[\hat{\mu}(R_+) < \frac{k}{N}\right] \leqslant \Pr\left[\hat{\mu}(R_+) < \mu(R_+) - \tilde{s}\right] < e^{-\frac{1}{4}\tilde{s}^2\frac{N}{\bar{\nu}_H}} = N^{-s}.$$

This proves that w.p. higher than $1 - N^{-s}$, the upper bound $\hat{R}(x) \leqslant R_+$ holds. We call the event $[\hat{R}(x) \leqslant R_+]$ the good event E_2.

Putting the above together, under the event $E_1 \cap E_2$, which happens w.p. greater than $1 - 2N^{-s}$,

$$\frac{|\hat{R}(x) - \bar{R}(x)|}{\bar{R}(x)} \leqslant \delta_r = \frac{c_p c_{p,1}^{d+2}}{0.8d}\left(\frac{k}{N}\right)^{2/d} + \frac{\frac{2\sqrt{1.5}}{0.9}}{d}\sqrt{\frac{s\log N}{k}},$$

which proves the claim of the proposition. □

Proof of Theorem 2.3. We restrict to when Y has distinct points, which, under Assumption 2.1, holds w.p. 1, and then Lemma 2.1 holds.

We cover $\mathcal{M}$ using r-Euclidean balls, where $r > 0$ is a constant of order $(k/N)^{3/d}$ with the implied constant to be determined. Suppose N is large enough such that $r < \delta_0$ in Lemma A.3, then by Lemma A.4, we can find an r-net $F := \{x_1, \cdots, x_n\}$ whose cardinal number is n, $n \leqslant V(\mathcal{M})r^{-d}$. We ask for the bound in Proposition 2.2 to hold at each x_i, where $s > 0$ will be chosen later as an $\Theta^{[1]}(1)$ constant. Then, when N exceeds a threshold depending on p and uniform for all x_i, by a union bound, under a good event $E_{\hat{\rho},net}$ which happens w.p. higher than $1 - 2nN^{-s}$, we have

$$\left|\frac{\hat{\rho}(x_i)}{\bar{\rho}(x_i)} - 1\right| \leqslant t_1\left(\frac{k}{N}\right)^{2/d} + \frac{t_2}{d}\sqrt{\frac{s\log N}{k}} := \varepsilon, \quad \text{for all } i = 1, \cdots, n, \tag{A.12}$$

where $t_1 = \Theta^{[p]}(1)$ and $t_2 = \Theta^{[1]}(1)$ are defined as in the proof of Proposition 2.2. Under the asymptotic condition on k, $\varepsilon = o^{[p]}(1)$ as $N \to \infty$.

We now consider $\hat{\rho}/\bar{\rho}$ on each $\bar{B}_r(x_i) \cap \mathcal{M}$. Because $\bar{\rho} = p^{-1/d}$ is C^∞ on $\mathcal{M}$, $\sup_{x\in\mathcal{M}} |\nabla_{\mathcal{M}}\bar{\rho}(x)| \leqslant L_p$. Then, for each x_i, by (A.55),

$$|\bar{\rho}(x) - \bar{\rho}(x_i)| \leqslant L_p d_{\mathcal{M}}(x, x_i) \leqslant 1.1L_p\|x - x_i\| \leqslant 1.1L_p r, \quad \forall x \in \bar{B}_r(x_i) \cap \mathcal{M}.$$

Meanwhile, Lemma 2.1 gives that

$$\mathrm{Lip}_{\mathbb{R}^D}(\hat{\rho}) = \left(\frac{1}{m_0[h]}\frac{k}{N}\right)^{-1/d}\mathrm{Lip}_{\mathbb{R}^D}(\hat{R}) \leqslant \left(\frac{1}{m_0[h]}\frac{k}{N}\right)^{-1/d},$$

so we have

$$|\hat{\rho}(x) - \hat{\rho}(x_i)| \leqslant \left(\frac{1}{m_0[h]}\frac{k}{N}\right)^{-1/d} r, \quad \forall\ \mathrm{x}]\ \in \bar{B}_r(x_i) \cap \mathcal{M}.$$

Together, we have that $\forall x \in \bar{B}_r(x_i) \cap \mathcal{M}$,

$$\left|\frac{\hat{\rho}(x)}{\bar{\rho}(x)} - \frac{\hat{\rho}(x_i)}{\bar{\rho}(x_i)}\right| \leqslant \frac{1}{\bar{\rho}(x)}\left|(\hat{\rho}(x)) - \hat{\rho}(x_i)) - \frac{\hat{\rho}(x_i)}{\bar{\rho}(x_i)}(\bar{\rho}(x) - \bar{\rho}(x_i))\right|$$
$$\leqslant \frac{1}{\rho_{min}}\left|\left(\frac{1}{m_0[h]}\frac{k}{N}\right)^{-1/d} + (1+\varepsilon)\cdot 1.1L_p\right| r = \Theta^{[p]}\left((k/N)^{-1/d}\right) r. \tag{A.13}$$

Thus, one can choose $r > 0$ to be $\Theta^{[p]}((k/N)^{3/d})$ so as to make (A.13) bounded by $t_1(\frac{k}{N})^{2/d}$ when N is sufficiently large, where the threshold of N depends on p only. This gives that $\left|\frac{\hat{\rho}(x)}{\bar{\rho}(x)} - \frac{\hat{\rho}(x_i)}{\bar{\rho}(x_i)}\right| \leqslant t_1(\frac{k}{N})^{2/d}$. Meanwhile, we already have (A.12) under $E_{\hat{\rho},net}$, and putting together,

$$\left|\frac{\hat{\rho}(x)}{\bar{\rho}(x)} - 1\right| \leqslant \left|\frac{\hat{\rho}(x)}{\bar{\rho}(x)} - \frac{\hat{\rho}(x_i)}{\bar{\rho}(x_i)}\right| + \left|\frac{\hat{\rho}(x_i)}{\bar{\rho}(x_i)} - 1\right| \leqslant t_1\left(\frac{k}{N}\right)^{2/d} + \varepsilon, \quad \forall x \in \bar{B}_r(x_i) \cap \mathcal{M}.$$

By that $\mathcal{M} \subset \cup_i \bar{B}_r(x_i)$, the above bound holds for all $x \in \mathcal{M}$. Recall the definition of ε in (A.12), we have that, under $E_{\hat{\rho},net}$,

$$\sup_{x\in\mathcal{M}}\left|\frac{\hat{\rho}(x)}{\bar{\rho}(x)} - 1\right| \leqslant 2t_1\left(\frac{k}{N}\right)^{2/d} + \frac{t_2}{d}\sqrt{\frac{s\log N}{k}}.$$

Finally, to show the high probability of $E_{\hat{\rho},net}$, by that $n \leqslant V(\mathcal{M})r^{-d}$,

$$2nN^{-s} \leqslant 2V(\mathcal{M})r^{-d}N^{-s} \leqslant c_p\left(\frac{k}{N}\right)^{-3}N^{-s} \quad \text{(constant } c_p \text{ depending on } p\text{)}$$
$$\leqslant N^{-s+3} \quad \text{(with large } N, \text{ because } k = \Omega(1)\text{)}.$$

so by setting $s = 13$, we have that $E_{\hat{\rho},net}$ happens w.p. higher than $> 1 - N^{-s+3} = 1 - N^{-10}$. □

A.1.2 *Proof of Proposition* 3.2

Proof of Proposition 3.2. To simplify the notation, when there is no danger of confusion, we omit the dependence of $m_l[h]$ on h and use the notation m_l, where $l = 0, 2$.

Under the condition that

$$\sup_{x\in\mathcal{M}}\frac{|\hat{\rho}(x) - \bar{\rho}(x)|}{|\bar{\rho}(x)|} < \varepsilon_\rho < 0.1, \tag{A.14}$$

we have that

$$0.9\bar{\rho}(x) < \hat{\rho}(x) < 1.1\bar{\rho}(x), \quad \forall x \in \mathcal{M}. \tag{A.15}$$

Recall that

$$\mathcal{E}^{(\alpha)}(f,f) = \frac{\epsilon^{-\frac{d}{2}-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2 k_0\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\hat{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\,\mathrm{d}V(x)\,\mathrm{d}V(y) =: ①,$$

and we consider the counterpart of ① where $\hat{\rho}(x)$ is replaced with $\bar{\rho}(x)$, namely,

$$② := \frac{\epsilon^{-\frac{d}{2}-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2 k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\bar{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

With the operator $G_\epsilon^{(\rho)}$ defined as in (3.6), writing the integration over $dV(x)$ via $G_\epsilon^{(\bar\rho)}$,

$$\textcircled{2} = \frac{1}{\epsilon m_2}\left(\int_{\mathcal{M}} (p\hat\rho^{d/2-\alpha})(y) G^{(\bar\rho)}_{\epsilon\hat\rho(y)} \frac{f^2 p}{\bar\rho^\alpha}(y)\,\mathrm{d}V(y) - 2\int_{\mathcal{M}} (fp\hat\rho^{d/2-\alpha})(y) G^{(\bar\rho)}_{\epsilon\hat\rho(y)} \frac{fp}{\bar\rho^\alpha}(y)\,\mathrm{d}V(y) \right.$$
$$\left. + \int_{\mathcal{M}} (f^2 p\hat\rho^{d/2-\alpha})(y) G^{(\bar\rho)}_{\epsilon\hat\rho(y)} \frac{p}{\bar\rho^\alpha}(y)\,\mathrm{d}V(y)\right). \tag{A.16}$$

Recall that $\bar\rho = p^{-1/d}$ is in $C^\infty(\mathcal{M})$ and uniformly bounded from below and above. By Lemma 3.1,

$$G^{(\bar\rho)}_{\epsilon\hat\rho(y)} \frac{f^2 p}{\bar\rho^\alpha} = m_0 f^2 p\bar\rho^{\frac{d}{2}-\alpha} + \epsilon\hat\rho\frac{m_2}{2}(\omega f^2 p\bar\rho^{1+\frac{d}{2}-\alpha} + \Delta(f^2 p\bar\rho^{1+\frac{d}{2}-\alpha})) + \hat\rho^2 r_1^{(2)},$$
$$G^{(\bar\rho)}_{\epsilon\hat\rho(y)} \frac{fp}{\bar\rho^\alpha} = m_0 fp\bar\rho^{\frac{d}{2}-\alpha} + \epsilon\hat\rho\frac{m_2}{2}(\omega fp\bar\rho^{1+\frac{d}{2}-\alpha} + \Delta(fp\bar\rho^{1+\frac{d}{2}-\alpha})) + \hat\rho^2 r_2^{(2)},$$
$$G^{(\bar\rho)}_{\epsilon\hat\rho(y)} \frac{p}{\bar\rho^\alpha} = m_0 p\bar\rho^{\frac{d}{2}-\alpha} + \epsilon\hat\rho\frac{m_2}{2}(\omega p\bar\rho^{1+\frac{d}{2}-\alpha} + \Delta(p\bar\rho^{1+\frac{d}{2}-\alpha})) + \hat\rho^2 r_3^{(2)},$$

where $\|r_1^{(2)}\|_\infty = O^{[f,p]}(\epsilon^2)$, $\|r_2^{(2)}\|_\infty = O^{[f,p]}(\epsilon^2)$ and $\|r_3^{(2)}\|_\infty = O^{[p]}(\epsilon^2)$ and we omit the evaluation of all functions at y in the notation. Then, (A.16) becomes

$$\begin{aligned}
\textcircled{2} &= \frac{1}{\epsilon m_2}\left(\int_{\mathcal{M}} (p\hat\rho^{d/2-\alpha})\left\{m_0 f^2 p\bar\rho^{\frac{d}{2}-\alpha} + \epsilon\hat\rho\frac{m_2}{2}(\omega f^2 p\bar\rho^{1+\frac{d}{2}-\alpha} + \Delta(f^2 p\bar\rho^{1+\frac{d}{2}-\alpha})) + \hat\rho^2 r_1^{(2)}\right\}\right.\\
&\quad - 2\int_{\mathcal{M}} (fp\hat\rho^{d/2-\alpha})\left\{m_0 fp\bar\rho^{\frac{d}{2}-\alpha} + \epsilon\hat\rho\frac{m_2}{2}(\omega fp\bar\rho^{1+\frac{d}{2}-\alpha} + \Delta(fp\bar\rho^{1+\frac{d}{2}-\alpha})) + \hat\rho^2 r_2^{(2)}\right\}\\
&\quad \left. + \int_{\mathcal{M}} (f^2 p\hat\rho^{d/2-\alpha})\left\{m_0 p\bar\rho^{\frac{d}{2}-\alpha} + \epsilon\hat\rho\frac{m_2}{2}(\omega p\bar\rho^{1+\frac{d}{2}-\alpha} + \Delta(p\bar\rho^{1+\frac{d}{2}-\alpha})) + \hat\rho^2 r_3^{(2)}\right\}\right)\\
&= \frac{1}{2}\int_{\mathcal{M}} p\hat\rho^{\frac{d}{2}-\alpha+1}\left\{\Delta(f^2 p\bar\rho^{1+\frac{d}{2}-\alpha}) - 2f\Delta(fp\bar\rho^{1+\frac{d}{2}-\alpha}) + f^2\Delta(p\bar\rho^{1+\frac{d}{2}-\alpha})\right\}\\
&\quad + \frac{1}{\epsilon m_2}\int_{\mathcal{M}} p\hat\rho^{\frac{d}{2}-\alpha+2}(r_1^{(2)} - 2fr_2^{(2)} + f^2 r_3^{(2)})\\
&=: \textcircled{2}_1 + \textcircled{2}_2,
\end{aligned}$$

where again, we omit the evaluation of all functions at y and the integration over $dV(y)$ in the notation, and same in below.

We first consider $\textcircled{2}_1$. Note that, by defining $g := p\bar\rho^{1+d/2-\alpha}$, the bracket inside the integrand becomes

$$\{\cdots\} = \Delta(f^2 g) - 2f\Delta(fg) + f^2\Delta g = 2g|\nabla f|^2 = 2p\bar\rho^{1+d/2-\alpha}|\nabla f|^2.$$

We define $\textcircled{2}_1'$ by substituting $\hat\rho$ with $\bar\rho$ in $\textcircled{2}_1$, namely,

$$\textcircled{2}_1' := \frac{1}{2}\int_{\mathcal{M}} p\bar\rho^{\frac{d}{2}-\alpha+1}\{\cdots\}, \quad \textcircled{2}_1 - \textcircled{2}_1' = \frac{1}{2}\int_{\mathcal{M}} p(\hat\rho^{\frac{d}{2}-\alpha+1} - \bar\rho^{\frac{d}{2}-\alpha+1})\{\cdots\}.$$

Inserting the expression of $\{\cdots\}$, by the definition of $\bar{\rho} = p^{-1/d}$ and p_α, we have

$$②'_1 = \int_{\mathcal{M}} p^2 \bar{\rho}^{2+d-2\alpha} |\nabla f|^2 = \int_{\mathcal{M}} p_\alpha |\nabla f|^2 = \mathcal{E}_{p_\alpha}(f,f),$$

and also

$$②_1 - ②'_1 = \int_{\mathcal{M}} p^2 \bar{\rho}^{1+d/2-\alpha} |\nabla f|^2 (\hat{\rho}^{\frac{d}{2}-\alpha+1} - \bar{\rho}^{\frac{d}{2}-\alpha+1}).$$

Since p and $\bar{\rho}$ are positive, we have

$$|②_1 - ②'_1| \leqslant \int_{\mathcal{M}} p^2 \bar{\rho}^{1+d/2-\alpha} |\nabla f|^2 |\hat{\rho}^{\frac{d}{2}-\alpha+1} - \bar{\rho}^{\frac{d}{2}-\alpha+1}|.$$

Let $\gamma := d/2 - \alpha + 1 \in \mathbb{R}$. By the mean value theorem and (A.15), for any y, there exists $\xi > 0$ between $\hat{\rho}(y)$ and $\bar{\rho}(y)$, $\xi^{\gamma-1} \leqslant \max\{1.1^{\gamma-1}, 0.9^{\gamma-1}\}\bar{\rho}(y)^{\gamma-1}$ such that

$$|\hat{\rho}(y)^\gamma - \bar{\rho}(y)^\gamma| = \gamma \xi^{\gamma-1} |\hat{\rho}(y) - \bar{\rho}(y)| \leqslant \gamma \max\{1.1^{\gamma-1}, 0.9^{\gamma-1}\} \bar{\rho}(y)^{\gamma-1} |\hat{\rho}(y) - \bar{\rho}(y)| \leqslant c_\gamma \bar{\rho}(y)^\gamma \varepsilon_\rho, \tag{A.17}$$

where the last inequality comes from (A.14) and $c_\gamma := \gamma \max\{1.1^{\gamma-1}, 0.9^{\gamma-1}\}$ is a constant determined by α and d. Then

$$|②_1 - ②'_1| \leqslant c_\gamma \varepsilon_\rho \int_{\mathcal{M}} p^2 \bar{\rho}^{1+d/2-\alpha} |\nabla f|^2 \bar{\rho}^\gamma = c_\gamma \varepsilon_\rho \cdot ②'_1,$$

which gives that

$$②_1 = ②'_1 (1 + O(c_\gamma \varepsilon_\rho)) = \mathcal{E}_{p_\alpha}(f,f)(1 + O^{[\alpha]}(\varepsilon_\rho)), \tag{A.18}$$

Next we bound $②_2$. By definition,

$$|②_2| \leqslant \frac{1}{\epsilon m_2} \int_{\mathcal{M}} p \hat{\rho}^{\frac{d}{2}-\alpha+2} (|r_1^{(2)}| + 2|f||r_2^{(2)}| + |f|^2 |r_3^{(2)}|).$$

Again, by (A.15), $\hat{\rho}(y)^{d/2-\alpha+2} \leqslant \max\{1.1^{d/2-\alpha+2}, 0.9^{d/2-\alpha+2}\}\bar{\rho}(y)^{d/2-\alpha+2} := c'_\alpha \bar{\rho}(y)^{d/2-\alpha+2}$. Then

$$|②_2| \leqslant O^{[f,p]}(\epsilon) \int_{\mathcal{M}} c'_\alpha p \bar{\rho}^{\frac{d}{2}-\alpha+2} = O^{[f,p]}(\epsilon).$$

Together with (A.18), we have that

$$② = \mathcal{E}_{p_\alpha}(f,f)(1 + O^{[\alpha]}(\varepsilon_\rho)) + O^{[f,p]}(\epsilon). \tag{A.19}$$

It remains to bound $|② - ①|$ to prove the same bound for ①. Define

$$③ := \frac{\epsilon^{-\frac{d}{2}-1}}{m_2} \int_{\mathcal{M}} \int_{\mathcal{M}} (f(x) - f(y))^2 k_0 \left(\frac{\|x-y\|^2}{\epsilon \bar{\rho}(x) \hat{\rho}(y)} \right) \frac{p(x)p(y)}{\hat{\rho}(x)^\alpha \hat{\rho}(y)^\alpha} \, \mathrm{d}V(x) \, \mathrm{d}V(y).$$

Then,

$$③ - ② = \frac{\epsilon^{-\frac{d}{2}-1}}{m_2} \int_{\mathcal{M}} \int_{\mathcal{M}} (f(x) - f(y))^2 \left(\frac{\bar{\rho}(x)^\alpha}{\hat{\rho}(x)^\alpha} - 1 \right) k_0 \left(\frac{\|x-y\|^2}{\epsilon \bar{\rho}(x) \hat{\rho}(y)} \right) \frac{p(x)p(y)}{\bar{\rho}(x)^\alpha \hat{\rho}(y)^\alpha} \, \mathrm{d}V(x) \, \mathrm{d}V(y). \tag{A.20}$$

By (A.14) and (A.15), we have that

$$\sup_{x\in\mathcal{M}}\left|\frac{\bar{\rho}(x)^{\alpha}}{\hat{\rho}(x)^{\alpha}}-1\right|=O^{[\alpha]}(\varepsilon_{\rho}),\quad \sup_{x\in\mathcal{M}}\left|\frac{\bar{\rho}(x)}{\hat{\rho}(x)}-1\right|=O^{[1]}(\varepsilon_{\rho}). \tag{A.21}$$

Note that by definition, ② $\geqslant 0$. Therefore, by that $k_0\geqslant 0$ and $p,\bar{\rho},\hat{\rho}>0$,

$$\begin{aligned}|③-②|&\leqslant O^{[\alpha]}(\varepsilon_{\rho})\cdot\frac{\epsilon^{-\frac{d}{2}-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\bar{\rho}(x)^{\alpha}\hat{\rho}(y)^{\alpha}}\,\mathrm{d}V(x)\,\mathrm{d}V(y)\\&=O^{[\alpha]}(\varepsilon_{\rho})②.\end{aligned}$$

Together with (A.19), this gives that

$$③=②(1+O^{[\alpha]}(\varepsilon_{\rho}))=\mathcal{E}_{p_{\alpha}}(f,f)(1+O^{[\alpha]}(\varepsilon_{\rho}))+O^{[f,p]}(\epsilon). \tag{A.22}$$

Meanwhile,

$$③-①=\frac{\epsilon^{-\frac{d}{2}-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2\left(k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)-k_0\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\right)\frac{p(x)p(y)}{\hat{\rho}(x)^{\alpha}\hat{\rho}(y)^{\alpha}}\,\mathrm{d}V(x)\,\mathrm{d}V(y),$$

and

$$k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)-k_0\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)=k_0'(\xi)\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\left(1-\frac{\bar{\rho}(x)}{\hat{\rho}(x)}\right),$$

where ξ is between $\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}$ and $\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}$. By (A.15), $\xi\geqslant\frac{\|x-y\|^2}{\epsilon 1.1\bar{\rho}(x)\hat{\rho}(y)}$, and then,

$$|k_0'(\xi)|\leqslant a_1e^{-a\xi}\leqslant a_1e^{-a\frac{\|x-y\|^2}{\epsilon 1.1\bar{\rho}(x)\hat{\rho}(y)}}.$$

By (A.21), we have that

$$\begin{aligned}|③-①|&\leqslant\frac{\epsilon^{-\frac{d}{2}-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2|k_0'(\xi)|\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\left|1-\frac{\bar{\rho}(x)}{\hat{\rho}(x)}\right|\frac{p(x)p(y)}{\hat{\rho}(x)^{\alpha}\hat{\rho}(y)^{\alpha}}\,\mathrm{d}V(x)\,\mathrm{d}V(y)\\&\leqslant O(\varepsilon_{\rho})\frac{\epsilon^{-\frac{d}{2}-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2a_1e^{-\frac{a}{1.1}\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}}\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\frac{p(x)p(y)}{\hat{\rho}(x)^{\alpha}\hat{\rho}(y)^{\alpha}}\,\mathrm{d}V(x)\,\mathrm{d}V(y)\\&=O(\varepsilon_{\rho})\frac{\epsilon^{-\frac{d}{2}-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^2k_1\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\hat{\rho}(x)^{\alpha}\hat{\rho}(y)^{\alpha}}\,\mathrm{d}V(x)\,\mathrm{d}V(y)\\&=O(\varepsilon_{\rho})\frac{m_2[k_1]}{m_2[k_0]}\cdot③',\end{aligned} \tag{A.23}$$

where ③′ is defined by replacing k_0 to be k_1 in ③, where

$$k_1(r):=a_1re^{-\frac{a}{1.1}r}\text{ and }r\geqslant 0.$$

Since k_1 satisfies Assumption 3.1, our analysis of ② and ③ with k_1 so far applies. Thus, by (A.19) and (A.22), we have

$$③'=\mathcal{E}_{p_{\alpha}}(f,f)(1+O^{[\alpha]}(\varepsilon_{\rho}))+O^{[f,p]}(\epsilon),$$

and then

$$|③ - ①| = O(\varepsilon_\rho)(\mathcal{E}_{p_\alpha}(f,f)(1 + O^{[\alpha]}(\varepsilon_\rho)) + O^{[f,p]}(\epsilon)) = \mathcal{E}_{p_\alpha}(f,f)O^{[\alpha]}(\varepsilon_\rho) + O^{[f,p]}(\epsilon\varepsilon_\rho).$$

Inserting (A.22) gives that

$$① = ③ + \mathcal{E}_{p_\alpha}(f,f)O^{[\alpha]}(\varepsilon_\rho) + O^{[f,p]}(\epsilon\varepsilon_\rho) = \mathcal{E}_{p_\alpha}(f,f)(1 + O^{[\alpha]}(\varepsilon_\rho)) + O^{[f,p]}(\epsilon).$$

This finishes the proof since $\mathcal{E}^{(\alpha)}(f,f) = ①$. □

A.1.3 *Proofs in Section 3.3 (Theorems 3.3 and 3.4)*

Proof of Theorem 3.3. Suppose f is not a constant function because otherwise $E_N(f,f) = \mathcal{E}_{p_\alpha}(f,f) = 0$ and the theorem holds. By definition,

$$E_N(f,f) = \frac{1}{N^2}\sum_{i,j=1}^{N}\frac{\epsilon^{-\frac{d}{2}-1}}{m_2}(f(x_i) - f(x_j))^2 W_{ij}^{(\alpha)} = \frac{1}{N^2}\sum_{i\neq j,\, i,j=1}^{N} V_{ij}, \tag{A.24}$$

where

$$V_{ij} := \frac{1}{\epsilon m_2}(f(x_i) - f(x_j))^2 \hat{K}(x_i, x_j).$$

As (A.24) is a V-statistic, we study $\mathbb{E}E_N(f,f)$ and its variation away from $\mathbb{E}E_N(f,f)$, respectively.

• Calculation of $\mathbb{E}E_N(f,f)$. By definition,

$$\mathbb{E}E_N(f,f) = \frac{N-1}{N}\mathbb{E}V_{1,2} \quad \text{and} \quad \mathbb{E}V_{1,2} = \mathcal{E}^{(\alpha)}(f,f).$$

Applying Proposition 3.2 gives

$$\mathbb{E}V_{1,2} = \mathcal{E}_{p_\alpha}(f,f)(1 + O^{[\alpha]}(\varepsilon_\rho)) + O^{[f,p]}(\epsilon). \tag{A.25}$$

• Bound the deviation of $E_N(f,f) - \mathbb{E}E_N(f,f)$. We use the decoupling trick to bound the deviation of a V-statistic by that of an independent sum over $\frac{N}{2}$ terms. Specifically, define $\tilde{V}_{ij} = V_{ij} - \mathbb{E}V_{ij}$. For any $t > 0$, the Markov inequality gives us

$$\Pr\left[\frac{1}{N(N-1)}\sum_{i\neq j,i,j=1}^{N}\tilde{V}_{ij} > t\right] \leqslant e^{-st}\mathbb{E}\exp\left\{s\frac{1}{N(N-1)}\sum_{i\neq j,i,j=1}^{N}\tilde{V}_{ij}\right\} \tag{A.26}$$

where $s > 0$ will be determined later. By a direct expansion, and denote by $\mathcal{S}_N$ the permutation group, we have

$$
\begin{aligned}
&e^{-st}\mathbb{E}\exp\left\{s\frac{1}{N(N-1)}\sum_{i\neq j,i,j=1}^{N}\tilde{V}_{ij}\right\} = e^{-st}\mathbb{E}\exp\left\{s\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\frac{1}{N(N-1)}\sum_{i\neq j,i,j=1}^{N}\tilde{V}_{\sigma(i),\sigma(j)}\right\}\\
=&e^{-st}\mathbb{E}\exp\left\{s\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\frac{1}{N/2}\sum_{l=1}^{N/2}\tilde{V}_{\sigma(2l-1),\sigma(2l)}\right\} \leqslant e^{-st}\mathbb{E}\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\exp\left\{s\frac{1}{N/2}\sum_{l=1}^{N/2}\tilde{V}_{\sigma(2l-1),\sigma(2l)}\right\}\\
=&e^{-st}\mathbb{E}\exp\left\{s\frac{1}{N/2}\sum_{l=1}^{N/2}\tilde{V}_{2l-1,2l}\right\},
\end{aligned}
$$

where we apply the Jensen's inequality in the inequality. Then, as in the derivation of the classical Bernstein's inequality, one can bound the probability in (A.26) by

$$
\Pr[\cdots] \leqslant \exp\left\{-\frac{\frac{N}{2}t^2}{2\nu+\frac{2}{3}tL}\right\}, \quad \text{where} \quad \nu := \mathbb{E}\tilde{V}_{1,2}^2, \quad |\tilde{V}_{1,2}| \leqslant L. \tag{A.27}
$$

Below, we control ν and L. We first show that we can make $L = \Theta^{[f,p]}(\epsilon^{-\frac{d}{2}})$. Recall that

$$
V_{1,2} = \frac{\epsilon^{-\frac{d}{2}-1}}{m_2}k_0\left(\frac{\|x_1-x_2\|^2}{\epsilon\hat{\rho}(x_1)\hat{\rho}(x_2)}\right)\frac{(f(x_1)-f(x_2))^2}{\hat{\rho}(x_1)^\alpha\hat{\rho}(x_2)^\alpha}.
$$

By Assumption A.3(C2') for the kernel k_0,

$$
|V_{1,2}| \leqslant \frac{\epsilon^{-\frac{d}{2}-1}}{m_2}a_0 e^{-a\left(\frac{\|x_1-x_2\|^2}{\epsilon\hat{\rho}(x_1)\hat{\rho}(x_2)}\right)}\frac{(f(x_1)-f(x_2))^2}{\hat{\rho}(x_1)^\alpha\hat{\rho}(x_2)^\alpha}.
$$

By the assumption and (A.15), $\hat{\rho}(x) \leqslant 1.1\bar{\rho}(x) < 1.1\rho_{max}$ for any x. Then when $\|x_1-x_2\| \geqslant \delta_\epsilon := \sqrt{\epsilon 1.1^2\rho_{max}^2\frac{5+d/2}{a}\log\frac{1}{\epsilon}}$,

$$
e^{-a\left(\frac{\|x_1-x_2\|^2}{\epsilon\hat{\rho}(x_1)\hat{\rho}(x_2)}\right)} \leqslant e^{-(5+d/2)\log\frac{1}{\epsilon}} = \epsilon^{5+d/2}
$$

and then

$$
|V_{1,2}| \leqslant \frac{a_0\epsilon^4}{m_2}\frac{(2\|f\|_\infty)^2}{(0.9\rho_{min})^{2\alpha}} = O^{[f,p]}(\epsilon^4), \quad \text{when } \|x_1-x_2\| \geqslant \delta_\epsilon.
$$

Note that δ_ϵ is of order $\sqrt{\epsilon\log(\epsilon^{-1})}$, since $\epsilon = o(1)$, when ϵ is small enough such that $\delta_\epsilon < \delta_0$ in Lemma A.3,

$$
|f(x_1)-f(x_2)| \leqslant \|\nabla_{\mathcal{M}}f\|_\infty 1.1\|x_1-x_2\| =: L_f\|x_1-x_2\|, \quad \text{for all } x_2 \in B_{\delta_\epsilon}(x_1)\cap\mathcal{M}. \tag{A.28}
$$

Then, when $\|x_1 - x_2\| < \delta_\epsilon$,

$$|V_{1,2}| \leqslant \frac{\epsilon^{-\frac{d}{2}-1}}{m_2} a_0 e^{-a\left(\frac{\|x_1-x_2\|^2}{\epsilon\hat{\rho}(x_1)\hat{\rho}(x_2)}\right)} L_f^2 \frac{\|x_1 - x_2\|^2}{\hat{\rho}(x_1)\hat{\rho}(x_2)} \hat{\rho}(x_1)^{1-\alpha}\hat{\rho}(x_2)^{1-\alpha}$$
$$\leqslant \epsilon^{-\frac{d}{2}} \frac{a_0 L_f^2}{m_2} a_1' \| \max\{0.9^{1-\alpha}, 1.1^{1-\alpha}\}\bar{\rho}^{1-\alpha}\|_\infty^2,$$

where a_1' equals an absolute constant times $\frac{a_0}{a}$. Combining both cases, $|V_{1,2}| = O^{[f,p]}(\epsilon^{-\frac{d}{2}})$, and we denote

$$|\tilde{V}_{1,2}| \leqslant L = \Theta^{[f,p]}(\epsilon^{-\frac{d}{2}}).$$

We now compute the variance ν, and show that $\nu \leqslant \epsilon^{-d/2} V_f$ where $V_f = \Theta^{[f,p]}(1)$. By definition,

$$\mathbb{E}V_{1,2}^2 = \int_{\mathcal{M}}\int_{\mathcal{M}} \frac{\epsilon^{-d-2}}{m_2[k_0]^2} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{(f(x)-f(y))^4}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y) = \frac{\epsilon^{-d/2-1}}{\frac{m_2[k_0]^2}{m_2[k_0^2]}} \cdot ④, \tag{A.29}$$

where

$$④ := \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{(f(x)-f(y))^4}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

Let δ_ϵ be as above, and we separate the integral within and outside $\{\|x-y\| < \delta_\epsilon\}$ and make $④ = ④_1 + ④_2$. Specifically, we define

$$④_2 := \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} \mathbf{1}_{\{\|x-y\|\geqslant\delta_\epsilon\}} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{(f(x)-f(y))^4}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

By a direct bound, we have

$$④_2 \leqslant \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} a_0^2 e^{-2a\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)}\mathbf{1}_{\{\|x-y\|\geqslant\delta_\epsilon\}}(f(x)-f(y))^4\hat{\rho}(x)^{-2\alpha}\hat{\rho}(y)^{-2\alpha}p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y)$$
$$\leqslant \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]} a_0^2\epsilon^{10+d}\int_{\mathcal{M}}\int_{\mathcal{M}}(f(x)-f(y))^4\max\{0.9^{-2\alpha}, 1.1^{-2\alpha}\}^2\bar{\rho}(x)^{-2\alpha}\bar{\rho}(y)^{-2\alpha}p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y)$$
$$= O^{[f,p]}(\epsilon^{\frac{d}{2}+9}).$$

Define

$$④_1 := \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} \mathbf{1}_{\{\|x-y\|<\delta_\epsilon\}} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{(f(x)-f(y))^4}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

To control $④_1$, which involves an integration over the δ_ϵ ball, note that for $y \in B_{\delta_\epsilon}(x) \cap \mathcal{M}$, by Lemma A.3,

$$f(y) = f(x) + \nabla f(x)\cdot(y-x) + O^{[f]}(\|x-y\|^2),$$

and thus,

$$(f(y)-f(x))^4 = (\nabla f(x)\cdot(y-x))^4 + O^{[f]}(\|x-y\|^5) \leqslant (|\nabla f(x)|\|y-x\|)^4 + O^{[f]}(\|x-y\|^5)\,.$$

We then have

$$\begin{aligned}
④_1 &\leqslant \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} \mathbf{1}_{\{\|x-y\|<\delta_\epsilon\}} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{|\nabla f(x)|^4\|x-y\|^4 + O^{[f]}(\|x-y\|^5)}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} \\
&\quad p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y) \\
&= \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} \mathbf{1}_{\{\|x-y\|<\delta_\epsilon\}} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{|\nabla f(x)|^4\|x-y\|^4}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y) \\
&\quad + \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} \mathbf{1}_{\{\|x-y\|<\delta_\epsilon\}} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{O^{[f]}(\|x-y\|^5)}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y) =: ⑤ + ⑥.
\end{aligned}$$

We establish a lemma, which can be proved similarly as in deriving the limit of ① above, namely, by replacing $\hat{\rho}(x)$ with $\bar{\rho}(x)$ first and then putting back. The proof is postponed to Appendix A.3.

LEMMA A.1 Under (A.14) and (A.15), suppose k_0 satisfies Assumption 3.1, $f \in C^\infty(\mathcal{M})$ and $\alpha \in \mathbb{R}$, then

$$\epsilon^{-\frac{d}{2}}\int_{\mathcal{M}}\int_{\mathcal{M}} f(x)^2 k_0\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\hat{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\,\mathrm{d}V(x)\,\mathrm{d}V(y) = m_0[k_0]\int p^2 f^2\bar{\rho}^{d-2\alpha} + O^{[f,p]}(\epsilon,\varepsilon_\rho).$$

We bound ⑤ and ⑥, respectively, where ⑤ will dominate. By a direct bound, we have

$$|⑥| \leqslant O^{[f]}(\epsilon^{\frac{3}{2}})\frac{\epsilon^{-d/2}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)^{\frac{5}{2}}\hat{\rho}(x)^{\frac{5}{2}-2\alpha}\hat{\rho}(y)^{\frac{5}{2}-2\alpha}p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

Note that there is $b_6 > 0$, determined by a_0 and a, such that

$$k_0(r)^2 r^{5/2} \leqslant a_0^2 e^{-2ar} r^{5/2} \leqslant b_6 e^{-ar}, \quad \forall r > 0.$$

Thus,

$$k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)^{\frac{5}{2}} \leqslant b_6 e^{-a\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}},$$

and then

$$|⑥| \leqslant O^{[f]}(\epsilon^{\frac{3}{2}})\frac{\epsilon^{-d/2}}{m_2[k_0^2]}\int_{\mathcal{M}}\int_{\mathcal{M}} b_6 e^{-a\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}}\frac{p(x)p(y)}{\hat{\rho}(x)^{2\alpha-\frac{5}{2}}\hat{\rho}(y)^{2\alpha-\frac{5}{2}}}\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

Since the kernel $k_6(r) := b_6 e^{-ar}$ satisfies Assumption 3.1, applying Lemma A.1 with f replaced by 1 and α replaced by $2\alpha - \frac{5}{2}$ gives that

$$|⑥| = O^{[f]}(\epsilon^{\frac{3}{2}})\frac{1}{m_2[k_0^2]}\left(m_0[k_6]\int p^2\bar{\rho}^{d-2(2\alpha-\frac{5}{2})} + O^{[p]}(\epsilon,\varepsilon_\rho)\right) = O^{[f,p]}(\epsilon^{\frac{3}{2}}).$$

Write $G(x) = |\nabla f(x)|^2$. Clearly, since $f \in C^\infty(\mathcal{M})$, G is in $C^\infty(\mathcal{M})$. Define

$$k_5(r) := k_0^2(r)r^2\,.$$

Then, k_5 satisfies Assumption 3.1 and $m_0[k_5] = dm_2[k_0^2]$. As a result,

$$
\begin{aligned}
⑤ &\leqslant \frac{\epsilon^{-d/2-1}}{m_2[k_0^2]} \int_{\mathcal{M}} \int_{\mathcal{M}} G(x)^2 k_0^2 \left(\frac{\|x-y\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(y)} \right) \frac{\|x-y\|^4}{\hat{\rho}(x)^{2\alpha}\hat{\rho}(y)^{2\alpha}} p(x)p(y)\, \mathrm{d}V(x)\, \mathrm{d}V(y) \\
&= \frac{\epsilon}{m_2[k_0^2]} \epsilon^{-d/2} \int_{\mathcal{M}} \int_{\mathcal{M}} G(x)^2 k_0^2 \left(\frac{\|x-y\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(y)} \right) \left(\frac{\|x-y\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(y)} \right)^2 \frac{p(x)p(y)}{\hat{\rho}(x)^{2\alpha-2}\hat{\rho}(y)^{2\alpha-2}}\, \mathrm{d}V(x)\, \mathrm{d}V(y) \\
&= \frac{\epsilon}{m_2[k_0^2]} \epsilon^{-d/2} \int_{\mathcal{M}} \int_{\mathcal{M}} G(x)^2 k_5 \left(\frac{\|x-y\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(y)} \right) \frac{p(x)p(y)}{\hat{\rho}(x)^{2\alpha-2}\hat{\rho}(y)^{2\alpha-2}}\, \mathrm{d}V(x)\, \mathrm{d}V(y) \\
&= \frac{\epsilon}{m_2[k_0^2]} \cdot \left(m_0[k_5] \int p^2 G^2 \bar{\rho}^{d-2(2\alpha-2)} + O^{[f,p]}(\epsilon, \varepsilon_\rho) \right) \\
&= \epsilon d \int p^2 |\nabla f|^4 \bar{\rho}^{d-2(2\alpha-2)} + O^{[f,p]}(\epsilon^2, \epsilon\varepsilon_\rho) \\
&= \epsilon d \int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} + O^{[f,p]}(\epsilon^2, \epsilon\varepsilon_\rho)\,,
\end{aligned}
$$

where the third equality holds by applying Lemma A.1 with f replaced by G and α replaced by $2\alpha - 2$. Putting together, we have that

$$
\begin{aligned}
④ &\leqslant ⑤ + ⑥ + ④_2 \\
&\leqslant \epsilon d \int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} + O^{[f,p]}(\epsilon^2, \epsilon\varepsilon_\rho) + O^{[f,p]}(\epsilon^{\frac{3}{2}}) + O^{[f,p]}(\epsilon^{\frac{d}{2}+9}) \\
&= \epsilon d \int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} + O^{[f,p]}(\epsilon^{\frac{3}{2}}, \epsilon\varepsilon_\rho)\,.
\end{aligned}
$$

Plugging the above bounds back to (A.29), this gives that

$$
\begin{aligned}
\mathrm{Var}(V_{1,2}) \leqslant \mathbb{E}V_{1,2}^2 &\leqslant \frac{\epsilon^{-d/2-1}}{\frac{m_2[k_0]^2}{m_2[k_0^2]}} \left(\epsilon d \int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} + O^{[f,p]}(\epsilon^{\frac{3}{2}}, \epsilon\varepsilon_\rho) \right) \\
&= \frac{m_2[k_0^2]}{m_2[k_0]^2} \epsilon^{-d/2} \left(d \int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} + O^{[f,p]}(\epsilon^{\frac{1}{2}}, \varepsilon_\rho) \right).
\end{aligned}
$$

Since f is not constant valued, $\int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} > 0$, and by that $\epsilon^{1/2}, \varepsilon_\rho = o(1)$, we have that with sufficiently large N,

$$
\nu = \mathrm{Var}(V_{1,2}) \leqslant \Theta^{[1]}(\epsilon^{-d/2}) \int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} = \epsilon^{-d/2} V_f,
$$

where

$$
V_f := \Theta^{[1]} \left(\int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}} \right).
$$

Back to (A.27), by that $v \leqslant \epsilon^{-d/2} V_f$ and then

$$\exp\{-\frac{\frac{N}{2}t^2}{2v + \frac{2}{3}tL}\} \leqslant \exp\{-\frac{\frac{N}{2}t^2}{2\epsilon^{-d/2}V_f + \frac{2}{3}tL}\}, \tag{A.30}$$

we now control the r.h.s. Let $s = \Theta(1)$ to be determined, and we set

$$t = \sqrt{\epsilon^{-d/2} V_f \frac{s \log N}{N}}.$$

Since $L = \Theta^{[f,p]}(\epsilon^{-d/2})$, and by the condition that $\epsilon^{d/2} N = \Omega(\log N)$, $\frac{\log N}{N\epsilon^{d/2}} = o(1)$, and hence $t = o^{[f,p]}(1)$. With large N,

$$\frac{tL}{\epsilon^{-d/2} V_f} = \Theta^{[f,p]}\left(\sqrt{\frac{\epsilon^{-d/2}}{V_f}\frac{s \log N}{N}}\right) = o^{[f,p]}(1).$$

Therefore, when N is sufficiently large, we have $\frac{tL}{\epsilon^{-d/2}V_f} < 3$. Then (A.30) bounds the tail probability in (A.27) to be less than $\exp\{-\frac{Nt^2}{8\epsilon^{-d/2}V_f}\} = N^{-s/8}$. Let $s = 80$, and use the same argument to bound $\Pr[\frac{1}{N(N-1)}\sum_{i\neq j}\tilde{V}_{ij} < -t]$. We have that w.p. greater than $1 - 2N^{-10}$

$$\left|\frac{1}{N(N-1)}\sum_{i\neq j,i,j=1}^{N}\tilde{V}_{ij}\right| \leqslant \sqrt{\epsilon^{-d/2}V_f\frac{80\log N}{N}} = O\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}}\int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}}}\right). \tag{A.31}$$

Call the event set that (A.31) holds the event E_{Dir}.

At last,

$$E_N(f,f) = \left(1 - \frac{1}{N}\right)\cdot\frac{1}{N(N-1)}\sum_{i\neq j} V_{ij},$$

and with (A.25) we have shown that under good event E_{Dir},

$$\frac{1}{N(N-1)}\sum_{i\neq j} V_{ij} = \mathcal{E}_{p_\alpha}(f,f)(1 + O^{[\alpha]}(\varepsilon_\rho)) + O^{[f,p]}(\epsilon) + O\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}}\int |\nabla f|^4 p^{1+\frac{4(\alpha-1)}{d}}}\right),$$

which is $O^{[f,p]}(1)$, thus $E_N(f,f)$ differs from it by $O^{[f,p]}(\frac{1}{N}) = o^{[f,p]}(\sqrt{\frac{1}{N}})$, which is dominated by the variance error. This finishes the proof of the theorem. □

Proof of Theorem 3.4. The proof uses same techniques as that of Theorem 3.3 and is simplified due to fixed bandwidth kernel. We track the influence of β and ε_p which differs from the proof of Theorem 3.3. Inherit the notations in the proof of Theorem 3.3, the random variable V_{ij} is now

$$V_{ij} := \frac{\epsilon^{-\frac{d}{2}}}{\epsilon m_2}(f(x_i) - f(x_j))^2 k_0\left(\frac{\|x_i - x_j\|^2}{\epsilon}\right)\frac{1}{\hat{p}(x_i)^\beta \hat{p}(x_j)^\beta}.$$

Similarly as before, we define

$$① := \mathbb{E}V_{ij} = \frac{1}{\epsilon m_2}\int\int (f(x)-f(y))^2\epsilon^{-\frac{d}{2}}k_0\left(\frac{\|x-y\|^2}{\epsilon}\right)\frac{p(x)p(y)}{\hat{p}(x)^\beta\hat{p}(y)^\beta}\,\mathrm{d}V(x)\,\mathrm{d}V(y),$$

and

$$② := \frac{1}{\epsilon m_2}\int\int (f(x)-f(y))^2\epsilon^{-\frac{d}{2}}k_0\left(\frac{\|x-y\|^2}{\epsilon}\right)\frac{p(x)^{1-\beta}p(y)}{\hat{p}(y)^\beta}\,\mathrm{d}V(x)\,\mathrm{d}V(y),$$

$$③ := \frac{1}{\epsilon m_2}\int\int (f(x)-f(y))^2\epsilon^{-\frac{d}{2}}k_0\left(\frac{\|x-y\|^2}{\epsilon}\right)p(x)^{1-\beta}p(y)^{1-\beta}\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

Define $q := p^{1-\beta}$ which is a non-negative power of p because $\beta \leqslant 1$, thus $q \in C^\infty(\mathcal{M})$. We then have

$$\begin{aligned}
③ &= \frac{1}{\epsilon m_2}\int\int (f(x)-f(y))^2\epsilon^{-\frac{d}{2}}k_0\left(\frac{\|x-y\|^2}{\epsilon}\right)q(x)q(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y)\\
&= \frac{2}{\epsilon m_2}\int\int (f(x)^2-f(x)f(y))\epsilon^{-\frac{d}{2}}k_0\left(\frac{\|x-y\|^2}{\epsilon}\right)q(x)q(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y)\\
&= \frac{2}{\epsilon m_2}\left(\int (qf^2)G_\epsilon(q)-\int (qf)G_\epsilon(fq)\right).
\end{aligned}$$

By Lemma A.5,

$$G_\epsilon(q) = m_0 q + \epsilon\frac{m_2}{2}(wq+\Delta q) + O^{[q^{(\leqslant 4)}]}(\epsilon^2),$$

$$G_\epsilon(fq) = m_0 fq + \epsilon\frac{m_2}{2}(wfq+\Delta(fq)) + O^{[(fq)^{(\leqslant 4)}]}(\epsilon^2),$$

and thus,

$$\begin{aligned}
③ &= \int qf(f\Delta q-\Delta(fq)) + O^{[f,q]}(\epsilon) = -\int q^2 f(\Delta f + 2\frac{\nabla q}{q}\cdot\nabla f) + O^{[f,q]}(\epsilon)\\
&= -\langle f, \Delta_{p_\beta} f\rangle_{p_\beta} + O^{[f,q]}(\epsilon) \quad (p_\beta := q^2 = p^{2-2\beta})\\
&= \mathcal{E}_{p_\beta}(f,f) + O^{[f,q]}(\epsilon).
\end{aligned} \tag{A.32}$$

To bound $|③ - ②|$: because

$$③ - ② = \frac{1}{\epsilon m_2}\int\int (f(x)-f(y))^2\epsilon^{-\frac{d}{2}}k_0\left(\frac{\|x-y\|^2}{\epsilon}\right)p(x)^{1-\beta}p(y)(p(y)^{-\beta}-\hat{p}(y)^{-\beta})\,\mathrm{d}V(x)\,\mathrm{d}V(y),$$

by the positivity of p and k_0,

$$|③ - ②| \leqslant \frac{1}{\epsilon m_2}\int\int (f(x)-f(y))^2\epsilon^{-\frac{d}{2}}k_0\left(\frac{\|x-y\|^2}{\epsilon}\right)p(x)^{1-\beta}p(y)|p(y)^{-\beta}-\hat{p}(y)^{-\beta}|\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

Note that for any y, by the mean value theorem and that $|\hat{p}(y)-p(y)|/p(y) < \varepsilon_p < 0.1$, $\exists \xi$ between $\hat{p}(y)$ and $p(y)$ and thus $\xi^{-\beta-1} \leqslant \max\{1.1^{-\beta-1}, 0.9^{-\beta-1}\} p(y)^{-\beta-1}$, and then we have

$$|\hat{p}(y)^{-\beta} - p(y)^{-\beta}| = |\beta|\xi^{-\beta-1}|\hat{p}(y)-p(y)| \leqslant |\beta| \max\{1.1^{-\beta-1}, 0.9^{-\beta-1}\} p(y)^{-\beta-1}|\hat{p}(y)-p(y)|$$
$$\leqslant c_\beta |\beta| p(y)^{-\beta}\varepsilon_p, \quad c_\beta := \max\{1.1^{-\beta-1}, 0.9^{-\beta-1}\}. \tag{A.33}$$

Thus,

$$|③-②| \leqslant |\beta| c_\beta \varepsilon_p \frac{1}{\epsilon m_2} \int\int (f(x)-f(y))^2 \epsilon^{-\frac{d}{2}} k_0\left(\frac{\|x-y\|^2}{\epsilon}\right) p(x)^{1-\beta} p(y)^{1-\beta}\, \mathrm{d}V(x)\, \mathrm{d}V(y) = |\beta| c_\beta \varepsilon_p ③.$$

Combined with (A.32), we have that

$$② = ③(1 + O^{[1]}(\beta c_\beta \varepsilon_p)) = (\mathcal{E}_{p_\beta}(f,f) + O^{[f,q]}(\epsilon))(1 + O^{[1]}(\beta c_\beta \varepsilon_p))$$
$$= \mathcal{E}_{p_\beta}(f,f)(1 + O^{[1]}(c_\beta \beta \varepsilon_p)) + O^{[f,q]}(\epsilon). \tag{A.34}$$

To bound $|② - ②|$: by (A.33)

$$|② - ①| \leqslant \frac{1}{\epsilon m_2} \int\int (f(x)-f(y))^2 \epsilon^{-\frac{d}{2}} k_0\left(\frac{\|x-y\|^2}{\epsilon}\right) p(y)\hat{p}(y)^{-\beta} p(x) |p(x)^{-\beta} - \hat{p}(x)^{-\beta}|\, \mathrm{d}V(x)\, \mathrm{d}V(y)$$
$$\leqslant |\beta| c_\beta \varepsilon_p \frac{1}{\epsilon m_2} \int\int (f(x)-f(y))^2 \epsilon^{-\frac{d}{2}} k_0\left(\frac{\|x-y\|^2}{\epsilon}\right) p(y)\hat{p}(y)^{-\beta} p(x)^{1-\beta}\, \mathrm{d}V(x)\, \mathrm{d}V(y) = |\beta| c_\beta \varepsilon_p ②.$$

Then, together with (A.34), we have that

$$\mathbb{E}V_{ij} = ① = ②(1 + O^{[1]}(\beta c_\beta \varepsilon_p)) = (\mathcal{E}_{p_\beta}(f,f)(1 + O^{[1]}(\beta c_\beta \varepsilon_p)) + O^{[f,q]}(\epsilon))(1 + O^{[1]}(\beta c_\beta \varepsilon_p))$$
$$= \mathcal{E}_{p_\beta}(f,f)(1 + O^{[1]}(\beta c_\beta \varepsilon_p)) + O^{[f,q]}(\epsilon), \quad \text{and } O^{[f,q]}(\epsilon) = O^{[f,p,\beta]}(\epsilon) \text{ by definition.}$$

In the special case where $\beta = 0$, we have $q = p$, and ① = ② = ③. Then (A.32) gives that

$$\mathbb{E}V_{ij} = ① = \mathcal{E}_{p^2}(f,f) + O^{[f,p]}(\epsilon).$$

The boundedness of $|V_{ij}|$ follows by the same argument of truncation on the δ_ϵ ball as in the proof of Theorem 3.3, which gives

$$|V_{ij}| \leqslant L = \Theta^{[f,p,\beta]}(\epsilon^{-d/2}).$$

The variance

$$\mathrm{Var}(V_{ij}) \leqslant \mathbb{E}V_{ij}^2 = \int\int \frac{\epsilon^{-d-2}}{m_2^2} (f(x)-f(y))^4 k_0^2\left(\frac{\|x_i - x_j\|^2}{\epsilon}\right) \frac{p(x)p(y)}{\hat{p}(x)^{2\beta}\hat{p}(y)^{2\beta}}\, \mathrm{d}V(x)\, \mathrm{d}V(y),$$

and, similarly as in the proof of Theorem 3.3, we can show that

$$\mathbb{E}V_{ij}^2 \leqslant \epsilon^{-d/2} V_f, \quad V_f = \Theta^{[1]}\Big(\int |\nabla f|^4 p^{2-4\beta}\Big).$$

Thus, by the V-statistics decoupling argument, w.p. higher than $1 - 2N^{-10}$, the variance error is

$$O^{[1]}\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}} \int |\nabla f|^4 p^{2-4\beta}}\right).$$

The normalization $\frac{1}{N^2}$ and $\frac{1}{N(N-1)}$ in the V-statistics incurs higher order error, and putting together bias and variance error proves the theorem. □

A.1.4 *Proofs in Section 3.4 (Theorems 3.5, 3.6, 3.7 and 3.8)*

Proof of Theorem 3.5. Define $\tilde{m} := \frac{m_2}{2m_0}$, and rewrite (3.10) as

$$L_{rw}^{(\alpha)} f(x) = \frac{1}{\epsilon \tilde{m} \hat{\rho}(x)^2} \left(\frac{\frac{1}{N}\sum_{j=1}^N F_j(x)}{\frac{1}{N}\sum_{j=1}^N G_j(x)} - f(x) \right), \tag{A.35}$$

where

$$F_j := \epsilon^{-d/2} k_0 \left(\frac{\|x - x_j\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(x_j)} \right) \frac{f(x_j)}{\hat{\rho}(x_j)^\alpha}, \quad G_j := \epsilon^{-d/2} k_0 \left(\frac{\|x - x_j\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(x_j)} \right) \frac{1}{\hat{\rho}(x_j)^\alpha}. \tag{A.36}$$

Note that since $x_j \sim p$ i.i.d., $\{F_j\}_{j=1}^N$ are i.i.d. rvs, so are $\{G_j\}_{j=1}^N$, while F_js and G_js are dependent. The expectations are

$$\mathbb{E}F(x) = \epsilon^{-d/2} \int_{\mathcal{M}} k_0 \left(\frac{\|x - y\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(y)} \right) \frac{f(y)}{\hat{\rho}(y)^\alpha} p(y)\, \mathrm{d}V(y),$$

$$\mathbb{E}G(x) = \epsilon^{-d/2} \int_{\mathcal{M}} k_0 \left(\frac{\|x - y\|^2}{\epsilon \hat{\rho}(x)\hat{\rho}(y)} \right) \frac{1}{\hat{\rho}(y)^\alpha} p(y)\, \mathrm{d}V(y).$$

Following the strategy in [5,42] to analyze the bias and variance errors respectively, we will show that

- The bias:

$$\frac{1}{\epsilon \tilde{m} \hat{\rho}(x)^2} \left(\frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} - f(x) \right) \stackrel{?}{=} \mathcal{L}^{(\alpha)} f(x) + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon} \right) \tag{A.37}$$

- The variance:

$$\frac{1}{\epsilon \tilde{m} \hat{\rho}(x)^2} \left(\frac{\frac{1}{N}\sum_{j=1}^N F_j(x)}{\frac{1}{N}\sum_{j=1}^N G_j(x)} - \frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} \right) \stackrel{?}{=} O^{[1]}\left(\|\nabla f\|_\infty p(x)^{1/d} \sqrt{\frac{\log N}{N\epsilon^{d/2+1}}} \right) \tag{A.38}$$

Proof of (A.37): by the definition of $G_\epsilon^{(\rho)}$ in (3.6),

$$\mathbb{E}F(x) = \hat{\rho}(x)^{d/2} G_{\epsilon\hat{\rho}(x)}^{(\hat{\rho})}\left(\frac{fp}{\hat{\rho}^\alpha}\right)(x),$$

yet $\frac{fp}{\hat{\rho}^\alpha}$ is not C^4. In order to apply Lemmas 3.1 and 3.2, we compare to replacing it with $\frac{fp}{\rho^\alpha}$:

Lemma A.2 Suppose notation and condition are the same as those in Lemmas 3.1 and 3.2. When ϵ is sufficiently small, for some constant c_ρ'' determined by $(\mathcal{M}, k_0, \rho_{max}, \rho_{min})$,

$$\sup_{x\in\mathcal{M}} |G_\epsilon^{(\tilde{\rho})}\left(\frac{f}{\tilde{\rho}^\alpha}\right)(x) - G_\epsilon^{(\tilde{\rho})}\left(\frac{f}{\rho^\alpha}\right)(x)| \leqslant c_\rho'' \|f\|_\infty \varepsilon = O^{[f,\rho]}(\varepsilon).$$

The proof of Lemma A.2 is postponed to Appendix A.3. Applying Lemmas 3.1, 3.2 and A.2, where $\rho = \bar{\rho}$, $\tilde{\rho} = \hat{\rho}$ and 'f' in the lemmas is replaced with fp, we have

$$\mathbb{E}F(x) = \hat{\rho}(x)^{d/2}\left(G^{(\hat{\rho})}_{\epsilon\hat{\rho}(x)}(\frac{fp}{\bar{\rho}^{\alpha}})(x) + O^{[f,p]}(\varepsilon_{\rho})\right) \quad \text{(by Lemma A.2)}$$

$$= \hat{\rho}(x)^{d/2}\left(G^{(\bar{\rho})}_{\epsilon\hat{\rho}(x)}(\frac{fp}{\bar{\rho}^{\alpha}})(x) + O^{[f,p]}(\varepsilon_{\rho}) + O^{[f,p]}(\varepsilon_{\rho})\right) \quad \text{(by Lemma 3.2)}$$

$$= \hat{\rho}(x)^{d/2}\left(m_0 fp\bar{\rho}^{\frac{d}{2}-\alpha}(x) + \epsilon\hat{\rho}\frac{m_2}{2}(\omega fp\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha}))(x) + \hat{\rho}^2(x)O^{[f,p]}(\epsilon^2) + O^{[f,p]}(\varepsilon_{\rho})\right),$$

where the residual terms in big-O are bounded uniformly for all $x \in \mathcal{M}$. Below, we omit the variable x in the notation when there is no confusion.

Because $0.9\bar{\rho} \leqslant \hat{\rho} \leqslant 1.1\bar{\rho}$ for all $x \in \mathcal{M}$, for any power $\gamma \in \mathbb{R}$, $\hat{\rho}^{\gamma}$ lies between $\bar{\rho}^{\gamma}1.1^{\gamma}$ and $\bar{\rho}^{\gamma}0.9^{\gamma}$ (the order depending on the sign of γ), and then uniformly bounded between $(0.9\rho_{min})^{\gamma}$ and $(1.1\rho_{max})^{\gamma}$, both of which are $\Theta^{[p]}(1)$ constants. We can also bound $|\hat{\rho}^{\gamma} - \bar{\rho}^{\gamma}|$ as in (A.17), and in summary we have

$$\sup_{x\in\mathcal{M}}|\hat{\rho}^{\gamma}(x)| = O^{[p]}(1), \quad \sup_{x\in\mathcal{M}}|\hat{\rho}^{\gamma}(x) - \bar{\rho}^{\gamma}(x)| \leqslant O^{[p]}(\varepsilon_{\rho}). \tag{A.39}$$

We proceed with these bounds. We have shown, omitting the evaluation of x in the notation, that

$$\mathbb{E}F = \hat{\rho}^{d/2}\left(m_0 fp\bar{\rho}^{\frac{d}{2}-\alpha} + \epsilon\hat{\rho}\frac{m_2}{2}(\omega fp\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha}))\right) + O^{[f,p]}(\epsilon^2, \varepsilon_{\rho}),$$

and by (A.39) with $\gamma = \frac{d}{2}$ and $\frac{d}{2}+1$,

$$\mathbb{E}F = \bar{\rho}^{d/2}\left(m_0 fp\bar{\rho}^{\frac{d}{2}-\alpha} + \epsilon\bar{\rho}\frac{m_2}{2}(\omega fp\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha}))\right) + O^{[f,p]}(\epsilon^2, \varepsilon_{\rho}). \tag{A.40}$$

Similarly, we have

$$\mathbb{E}G(x) = \hat{\rho}(x)^{d/2}G^{(\hat{\rho})}_{\epsilon\hat{\rho}(x)}(\frac{p}{\hat{\rho}^{\alpha}})(x)$$

$$= \bar{\rho}^{d/2}\left(m_0 p\bar{\rho}^{\frac{d}{2}-\alpha} + \epsilon\bar{\rho}\frac{m_2}{2}(\omega p\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(p\bar{\rho}^{1+\frac{d}{2}-\alpha}))\right) + O^{[p]}(\epsilon^2, \varepsilon_{\rho}), \tag{A.41}$$

and expanding $\mathbb{E}G$ to the $O(\epsilon)$ term only gives that

$$\mathbb{E}G(x) = m_0 p\bar{\rho}^{d-\alpha}(x) + r_G(x), \quad \|r_G\|_{\infty} = O^{[p]}(\epsilon, \varepsilon_{\rho}).$$

Because $\inf_{x\in\mathcal{M}} m_0 p\bar{\rho}^{d-\alpha}(x)$ is a strictly positive constant depending on p, and ϵ and ε_{ρ} are $o(1)$, thus when N is large and the threshold depends on $(\mathcal{M}, p, \alpha)$, $\|r_G\|_{\infty} < \inf_{x\in\mathcal{M}} m_0 p\bar{\rho}^{d-\alpha}(x)$. Then for any $x \in \mathcal{M}$, $|r_G(x)| < m_0 p\bar{\rho}^{d-\alpha}(x)$, and we have

$$\frac{1}{\mathbb{E}G(x)} = \frac{1}{m_0 p\bar{\rho}^{d-\alpha}(x)}\sum_{l=0}^{\infty}\left(-\frac{r_G(x)}{m_0 p\bar{\rho}^{d-\alpha}(x)}\right)^{l} = \frac{1}{m_0 p\bar{\rho}^{d-\alpha}(x)} + O^{[p]}(\epsilon, \varepsilon_{\rho}). \tag{A.42}$$

Meanwhile,

$$\mathbb{E}F(x) - f(x)\mathbb{E}G(x) = \epsilon\frac{m_2}{2}\bar{\rho}^{d/2+1}\left[\Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha}) - f\Delta(p\bar{\rho}^{1+\frac{d}{2}-\alpha})\right] + O^{[f,p]}(\epsilon^2, \varepsilon_{\rho}).$$

Note that the quantity in the square brackets

$$[\cdots] = p\bar{\rho}^{1+\frac{d}{2}-\alpha}\left(\Delta f + 2\frac{\nabla p}{p}\cdot\nabla f + (2+d-2\alpha)\frac{\nabla\bar{\rho}}{\bar{\rho}}\cdot\nabla f\right),$$

and then by the definition of $\mathcal{L}_\rho^{(\alpha)}$ in (3.1), and that $\mathcal{L}_{\bar{\rho}}^{(\alpha)} = \mathcal{L}^{(\alpha)}$, we have

$$\mathbb{E}F(x) - f(x)\mathbb{E}G(x) = \epsilon\frac{m_2}{2}p\bar{\rho}^{d+2-\alpha}\mathcal{L}^{(\alpha)}f + O^{[f,p]}(\epsilon^2, \varepsilon_\rho). \tag{A.43}$$

Putting together, we have

$$\begin{aligned}\frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} - f(x) &= \frac{\mathbb{E}F(x) - f(x)\mathbb{E}G(x)}{\mathbb{E}G(x)}\\ &= \left(\epsilon\frac{m_2}{2}p\bar{\rho}^{d+2-\alpha}\mathcal{L}^{(\alpha)}f + O^{[f,p]}(\epsilon^2, \varepsilon_\rho)\right)\left(\frac{1}{m_0 p\bar{\rho}^{d-\alpha}(x)} + O^{[p]}(\epsilon, \varepsilon_\rho)\right)\\ &= \epsilon\frac{m_2}{2m_0}\bar{\rho}^2\mathcal{L}^{(\alpha)}f(x) + O^{[f,p]}\left(\epsilon^2, \varepsilon_\rho\right) + O^{[f,p]}\left(\epsilon^2, \epsilon\varepsilon_\rho\right)\\ &= \epsilon\tilde{m}\bar{\rho}^2\mathcal{L}^{(\alpha)}f(x) + O^{[f,p]}\left(\epsilon^2, \varepsilon_\rho\right),\end{aligned} \tag{A.44}$$

and then

$$\frac{1}{\epsilon\tilde{m}\bar{\rho}^2}\left(\frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} - f(x)\right) = \mathcal{L}^{(\alpha)}f(x) + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right) = O^{[f,p]}(1).$$

Finally, by (A.39) with $\gamma = -2$,

$$\left|\frac{1}{\epsilon\tilde{m}}\left(\frac{1}{\hat{\rho}(x)^2} - \frac{1}{\bar{\rho}(x)^2}\right)\left(\frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} - f(x)\right)\right| = O^{[p]}(\varepsilon_\rho)\frac{1}{\epsilon\tilde{m}\bar{\rho}^2}\left|\frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} - f(x)\right| = O^{[p]}(\varepsilon_\rho)O^{[f,p]}(1),$$

and, using triangle inequality, this together with (A.44) gives the following

$$\frac{1}{\epsilon\tilde{m}\hat{\rho}^2(x)}\left(\frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} - f(x)\right) = O^{[f,p]}(\varepsilon_\rho) + \mathcal{L}^{(\alpha)}f(x) + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right),$$

where the constants in $O^{[f,p]}(\cdot)$ are uniform for $x \in \mathcal{M}$. This proves (A.37).

Proof of (A.38): by definition, we have

$$\frac{\frac{1}{N}\sum_{j=1}^N F_j(x)}{\frac{1}{N}\sum_{j=1}^N G_j(x)} - \frac{\mathbb{E}F(x)}{\mathbb{E}G(x)} = \frac{\frac{1}{N}\sum_{j=1}^N (F_j(x)\mathbb{E}G(x) - G_j(x)\mathbb{E}F(x))}{\mathbb{E}G(x)\cdot\frac{1}{N}\sum_{j=1}^N G_j(x)} =: \frac{\frac{1}{N}\sum_{j=1}^N Y_j(x)}{\mathbb{E}G(x)\cdot\frac{1}{N}\sum_{j=1}^N G_j(x)}, \tag{A.45}$$

where $Y_j(x) = F_j(x)\mathbb{E}G(x) - G_j(x)\mathbb{E}F(x)$. Below we omit x, which is fixed, in the notation. We consider the concentration of $\frac{1}{N}\sum_{j=1}^N G_j$ and $\frac{1}{N}\sum_{j=1}^N Y_j$, respectively.

We have shown in (A.41) that

$$\mathbb{E}G = m_0 p\bar{\rho}^{d-\alpha} + O^{[p]}(\epsilon, \varepsilon_\rho) = \Theta^{[1]}(m_0 p\bar{\rho}^{d-\alpha}),$$

and by the boundedness of k_0 and uniform boundedness of $\hat{\rho}$, $|G_j| \leqslant L_G = \Theta^{[p]}(\epsilon^{-d/2})$. Using the same argument to analyze the operator $G^{(\hat{\rho})}_{\epsilon\hat{\rho}(x)}$ as that in Lemmas 3.1, 3.2 and A.2, the variance

$$\begin{aligned}
\mathrm{Var}(G_j) \leqslant \mathbb{E}G_j^2 \quad &= \int_{\mathcal{M}} \epsilon^{-d} k_0^2\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{p(y)}{\hat{\rho}(y)^{2\alpha}}\,\mathrm{d}V(y)\\
&= \epsilon^{-d/2}\hat{\rho}(x)^{d/2} G^{(\hat{\rho})}_{\epsilon\hat{\rho}(x)}[k_0^2]\left(\frac{p}{\hat{\rho}^{2\alpha}}\right)(x)\\
&= \epsilon^{-d/2}\left\{m_0[k_0^2]p\bar{\rho}^{d-2\alpha} + \epsilon\frac{m_2[k_0^2]}{2}\bar{\rho}^{1+\frac{d}{2}}(\omega p\bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(p\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[p]}\left(\epsilon^2,\, \varepsilon_\rho\right)\right\}\\
&= \epsilon^{-d/2}\left\{m_0[k_0^2]p\bar{\rho}^{d-2\alpha} + O^{[p]}\left(\epsilon,\, \varepsilon_\rho\right)\right\} \qquad \text{(A.46)}\\
&\leqslant \bar{\nu}_G(x) = \Theta(\epsilon^{-d/2}m_0[k_0^2]p\bar{\rho}^{d-2\alpha}(x)) \ \text{ with large } N, \text{ since } p\bar{\rho}^{d-2\alpha}(x) \geqslant \Theta^{[p]}(1) > 0 \text{ for all } x.
\end{aligned}$$

By that $\frac{\log N}{N\epsilon^{d/2}} = o(1)$, when N is large, $\sqrt{40\frac{\log N}{N}\bar{\nu}_G(x)} < \frac{3\bar{\nu}_G(x)}{L_G}$ for all x, and then w.p. higher than $> 1 - 2N^{-10}$,

$$\left|\frac{1}{N}\sum_{j=1}^N G_j - \mathbb{E}G\right| \leqslant \sqrt{\frac{40\log N}{N}\bar{\nu}_G} = O^{[p]}\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}}}\right),$$

which we define as the good event E_2. The threshold of large N needed for $\mathrm{Var}(G_j) \leqslant \bar{\nu}_G(x)$ and for applying the sub-Gaussian tail in Bernstein inequality depends on $(\mathcal{M}, p, \alpha)$. Under E_2,

$$\frac{1}{N}\sum_{j=1}^N G_j = m_0 p\bar{\rho}^{d-\alpha} + O^{[p]}(\epsilon,\, \varepsilon_\rho) + O^{[p]}\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}}}\right) = \Theta\left(m_0 p\bar{\rho}^{d-\alpha}\right),$$

and then

$$\mathbb{E}G\cdot\frac{1}{N}\sum_{j=1}^N G_j = \Theta((m_0 p\bar{\rho}^{d-\alpha})^2)\,. \qquad \text{(A.47)}$$

To analyze the independent sum $\frac{1}{N}\sum_{j=1}^N Y_j$, first note that $\mathbb{E}Y_j = 0$. For boundedness of Y_j, because

$$\mathbb{E}G = O^{[p]}(1), \quad \mathbb{E}F = O^{[f,p]}(1), \quad |G_j| \leqslant L_G = \Theta^{[p]}(\epsilon^{-d/2}), \quad |F_j| \leqslant L_F = \Theta^{[f,p]}(\epsilon^{-d/2}), \quad \text{(A.48)}$$

we have that

$$|Y_j| \leqslant |F_j||\mathbb{E}G| + |G_j||\mathbb{E}F| \leqslant L_Y = \Theta^{[f,p]}(\epsilon^{-d/2}).$$

For the variance of Y_j,

$$\mathbb{E}Y_j^2 = \mathbb{E}(F_j\mathbb{E}G - G_j\mathbb{E}F)^2 = \mathbb{E}F^2(\mathbb{E}G)^2 + \mathbb{E}G^2(\mathbb{E}F)^2 - 2\mathbb{E}(FG)\mathbb{E}F\mathbb{E}G,$$

and we have

$$\begin{aligned}
\mathbb{E}F^2 &= \int_{\mathcal{M}} \epsilon^{-d} k_0^2 \left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{f(y)^2 p(y)}{\hat{\rho}(y)^{2\alpha}}\,\mathrm{d}V(y) \\
&= \epsilon^{-d/2}\hat{\rho}(x)^{d/2} G^{(\hat{\rho})}_{\epsilon\hat{\rho}(x)}[k_0^2](\frac{f^2 p}{\hat{\rho}^{2\alpha}})(x) \\
&= \epsilon^{-d/2}\left\{ m_0[k_0^2] f^2 p\bar{\rho}^{d-2\alpha} + \epsilon\frac{m_2[k_0^2]}{2}\bar{\rho}^{1+\frac{d}{2}}(\omega f^2 p\bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(f^2 p\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[f,p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\}, \\
\mathbb{E}[FG] &= \int_{\mathcal{M}} \epsilon^{-d} k_0^2 \left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{f(y) p(y)}{\hat{\rho}(y)^{2\alpha}}\,\mathrm{d}V(y) \\
&= \epsilon^{-d/2}\hat{\rho}(x)^{d/2} G^{(\hat{\rho})}_{\epsilon\hat{\rho}(x)}[k_0^2](\frac{fp}{\hat{\rho}^{2\alpha}})(x) \\
&= \epsilon^{-d/2}\left\{ m_0[k_0^2] fp\bar{\rho}^{d-2\alpha} + \epsilon\frac{m_2[k_0^2]}{2}\bar{\rho}^{1+\frac{d}{2}}(\omega fp\bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(fp\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[f,p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\}.
\end{aligned}$$

Together with (A.46), (A.40) and (A.41), and defining $m_0' := m_0[k_0^2]$ and $m_2' := m_2[k_0^2]$, we have that

$$\begin{aligned}
\mathbb{E}Y_j^2 &= \epsilon^{-d/2}\left\{ m_0' f^2 p\bar{\rho}^{d-2\alpha} + \epsilon\frac{m_2'}{2}\bar{\rho}^{1+\frac{d}{2}}(\omega f^2 p\bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(f^2 p\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[f,p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\} \\
&\quad\cdot\left\{ m_0 p\bar{\rho}^{d-\alpha} + \epsilon\frac{m_2}{2}\bar{\rho}^{d/2+1}(\omega p\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(p\bar{\rho}^{1+\frac{d}{2}-\alpha})) + O^{[p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\}^2 \\
&\quad+ \epsilon^{-d/2}\left\{ m_0' p\bar{\rho}^{d-2\alpha} + \epsilon\frac{m_2'}{2}\bar{\rho}^{1+\frac{d}{2}}(\omega p\bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(p\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\} \\
&\quad\cdot\left\{ m_0 fp\bar{\rho}^{d-\alpha} + \epsilon\frac{m_2}{2}\bar{\rho}^{d/2+1}(\omega fp\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha})) + O^{[f,p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\}^2 \\
&\quad- 2\epsilon^{-d/2}\left\{ m_0' fp\bar{\rho}^{d-2\alpha} + \epsilon\frac{m_2'}{2}\bar{\rho}^{1+\frac{d}{2}}(\omega fp\bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(fp\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[f,p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\} \\
&\quad\cdot\left\{ m_0 fp\bar{\rho}^{d-\alpha} + \epsilon\frac{m_2}{2}\bar{\rho}^{d/2+1}(\omega fp\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha})) + O^{[f,p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\} \\
&\quad\cdot\left\{ m_0 p\bar{\rho}^{d-\alpha} + \epsilon\frac{m_2}{2}\bar{\rho}^{d/2+1}(\omega p\bar{\rho}^{1+\frac{d}{2}-\alpha} + \Delta(p\bar{\rho}^{1+\frac{d}{2}-\alpha})) + O^{[p]}\left(\epsilon^2,\ \varepsilon_\rho\right)\right\} \\
&= \epsilon^{-d/2}\left\{ \epsilon\frac{m_2' m_0^2}{2} p^2\bar{\rho}^{\frac{5}{2}d-2\alpha+1}(\Delta(f^2 p\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + \epsilon m_2 m_0' m_0 p^2\bar{\rho}^{\frac{5}{2}d-3\alpha+1}(f^2\Delta(p\bar{\rho}^{1+\frac{d}{2}-\alpha}))\right. \\
&\quad+ \epsilon\frac{m_2' m_0^2}{2} f^2 p^2\bar{\rho}^{\frac{5}{2}d-2\alpha+1}(\Delta(p\bar{\rho}^{1+\frac{d}{2}-2\alpha})) + \epsilon m_2 m_0' m_0 p^2\bar{\rho}^{\frac{5}{2}d-3\alpha+1}(f\Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha})) \\
&\quad\left.- 2\epsilon\frac{m_2' m_0^2}{2} fp^2\bar{\rho}^{\frac{5}{2}d-2\alpha+1}(\Delta(fp\bar{\rho}^{1+\frac{d}{2}-2\alpha})) - \epsilon m_2 m_0' m_0 p^2\bar{\rho}^{\frac{5}{2}d-3\alpha+1}(f\Delta(fp\bar{\rho}^{1+\frac{d}{2}-\alpha}))\right.
\end{aligned}$$

$$-\epsilon m_2 m_0' m_0 p^2 \bar{\rho}^{\frac{5}{2}d-3\alpha+1}(f^2\Delta(p\bar{\rho}^{1+\frac{d}{2}-\alpha})) + O^{[f,p]}\left(\epsilon^2, \varepsilon_\rho\right)\Big\}$$

$$= \epsilon^{-d/2}\left\{\epsilon\frac{m_2' m_0^2}{2}p^2\bar{\rho}^{\frac{5}{2}d-2\alpha+1}\left[\Delta(f^2 p\bar{\rho}^{1+\frac{d}{2}-2\alpha}) + f^2\Delta(p\bar{\rho}^{1+\frac{d}{2}-2\alpha}) - 2f\Delta(fp\bar{\rho}^{1+\frac{d}{2}-2\alpha})\right]\right.$$

$$\left. + O^{[f,p]}\left(\epsilon^2, \varepsilon_\rho\right)\right\}.$$

Note that the quantity in the square brackets

$$[\cdots] = 2|\nabla f|^2 p\bar{\rho}^{1+\frac{d}{2}-2\alpha}. \tag{A.49}$$

Then, also by the assumption that ϵ, $\frac{\varepsilon_\rho}{\epsilon} = o(1)$, we have with large enough N,

$$\mathbb{E}Y_j^2 = \epsilon^{-d/2+1}\left\{m_2' m_0^2 p^3\bar{\rho}^{3d-4\alpha+2}|\nabla f|^2 + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right)\right\} \tag{A.50}$$

$$\leqslant \bar{\nu}_Y(x) \sim \epsilon^{-d/2+1}p^3\bar{\rho}^{3d-4\alpha+2}(x)\|\nabla f\|_\infty^2,$$

where in obtaining the last row we assume that $\|\nabla f\|_\infty > 0$ (because otherwise the theorem holds trivially) and use that $p(x) > p_{min}$ for all x. Since $\bar{\nu}_Y(x) \geqslant c_{f,p,\alpha}\epsilon^{-d/2+1}$ for $c_{f,p,\alpha} > 0$, the needed threshold of large N for $\mathbb{E}Y_j^2 \leqslant \bar{\nu}_Y(x)$ is determined by $(\mathcal{M}, p, f, \alpha)$. Meanwhile, under the condition that $\frac{\log N}{N} = o(\epsilon^{d/2+1})$, with sufficiently large N and the threshold is determined by $(\mathcal{M}, p, f, \alpha)$, we have $40\frac{\log N}{N} < \frac{9c_{f,p,\alpha}\epsilon^{-d/2+1}}{L_Y^2} \leqslant \frac{9\bar{\nu}_Y(x)}{L_Y^2}$, i.e., $\sqrt{40\frac{\log N}{N}\bar{\nu}_Y(x)} < \frac{3\bar{\nu}_Y(x)}{L_Y}$ for any $x \in \mathcal{M}$. Then, by the classical Bernstein, w.p. higher than $1 - 2N^{-10}$,

$$|\frac{1}{N}\sum_{j=1}^N Y_j| \leqslant \sqrt{\frac{40\log N}{N}\bar{\nu}_Y(x)} = O^{[1]}\left(\|\nabla f\|_\infty p^{3/2}\bar{\rho}^{\frac{3}{2}d-2\alpha+1}(x)\epsilon^{-d/4+1/2}\sqrt{\frac{\log N}{N}}\right),$$

and we call the event the good event E_3.

Note that in (A.50), when $|\nabla f(x)| > 0$, one can bound the variance of Y_j at x by $\bar{\nu}_Y(x) \sim \epsilon^{-d/2+1}p^3\bar{\rho}^{3d-4\alpha+2}(x)|\nabla f(x)|^2$ and obtain the same large deviation bound where $\|\nabla f\|_\infty$ is replaced with $|\nabla f(x)|$, allowing the large N threshold to depend on x (such that the $O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right)$ term in (A.50) is dominated by $\Theta^{[1]}(1)$ multiplied the first term, and $40\frac{\log N}{N} < \frac{9\bar{\nu}_Y(x)}{L_Y^2}$). An alternative way to obtain an x-uniform threshold of large N is by adding 0.1^2 to $|\nabla f(x)|^2$ in setting $\bar{\nu}_Y(x)$, so that the x-dependent constant in front of $\epsilon^{-d/2+1}$ is uniformly bounded from below. This leads to the same variance error bound where $\|\nabla f\|_\infty$ is replaced with $|\nabla f(x)| + 0.1$. The above verifies Remark 3.3.

Back to (A.45), with (A.47), we have that under good events E_2 and E_3,

$$\frac{1}{\epsilon\tilde{m}\hat{\rho}^2}\left|\frac{\frac{1}{N}\sum_{j=1}^N F_j}{\frac{1}{N}\sum_{j=1}^N G_j} - \frac{\mathbb{E}F}{\mathbb{E}G}\right| = \frac{1}{\epsilon\tilde{m}\hat{\rho}^2}\frac{|\frac{1}{N}\sum_{j=1}^N Y_j|}{\mathbb{E}G\cdot\frac{1}{N}\sum_{j=1}^N G_j} = \frac{O^{[1]}\left(\|\nabla f\|_\infty p^{3/2}\bar{\rho}^{\frac{3}{2}d-2\alpha+1}\epsilon^{-d/4+1/2}\sqrt{\frac{\log N}{N}}\right)}{\epsilon\bar{\rho}^2(m_0 p\bar{\rho}^{d-\alpha})^2}$$

$$= O^{[1]}\left(\|\nabla f\|_\infty p^{-1/2}\bar{\rho}^{-\frac{d}{2}-1}\epsilon^{-d/4-1/2}\sqrt{\frac{\log N}{N}}\right),$$

where the location x is omitted in the notation. By that $\bar{\rho} = p^{-1/d}$, this proves (A.38). □

Proof of Theorem 3.6. By the definition of $L_{un}^{(\alpha)} f$ and that of F_j, G_j in (A.36),

$$L_{un}^{(\alpha)} f(x) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\epsilon^{\frac{m_2}{2}}} \frac{1}{\hat{\rho}(x)^{\alpha}} (F_j(x) - f(x) G_j(x)) =: \frac{1}{N} \sum_{j=1}^{N} H_j.$$

We have computed $\mathbb{E}F - f(x)\mathbb{E}G$ in (A.43), and (A.39) gives that $\sup_{x \in \mathcal{M}} |\hat{\rho}(x)^{-\alpha} - \hat{\rho}(x)^{-\alpha}| = O^{[p]}(\varepsilon_\rho)$. Then,

$$\begin{aligned}
\mathbb{E}H_j &= \frac{1}{\epsilon^{\frac{m_2}{2}}} \hat{\rho}(x)^{-\alpha} (\mathbb{E}F(x) - f(x)\mathbb{E}G(x)) \\
&= \frac{1}{\epsilon^{\frac{m_2}{2}}} \hat{\rho}(x)^{-\alpha} \left(\epsilon \frac{m_2}{2} p \bar{\rho}^{d+2-\alpha} \mathcal{L}^{(\alpha)} f + O^{[f,p]}(\epsilon^2, \varepsilon_\rho) \right) \\
&= p \bar{\rho}^{d+2-2\alpha} \mathcal{L}^{(\alpha)} f(x) + O^{[f,p]}(\epsilon, \frac{\varepsilon_\rho}{\epsilon}).
\end{aligned}$$

By $\bar{\rho} = p^{-1/d}$, this proves that

$$\mathbb{E}L_{un}^{(\alpha)} f(x) = p^{\frac{2\alpha-2}{d}} \mathcal{L}^{(\alpha)} f(x) + O^{[f,p]}(\epsilon, \frac{\varepsilon_\rho}{\epsilon}), \tag{A.51}$$

where the constant in $O^{[f,p]}(\cdot)$ is uniform for all $x \in \mathcal{M}$.

To analyze the variance, first note the boundedness of $|H_j|$ as

$$|H_j| \leqslant L_H = \Theta^{[f,p]}(\epsilon^{-d/2-1}),$$

which follows by the boundedness of G_j, F_j in (A.48) and the uniform boundedness of $\hat{\rho}$. For the variance of H_j, we have

$$\mathbb{E}H_j^2 = (\frac{1}{\epsilon^{\frac{m_2}{2}}} \hat{\rho}(x)^{-\alpha})^2 (\mathbb{E}F_j^2 + f(x)^2 \mathbb{E}G_j^2 - 2f(x)\mathbb{E}F_j G_j),$$

and we have computed $\mathbb{E}F^2$, $\mathbb{E}G^2$ and $\mathbb{E}FG$ in the proof of Theorem 3.5. Specifically, with notation the same as therein, we have

$$\begin{aligned}
\mathbb{E}F^2 &= \epsilon^{-d/2} \left\{ m_0' f^2 p \bar{\rho}^{d-2\alpha} + \epsilon \frac{m_2'}{2} \bar{\rho}^{1+\frac{d}{2}} (\omega f^2 p \bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(f^2 p \bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[f,p]}\left(\epsilon^2, \varepsilon_\rho\right) \right\}, \\
\mathbb{E}G^2 &= \epsilon^{-d/2} \left\{ m_0' p \bar{\rho}^{d-2\alpha} + \epsilon \frac{m_2'}{2} \bar{\rho}^{1+\frac{d}{2}} (\omega p \bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(p \bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[p]}\left(\epsilon^2, \varepsilon_\rho\right) \right\}, \\
\mathbb{E}FG &= \epsilon^{-d/2} \left\{ m_0' f p \bar{\rho}^{d-2\alpha} + \epsilon \frac{m_2'}{2} \bar{\rho}^{1+\frac{d}{2}} (\omega f p \bar{\rho}^{1+\frac{d}{2}-2\alpha} + \Delta(f p \bar{\rho}^{1+\frac{d}{2}-2\alpha})) + O^{[f,p]}\left(\epsilon^2, \varepsilon_\rho\right) \right\}.
\end{aligned}$$

Also, by (A.49), where the square brackets denote the same quantity as before, we have

$$\begin{aligned}
\mathbb{E}F^2 + f^2 \mathbb{E}G^2 - 2f\mathbb{E}FG &= \epsilon^{-d/2} \epsilon \frac{m_2'}{2} \bar{\rho}^{1+\frac{d}{2}} [\cdots] \\
&= \epsilon^{-d/2} \left\{ \epsilon m_2' |\nabla f|^2 p \bar{\rho}^{2+d-2\alpha} + O^{[f,p]}\left(\epsilon^2, \varepsilon_\rho\right) \right\}.
\end{aligned}$$

Then, also by (A.39), we have

$$\begin{aligned}\mathbb{E}H_j^2 &= \frac{4}{m_2^2}\epsilon^{-1-d/2}\hat{\rho}^{-2\alpha}(x)\left\{m_2'|\nabla f|^2 p\bar{\rho}^{2+d-2\alpha}(x) + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right)\right\}\\ &= \frac{4}{m_2^2}\epsilon^{-1-d/2}\left\{m_2'|\nabla f|^2 p\bar{\rho}^{2+d-4\alpha}(x) + O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right)\right\} \quad (\text{by that } \hat{\rho}^{-2\alpha}(x) = \bar{\rho}^{-2\alpha}(x) + O^{[p]}(\varepsilon_\rho))\\ &\leqslant \bar{v}_H(x) = \Theta^{[1]}(\epsilon^{-1-d/2}\|\nabla f\|_\infty^2 p^{\frac{4\alpha-2}{d}}(x)), \text{ with large } N, \text{ the threshold depending on } (\mathcal{M}, p, f, \alpha).\end{aligned}$$

In obtaining the last row, we assumed $\|\nabla f\|_\infty > 0$ (when $\|\nabla f\|_\infty = 0$, the theorem holds trivially), and used that $O^{[f,p]}\left(\epsilon, \frac{\varepsilon_\rho}{\epsilon}\right) = o(1)$, $\bar{\rho} = p^{-1/d}$, and that p is uniformly bounded from below. Then same as in the proof of Theorem 3.5, the threshold of large N to achieve the sub-Gaussian tail in Bernstein inequality is determined by $(\mathcal{M}, p, f, \alpha)$. As a result, when N is large enough, we have that w.p. higher than $1 - 2N^{-10}$,

$$\left|\frac{1}{N}\sum_{j=1}^N H_j - \mathbb{E}H_j\right| \leqslant \sqrt{\frac{40\log N}{N}\bar{v}_H(x)} = O^{[1]}\left(\|\nabla f\|_\infty p^{\frac{2\alpha-1}{d}}(x)\sqrt{\frac{\log N}{N\epsilon^{d/2+1}}}\right).$$

To replace $\|\nabla f\|_\infty$ with $|\nabla f(x)|$ when strictly positive, or with $+0.1$, as in Remark 3.3, the same argument by re-defining $\bar{v}_H(x)$ similarly as in the proof of Theorem 3.5 applies. Combined with (A.51), this finishes the proof. □

Proof of Theorem 3.7. Suppose $\varphi \neq 0$ and $\|\nabla f\|_\infty > 0$, otherwise the theorem trivially holds. By definition (3.9), we have that

$$\langle \varphi, L_{un}^{(\alpha)} f\rangle_p = \frac{1}{N}\sum_{j=1}^N H_j,$$

where

$$H_j := \frac{2\epsilon^{-1}}{m_2}\int_{\mathcal{M}} \hat{K}(x, x_j)(f(x_j) - f(x))\varphi(x)p(x)\,\mathrm{d}V(x),$$

and $\hat{K}(x,y)$ is defined as in (3.4). We define

$$B(g,f) := \frac{2\epsilon^{-1}}{m_2}\int_{\mathcal{M}}\int_{\mathcal{M}} \hat{K}(x,y)(f(y) - f(x))g(x)p(x)p(y)\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

By $\hat{K}(x,y) = \hat{K}(y,x)$, $B(g,f) = B(f,g)$ i.e., $B(g,f)$ is a symmetric bilinear form. Meanwhile,

$$B(f,f) = -\mathcal{E}^{(\alpha)}(f,f),$$

where, by Proposition 3.2, $\mathcal{E}^{(\alpha)}(f,f) = -\langle f, \Delta_{p_\alpha} f\rangle_{p_\alpha} + O^{[f,p]}(\epsilon, \varepsilon_\rho)$. Thus,

$$\begin{aligned}\mathbb{E}H_j = B(\varphi,f) &= \frac{1}{4}(B(\varphi+f,\varphi+f) - B(\varphi-f,\varphi-f))\\ &= \frac{1}{4}(-\mathcal{E}^{(\alpha)}(\varphi+f,\varphi+f) + \mathcal{E}^{(\alpha)}(\varphi-f,\varphi-f))\\ &= \frac{1}{4}\left(\langle \varphi+f, \Delta_{p_\alpha}(\varphi+f)\rangle_{p_\alpha} - \langle \varphi-f, \Delta_{p_\alpha}(\varphi-f)\rangle_{p_\alpha} + O^{[\varphi,f,p]}(\epsilon, \varepsilon_\rho)\right)\\ &= \langle \varphi, \Delta_{p_\alpha} f\rangle_{p_\alpha} + O^{[\varphi,f,p]}(\epsilon, \varepsilon_\rho).\end{aligned}$$

To analyze the variance, we compute the boundedness and variance of H_j. To avoid obtaining $\frac{\varepsilon_\rho}{\epsilon}$, we cannot directly apply Lemma 3.1 and Lemma A.2 as in the proof of Theorem 3.6. By Cauchy–Schwartz inequality and that $\hat{K}(x,y) \geqslant 0$, we have

$$|H_j| \leqslant \frac{2\epsilon^{-1}}{m_2}\left(\int_{\mathcal{M}} K(x,x_j)(f(x_j)-f(x))^2 p(x)\,\mathrm{d}V(x)\right)^{1/2}\left(\int_{\mathcal{M}} K(x,x_j)\varphi(x)^2 p(x)\,\mathrm{d}V(x)\right)^{1/2}.$$

We define ①(y) and ②(y) as below and claim the following: for any $y \in \mathcal{M}$,

$$①(y) := \int_{\mathcal{M}} \hat{K}(x,y)\varphi(x)^2 p(x)\,\mathrm{d}V(x) \leqslant c_1\|\varphi p^{\alpha/d}\|_\infty^2, \quad c_1 = \Theta^{[1]}(1), \tag{A.52}$$

$$②(y) := \int_{\mathcal{M}} \hat{K}(x,y)(f(y)-f(x))^2 p(x)\,\mathrm{d}V(x) = O^{[f,p]}(\epsilon). \tag{A.53}$$

If true, then we have

$$|H_j| = \epsilon^{-1} O^{p,f,\varphi}(\sqrt{\epsilon}) = O^{[p,f,\varphi]}(\epsilon^{-1/2}),$$

and at the same time, using the upper bound (A.52), we have

$$\begin{aligned}\mathbb{E}H_j^2 &\leqslant \left(\frac{4\epsilon^{-1}}{m_2}\right) c_1\|\varphi p^{\alpha/d}\|_\infty^2\left(\frac{1}{\epsilon m_2}\int_{\mathcal{M}}\int_{\mathcal{M}} \hat{K}(x,y)(f(y)-f(x))^2 p(x)\,\mathrm{d}V(x)p(y)\,\mathrm{d}V(y)\right)\\ &= \epsilon^{-1}c_1'\|\varphi p^{\alpha/d}\|_\infty^2 \mathcal{E}^{(\alpha)}(f,f), \quad c_1' = \Theta^{[1]}(1).\end{aligned}$$

Again, $\mathcal{E}^{(\alpha)}(f,f) = -\langle f, \Delta_{p_\alpha} f\rangle_{p_\alpha} + O^{[f,p]}(\epsilon, \varepsilon_\rho)$, where $-\langle f, \Delta_{p_\alpha} f\rangle_{p_\alpha} = \int p_\alpha |\nabla f|^2 > 0$. We then have that

$$\mathrm{Var}(H_j) \leqslant \mathbb{E}H_j^2 \leqslant \bar{v}_H = \Theta^{[1]}\left(\epsilon^{-1}\|\varphi p^{\alpha/d}\|_\infty^2 \int p_\alpha |\nabla f|^2\right).$$

Thus, when N is large enough, w.p. higher than $1 - 2N^{-10}$, we have

$$\left|\frac{1}{N}\sum_{j=1}^N H_j - \mathbb{E}H_j\right| \leqslant \sqrt{\frac{40\log N}{N}\bar{v}_H} = O^{[1]}\left(\|\varphi\|_\infty \|p^{\alpha/d}\|_\infty \sqrt{\frac{\log N}{N\epsilon}\int p_\alpha |\nabla f|^2}\right).$$

It remains to show (A.52) and (A.53) to finish the proof of the theorem.

Proof of (A.52): by definition,

$$①(y) = \frac{1}{\hat\rho(y)^\alpha}\epsilon^{-\frac{d}{2}}\int_{\mathcal M} k_0\left(\frac{\|x-y\|^2}{\epsilon\hat\rho(x)\hat\rho(y)}\right)\frac{\varphi(x)^2 p(x)}{\hat\rho(x)^\alpha}\,\mathrm{d}V(x),$$

where by that $\sup_{x\in\mathcal M}\frac{|\hat\rho(x)-\bar\rho(x)|}{|\bar\rho(x)|} < \varepsilon_\rho < 0.1$, we have $\hat\rho(x)\hat\rho(y) \leqslant 1.1^2\bar\rho(x)\bar\rho(y)$, and then by Assumption (3.1)(C2) we have

$$k_0\left(\frac{\|x-y\|^2}{\epsilon\hat\rho(x)\hat\rho(y)}\right) \leqslant a_0 e^{-a\frac{\|x-y\|^2}{\epsilon\hat\rho(x)\hat\rho(y)}} \leqslant a_0 e^{-\frac{a}{1.1^2}\frac{\|x-y\|^2}{\epsilon\bar\rho(x)\bar\rho(y)}} = \bar k_1\left(\frac{\|x-y\|^2}{\epsilon\bar\rho(x)\bar\rho(y)}\right),$$

where $\bar k_1(r) := a_0 e^{-\frac{a}{1.1^2}r}$ and satisfies Assumption 3.1. We introduce $G^{(\rho)}_\epsilon[h]$ when in the definition (3.6) the kernel function k_0 is replaced with some h that satisfies Assumption 3.1. That is, $G^{(\rho)}_\epsilon = G^{(\rho)}_\epsilon[k_0]$, and the notation $[k_0]$ is to declare the kernel function being used. To proceed, by that $\hat\rho(x)^{-\alpha} \leqslant \max\{0.9^\alpha, 1.1^\alpha\}\bar\rho(x)^{-\alpha} := c_2\bar\rho(x)^{-\alpha}$, we have that for any $x\in\mathcal M$,

$$①(y) \leqslant \frac{c_2^2}{\bar\rho(y)^\alpha}\epsilon^{-\frac{d}{2}}\int_{\mathcal M}\bar k_1\left(\frac{\|x-y\|^2}{\epsilon\bar\rho(x)\bar\rho(y)}\right)\frac{\varphi(x)^2p(x)}{\bar\rho(x)^\alpha}\,\mathrm{d}V(x) = \frac{c_2^2}{\bar\rho(y)^\alpha}\bar\rho(y)^{d/2}G^{(\bar\rho)}_{\epsilon\bar\rho(y)}[\bar k_1](\frac{\varphi^2 p}{\bar\rho^\alpha})(y).$$

By Lemma 3.1,

$$G^{(\bar\rho)}_{\epsilon\bar\rho(y)}[\bar k_1](\frac{\varphi^2p}{\bar\rho^\alpha})(y) = m_0[\bar k_1]\bar\rho^{d/2}(\frac{\varphi^2p}{\bar\rho^\alpha})(y) + \bar\rho(y)O^{[p,f,\varphi]}(\epsilon),$$

and then

$$①(y) \leqslant c_2^2 m_0[\bar k_1](p\bar\rho^{d-2\alpha}\varphi^2)(y) + O^{[p,f,\varphi]}(\epsilon) = \Theta^{[1]}((p\bar\rho^{d-2\alpha}\varphi^2)(y)).$$

By that $\bar\rho = p^{1/d}$, we have shown that $①(y) \leqslant c_1 p(y)^{2\alpha/d}\varphi(y)^2$, where $c_1 = \Theta^{[1]}(1)$, and this proves (A.52).

Proof of (A.53): similarly, we have

$$②(y) \leqslant \frac{c_2^2}{\bar\rho(y)^\alpha}\epsilon^{-\frac{d}{2}}\int_{\mathcal M}\bar k_1\left(\frac{\|x-y\|^2}{\epsilon\bar\rho(x)\bar\rho(y)}\right)\frac{(f(y)-f(x))^2}{\bar\rho(x)^\alpha}p(x)\,\mathrm{d}V(x)$$

$$= \frac{c_2^2}{\bar\rho(y)^\alpha}\bar\rho(y)^{d/2}G^{(\bar\rho)}_{\epsilon\bar\rho(y)}[\bar k_1](g)(y), \quad g(x) := (f(y)-f(x))^2(\frac{p}{\bar\rho^\alpha})(x)\,.$$

Note that $g\in C^\infty(\mathcal M)$ and $g(y)=0$. By Lemma 3.1, we have

$$G^{(\bar\rho)}_{\epsilon\bar\rho(y)}[\bar k_1](g)(y) = \epsilon\bar\rho(y)\frac{m_2[\bar k_1]}{2}\Delta g(y) + \bar\rho(y)^2O^{[f,p]}(\epsilon^2),$$

and then

$$②(y) \leqslant \frac{c_2^2}{\bar\rho(y)^\alpha}\bar\rho(y)^{d/2}\left(\epsilon\bar\rho(y)\frac{m_2[\bar k_1]}{2}\Delta g(y) + \bar\rho(y)^2O^{[f,p]}(\epsilon^2)\right)$$

$$= c_2^2\frac{m_2[\bar k_1]}{2}\epsilon\bar\rho(y)^{d/2-\alpha+1}\Delta g(y) + O^{[f,p]}(\epsilon^2) = O^{[f,p]}(\epsilon),$$

which proves (A.53). □

Proof of Theorem 3.8. The proof combines the approach in the proof of Theorem 3.5 and the computation in that of Theorem 3.4. Define

$$F_j := \epsilon^{-d/2} k_0\left(\frac{\|x - x_j\|^2}{\epsilon}\right)\frac{f(x_j)}{\hat{p}(x_j)^\beta}, \quad G_j := \epsilon^{-d/2} k_0\left(\frac{\|x - x_j\|^2}{\epsilon}\right)\frac{1}{\hat{p}(x_j)^\beta},$$

then we have

$$\mathbb{E}F = G_\epsilon\left(\frac{fp}{\hat{p}^\beta}\right) = \int_{\mathcal{M}} \epsilon^{-d/2} k_0\left(\frac{\|x - y\|^2}{\epsilon}\right) f(y)\hat{p}(y)^{-\beta} p(y)\, \mathrm{d}V(y).$$

By (A.33) and the constant c_β defined as therein, again defining $q = p^{1-\beta}$, we have

$$\left|G_\epsilon\left(\frac{fp}{\hat{p}^\beta}\right) - G_\epsilon(fp^{1-\beta})\right| \leqslant \|f\|_\infty \int \epsilon^{-d/2} k_0\left(\frac{\|x - y\|^2}{\epsilon}\right) p(y)|\hat{p}(y)^{-\beta} - p(y)^{-\beta}|\, \mathrm{d}V(y)$$

$$\leqslant c_\beta |\beta| \varepsilon_p \|f\|_\infty \int \epsilon^{-d/2} k_0\left(\frac{\|x - y\|^2}{\epsilon}\right) p(y)^{1-\beta}\, \mathrm{d}V(y) = O^{[f,p,\beta]}(\beta\varepsilon_p),$$

where we apply Lemma A.5 to obtain that $\int \epsilon^{-d/2} k_0\left(\frac{\|x-y\|^2}{\epsilon}\right) q(y)\, \mathrm{d}V(y) \leqslant \|q\|_\infty O(1)$, and absorb the constants $\|q\|_\infty$, $\|f\|_\infty$ and c_β in to the notation $O^{[f,p,\beta]}(\cdot)$. In the rest of the proof, we omit the dependence on β in the superscript and write it as $O^{[f,p]}(\beta\varepsilon_p)$, while we keep β in $(\cdot)$ to indicate that the term vanishes when $\beta = 0$. Then, using Lemma A.5 to expand $G_\epsilon(fp^{1-\beta})$, we have that

$$\mathbb{E}F = G_\epsilon(fp^{1-\beta}) + O^{[f,p]}(\beta\varepsilon_p) = m_0 fp^{1-\beta} + \epsilon\frac{m_2}{2}(\omega fp^{1-\beta} + \Delta(fp^{1-\beta})) + O^{[f,p]}(\epsilon^2, \beta\varepsilon_p).$$

Taking $f = 1$ then gives

$$\mathbb{E}G = m_0 p^{1-\beta} + \epsilon\frac{m_2}{2}(\omega p^{1-\beta} + \Delta(p^{1-\beta})) + O^{[p]}(\epsilon^2, \beta\varepsilon_p) = m_0 p^{1-\beta} + O^{[p]}(\epsilon, \beta\varepsilon_p).$$

We can then compute and bound the bias error as

$$\frac{1}{\epsilon\tilde{m}}\frac{\mathbb{E}F - f(x)\mathbb{E}G}{\mathbb{E}G} = \frac{1}{\epsilon\tilde{m}}\frac{\epsilon\frac{m_2}{2}(\Delta(fp^{1-\beta}) - f\Delta(p^{1-\beta})) + O^{[f,p]}(\epsilon^2, \beta\varepsilon_p)}{m_0 p^{1-\beta} + O^{[p]}(\epsilon, \beta\varepsilon_p)}$$

$$= \Delta f + 2\frac{\nabla p^{1-\beta}}{p^{1-\beta}} \cdot \nabla f + O^{[f,p]}\left(\epsilon, \beta\frac{\varepsilon_p}{\epsilon}\right),$$

which, similarly as in (A.42), holds when N exceeds a threshold depending on $(\mathcal{M}, p, \beta)$. The variance analysis follows a similar computation as before, specifically the computation of the quantities of $\mathbb{E}F^2$, $\mathbb{E}G^2$ and $\mathbb{E}(FG)$. First, observe that

$$\mathbb{E}G^2 = \epsilon^{-d/2}\{m_0[k_0^2]p^{1-2\beta} + O^{[p]}(\epsilon, \beta\varepsilon_p)\},$$

and then, using that $p^{1-2\beta}(x) \geqslant \Theta^{[p,\beta]}(1) > 0$ for all x, one verifies that w.p. higher than $1 - 2N^{-10}$,

$$\frac{1}{N}\sum_j G_j = \mathbb{E}G + O^{[p]}\left(\sqrt{\frac{\log N}{N\epsilon^{d/2}}}\right).$$

Define $Y_j := F_j\mathbb{E}G - G_j\mathbb{E}F$, then $\mathbb{E}Y = 0$. Following the same method as before, one verifies that, with $m_2' := m_2[k_0^2]$,

$$\mathbb{E}Y^2 = \mathbb{E}F^2(\mathbb{E}G)^2 + EG^2(\mathbb{E}F)^2 - 2\mathbb{E}(FG)\mathbb{E}F\mathbb{E}G = \epsilon^{-d/2+1}\left\{m_2' m_0^2 p^{3-4\beta}|\nabla f|^2 + O^{[f,p]}(\epsilon, \beta\frac{\varepsilon_p}{\epsilon})\right\}.$$

Similarly as in the proof of Theorem 3.5, this gives that (assuming $\|\nabla f\|_\infty > 0$ otherwise the theorem holds trivially) when N exceeds a threshold determined by $(\mathcal{M}, p, f, \beta)$, w.p. higher than $1 - 2N^{-10}$,

$$\left|\frac{1}{N}\sum_{j=1}^N Y_j\right| = O\left(\|\nabla f\|_\infty p^{3/2-2\beta}\epsilon^{-d/4+1/2}\sqrt{\frac{\log N}{N}}\right).$$

One can also replace $\|\nabla f\|_\infty$ with $|\nabla f(x)|$ when strictly positive, or with $+0.1$, as in Remark 3.3.

Putting together, we have that

$$\frac{1}{\epsilon\tilde{m}}\left|\frac{\frac{1}{N}\sum_{j=1}^N F_j}{\frac{1}{N}\sum_{j=1}^N G_j} - \frac{\mathbb{E}F}{\mathbb{E}G}\right| = \frac{1}{\epsilon\tilde{m}}\frac{|\frac{1}{N}\sum_{j=1}^N Y_j|}{\mathbb{E}G\cdot\frac{1}{N}\sum_{j=1}^N G_j} \leqslant \frac{O^{[1]}\left(\|\nabla f\|_\infty p^{3/2-2\beta}\epsilon^{-d/4+1/2}\sqrt{\frac{\log N}{N}}\right)}{\epsilon(m_0 p^{1-\beta})^2}$$
$$= O^{[1]}\left(\|\nabla f\|_\infty p^{-1/2}\epsilon^{-d/4-1/2}\sqrt{\frac{\log N}{N}}\right).$$

Combining the bias and variance error bounds proves the theorem. □

A.2 *Technical lemmas of differential geometry*

A.2.1 *Local charting on $\mathcal{M}$.* The following lemma is about manifold local charting, where we have metric and volume comparisons between the manifold and the ambient Euclidean space $\mathbb{R}^D$.

LEMMA A.3 ([11, Lemmas 6 and 7]). Suppose $\mathcal{M}$ is a d-dimensional C^3, boundaryless (thus closed) manifold that is isometrically embedded in $\mathbb{R}^D$. Then there exists some $\delta_0(\mathcal{M}) > 0$ such that for any $\delta < \delta_0$ and any $x \in \mathcal{M}$,

(i) $\mathcal{M} \cap B_\delta(x)$ is isomorphic to a ball in $\mathbb{R}^d$.

(ii) On the local chart at each x, let ϕ_x be the orthogonal projection to the tangent plane $T_x\mathcal{M}$ embedded as an affine subspace of $\mathbb{R}^D$, and call $u(y) := \phi_x(y)$ the tangent coordinate of y, then

$$0.9\|y-x\|_{\mathbb{R}^D} < \|u(y)\|_{\mathbb{R}^d} < 1.1\|y-x\|_{\mathbb{R}^D}, \quad 0.9 < \left|\det\left(\frac{dy}{du}\right)\right| < 1.1, \quad \forall y \in \mathcal{M}\cap B_\delta(x). \tag{A.54}$$

(iii) Let $d_{\mathcal{M}}$ denote the manifold geodesic distance, then

$$\|x-y\|_{\mathbb{R}^D} \leqslant d_{\mathcal{M}}(x,y) \leqslant 1.1\|x-y\|_{\mathbb{R}^D}, \quad \forall y \in B_\delta(x)\cap\mathcal{M}. \tag{A.55}$$

Proof. At every point x, (i) holds when $\delta < \delta_{x,1}$ for some $\delta_{x,1} > 0$, and then the local chart can be defined where the normal coordinates s ($\exp_x(s) = y$) and the tangent coordinates u match up to $O(\|u\|^3)$ [11, Lemma 6], the squared metric of $\|y-x\|_{\mathbb{R}^D}^2$ and $\|u\|^2$ match up to $O(\|u\|^4)$, and the Jacobian's match via

$$\left|\det\left(\frac{dy}{du}\right)\right| = 1 + b_x^{(v)}(u) + c_x^{(v)}(u) + O(\|u\|^4), \tag{A.56}$$

where $b_x^{(v)}$ ($c_x^{(v)}$) is a homogeneous polynomial of degree 2 (3) of the variable $u = (u_1, \cdots, u_d)$ [**11**, Lemma 7]. Thus, (A.54) and (A.55) hold on $\mathcal{M} \cap B_\delta(x)$ when $\delta < \delta_{x,2}$ for some $0 < \delta_{x,2} \leqslant \delta_{x,1}$. The $\min_{x\in\mathcal{M}} \delta_{x,2}$ exists due to the smoothness and compactness of $\mathcal{M}$, and the minimum can be used as $\delta_0(\mathcal{M})$. □

A.2.2 *Covering number of* $\mathcal{M}$ Introduce the definitions:

DEFINITION A.1 Let (X, d) be a metric space and $Y \subset X$. Let $\epsilon > 0$, then $P \subset X$ is called a ϵ-net of Y if $\forall x \in Y$, $\exists x_0 \in P$, s.t. $d(x, x_0) \leqslant \epsilon$. The covering number of Y, denoted by $\mathcal{N}(Y, d, \epsilon)$, is defined to be the smallest cardinality of an ϵ-net of Y.

DEFINITION A.2 Let (X, d) be a metric space. Let $\epsilon > 0$, then $P \subset X$ is said to be ϵ-separated if $d(x, y) > \epsilon$ for all distinct points $x, y \in P$. The packing number of $Y \subset X$ denoted by $\mathcal{P}(Y, d, \epsilon)$ is defined to be the largest cardinality of an ϵ-separated subset of Y.

The following lemma bounds the covering number of $\mathcal{M}$ using Euclidean balls in $\mathbb{R}^D$, which has been established in literature. We reproduce under our setting for completeness.

LEMMA A.4 For any $r < \delta_0$, where δ_0 is defined in Lemma A.3, $\mathcal{N}(\mathcal{M}, \|\cdot\|_{\mathbb{R}^D}, r) \leqslant V(\mathcal{M}) r^{-d}$, where $V(\mathcal{M})$ equals an $O_d(1)$ constant times the Riemannian volume of $\mathcal{M}$.

Proof of Lemma A.4. The proof uses Lemma A.3 and standard arguments as in [**53**, Section 4.2]. Let d_E be the Euclidean distance in $\mathbb{R}^D$.

Let d_E denote the metric on $\mathcal{M}$ induced by the Euclidean metric in $\mathbb{R}^D$, that is, $d_E(x, y) = \|x-y\|_{\mathbb{R}^D}$, where $x, y \in \mathcal{M}$. The packing number $\mathcal{P}(\mathcal{M}, d_E, r)$ always upper bounds the covering number (see e.g., [**53**, Lemma 4.2.6 and Lemma 4.2.8]); thus, it suffices to upper bound $\mathcal{P}(\mathcal{M}, d_E, r)$.

Denote by $B_r(x, d_E)$ the open ball on $(\mathcal{M}, d_E)$ centered at x, and $B_{r,\mathbb{R}^m}(x)$ the open Euclidean ball of radius r centered at x in $\mathbb{R}^m$. Without declaring m, $B_r(x)$ means $B_{r,\mathbb{R}^D}(x)$. By definition, $B_r(x, d_E) = B_r(x) \cap \mathcal{M}$. Suppose $r < \delta_0$ in Lemma A.3, we consider the manifold volume Vol of these Euclidean balls, where for $Y \subset \mathcal{M}$, $\mathrm{Vol}(Y) := \int_{\mathcal{M}} \mathbf{1}_Y \, \mathrm{d}V$ when integrable. By Lemma A.3(ii), on $T_x(\mathcal{M})$ which is viewed as $\mathbb{R}^d$,

$$B_{0.9r,\mathbb{R}^d}(0) \subset \phi_x(B_r(x, d_E)),$$

and $\left|\det(\frac{dy}{du})\right| > 0.9$ on $B_r(x, d_E)$, then

$$\begin{aligned} Vol(B_r(x, d_E)) &= \int_{\phi_x(B_r(x,d_E))} \left|\det(\frac{\mathrm{d}y}{\mathrm{d}u})\right| \mathrm{d}u \geqslant \int_{\{u,\, \|u\|<0.9r\}} \left|\det(\frac{\mathrm{d}y}{\mathrm{d}u})\right| \mathrm{d}u \\ &\geqslant 0.9 \int_{\{u,\, \|u\|<0.9r\}} \mathrm{d}u = 0.9 v_d (0.9r)^d, \end{aligned}$$

where v_d is the Euclidean volume of a unit d-sphere.

Now let P be a maximal r-separated subset of $\mathcal{M}$ (under d_E) such that $\mathrm{Card}(P) = n = \mathcal{P}(\mathcal{M}, d_E, r)$, and $P = \{x_1, \cdots, x_n\}$. By definition of r-separateness, $B_{\frac{r}{2}}(x_i, d_E)$ are disjoint, thus

$$Vol(\mathcal{M}) \geqslant \sum_{i=1}^{n} Vol(B_{\frac{r}{2}}(x_i, d_E)) \geqslant n \cdot 0.9 v_d (0.9 \frac{r}{2})^d,$$

that is, for $V(\mathcal{M})$ which is an $O_d(1)$ constant times the Riemannian volume of $\mathcal{M}$,

$$n \leqslant \frac{V(\mathcal{M})}{r^d}.$$

This proves that $\mathcal{N}(\mathcal{M}, d_E, r) \leqslant \mathcal{P}(\mathcal{M}, d_E, r) \leqslant \frac{V(\mathcal{M})}{r^d}$. □

A.2.3 *Fixed-bandwidth integral operator*

ASSUMPTION A.3 (Assumption on k_0 in [11]). (C1') Regularity. k_0 is continuous on $[0, \infty)$, C^2 on $(0, \infty)$.

(C2') Decay condition. k_0 and up to its second derivatives are bounded on $(0, \infty)$ and have sub-exponential tail, specifically, $\exists a, a_l > 0$, s.t., $|k_0^{(l)}(\xi)| \leq a_l e^{-a\xi}$ for all $\xi > 0$, $l = 0, 1, 2$. To exclude the case that $k_0 \equiv 0$, suppose $\|k_0\|_\infty > 0$.

LEMMA A.5 ([**11**, Lemma 8]). Suppose h satisfies Assumption A.3. For any $f \in C^\infty(\mathcal{M})$, define

$$G_\epsilon f(x) := \int_{\mathcal{M}} h\left(\frac{\|x-y\|^2}{\epsilon}\right) f(y)\, \mathrm{d}V(y). \tag{A.57}$$

Then there is $\epsilon_0(\mathcal{M}, h) > 0$ such that when $0 < \epsilon < \epsilon_0$,

$$G_\epsilon[h]f = \epsilon^{\frac{d}{2}} \left(m_0[h] f + \epsilon \frac{m_2[h]}{2} (\omega f + \Delta_{\mathcal{M}} f) + O^{[f^{(\leqslant 4)}]}(\epsilon^2) \right),$$

where $\omega(x)$ is determined by local derivatives of the extrinsic manifold coordinates at x, the residual term denoted by big-O with superscript $f^{(\leqslant 4)}$ means that the constant involves up to the fourth derivative of f on $\mathcal{M}$. Specifically, if the residual term is denoted as $r_{f,\epsilon}(x)$, it satisfies $\sup_{x\in\mathcal{M}} |r_{f,\epsilon}(x)| \leqslant C(f)\epsilon^2$, where $C(f) = c(\mathcal{M}, h)(1 + \sum_{l=0}^4 \|D^{(l)} f\|_\infty)$.

For the sake of self-containedness and specifically quantifying the constant in the error term, we provide a proof of this lemma below.

Proof of Lemma A.5. The original proof is in Appendix B of [**11**]. We made slightly more precise the truncation argument of the integral, as well as under the formal statement of assumptions on k_0 as in Assumption A.3.

The proof uses the exponential decay of h to truncate the integral of $dV(y)$ on $\mathcal{M} \cap B_{\delta_\epsilon}(x)$, where the Euclidean ball radius δ_ϵ can be chosen to be $\sqrt{\alpha_0 \epsilon \log \frac{1}{\epsilon}}$ for some $O_d(1)$ constant α_0. For example, let $\alpha_0 = \frac{d+10}{a}$, where a is the sub-exponential decay constant of h in Assumption A.3(C2), then the truncations of integrals used in the proof all incur an error of order $O(\epsilon^{10})$. For the truncation tail bounds to hold, the radius δ_ϵ needs to be smaller than $\delta_0(\mathcal{M})$ in Lemma A.3. The requirement $\delta_\epsilon < \delta_0$ gives rise to the condition that $\epsilon < \epsilon_0$ in the lemma.

Restricting on a local ball, the integrals in the proof are computed via local projected coordinates on $T_x\mathcal{M}$, and using the volume and metric comparison lemmas, [**11**, Lemmas 6 and 7], as detailed in [**11**, Appendix B]. In particular, only the differentiability and sub-exponential decay of up to second derivatives of h and the isometry of the kernel (h is a function of $\|x-y\|^2$) are used; thus, the lemma holds for any h satisfying Assumption A.3. □

The following lemma is the counterpart of Lemma A.5 when h is the indicator function (only to the '$O(\epsilon)$' term, $\epsilon = r^2$ here). It can be implied by [22, Lemma 4] (without proof) and was also given in a different setting for uniform p in [50, Lemma 7]. We include a proof for completeness.

LEMMA A.6 Under Assumption 2.1, $h = \mathbf{1}_{[0,1)}$, there is a constant $\delta_1(\mathcal{M}) < \delta_0$ in Lemma A.3 such that when $r < \delta_1$, for any $x \in \mathcal{M}$,

$$r^{-d}\int_{\mathcal{M}} h\left(\frac{\|x-y\|^2}{r^2}\right)p(y)\,\mathrm{d}V(y) = r^{-d}\int_{\mathcal{M}} \mathbf{1}_{\{\|x-y\|<r\}}p(y)\,\mathrm{d}V(y) = m_0[h]p(x) + O^{[p]}(r^2),$$

and the constant in big-O is uniform for all x.

Proof of Lemma A.6. The proof uses the same technique of that in Lemma A.5. Because $r < \delta_0$ in Lemma A.3, using the local chart, we have that

$$I_r := \int_{\mathcal{M}} \mathbf{1}_{\{\|x-y\|<r\}}p(y)\,\mathrm{d}V(y) = \int_{B'} p(y(u))\left|\det\left(\frac{\mathrm{d}y}{\mathrm{d}u}\right)\right|\mathrm{d}u, \quad B' := \phi_x(B_r(x)\cap\mathcal{M}) \subset \mathbb{R}^d.$$

By that

$$\|y-x\|^2 = |u|^2 + O(|u|^4),$$

where the constant in big-O depends on local derivatives of manifold extrinsic coordinates at x and by compactness of $\mathcal{M}$ is uniform for all x, there is $\delta_1 = \delta_1(\mathcal{M})$ and constant $c_M > 0$ uniform for all x such that when $r < \delta_1$, for any x,

$$B_{r^-} \subset B' \subset B_{r^+}, \quad r^{\pm} = r(1 \pm c_M r^2), \quad B_r := \{u \in \mathbb{R}^d,\ |u| < r\}.$$

We consider upper and lower bounds of I_r, respectively. By that $p > 0$,

$$I_r \leqslant \int_{B_{r^+}} p(y(u))\left|\det\left(\frac{\mathrm{d}y}{\mathrm{d}u}\right)\right|\mathrm{d}u =: I_+.$$

Similarly as in the proof of Lemma A.5,

$$p(y(u)) = p(x) + \nabla_{\mathcal{M}}p(x)\cdot u + O^{[p]}(|u|^2),$$

and by (A.56), $\left|\det\left(\frac{dy}{du}\right)\right| = 1 + O(|u|^2)$, where the constant in big-O depends on local derivatives of manifold extrinsic coordinates at x and by compactness of $\mathcal{M}$ is uniform for all x. This gives that

$$I_+ = \int_{B_{r^+}} \left(p(x) + \nabla_{\mathcal{M}}p(x)\cdot u + O^{[p]}(|u|^2)\right)(1 + O(|u|^2))\mathrm{d}u = Vol(B_{r^+})(p(x) + O^{[p]}(r^2)),$$

where the odd-order term of u does not contribute to integral because B_{r^+} is a d-sphere, and $Vol(B_{r^+}) = v_d r^d(1 + c_M r^2)^d = m_0[h]r^d(1 + O(r^2))$. Thus,

$$I_r \leqslant I_+ = m_0[h]r^d(1 + O(r^2))(p(x) + O^{[p]}(r^2)) = m_0[h]r^d(p(x) + O^{[p]}(r^2)).$$

Similarly,

$$I_r \geqslant \int_{B_{r^-}} p(y(u))\left|\det(\frac{\mathrm{d}y}{\mathrm{d}u})\right|\mathrm{d}u = Vol(B_{r^-})(p(x) + O^{[p]}(r^2)) = m_0[h]r^d(p(x) + O^{[p]}(r^2)).$$

Putting together upper and lower bounds proves the lemma. □

A.3 *Other lemmas and proofs*

A.3.1 *Proofs of Lemmas 3.1 and 3.2*

REMARK A.1 The expansion of $G_\epsilon^{(\rho)} f$ for differentiable ρ was derived in [5, Appendix A.3], where duality was to analyze the 'right operator' $G_\epsilon^{(\rho)}$ by its 'left operator' which were defined in [5]. That bounds the error in the weak sense but not in the strong sense. Here we give a direct proof of a more precise bound of the error in the point-wise strong sense, which is important for analyzing the point-wise convergence of $L_N f(x)$.

Proof of Lemma 3.1. For a fixed $x \in \mathcal{M}$, define $\delta_{r_\epsilon}(x, y)$ as the following:

$$\frac{\|x-y\|^2}{\epsilon\rho(y)} = \frac{\|x-y\|^2}{\epsilon\rho(x)} + \frac{\|x-y\|^2}{\epsilon\rho(x)}\left(\frac{\rho(x)}{\rho(y)} - 1\right) =: \frac{\|x-y\|^2}{\epsilon\rho(x)} + \delta_{r_\epsilon}(x, y).$$

By that k_0 is C^4 on $(0, \infty)$, Taylor expansion up to the fourth order at $\frac{\|x-y\|^2}{\epsilon\rho(x)}$ gives

$$\begin{aligned} k_0\left(\frac{\|x-y\|^2}{\epsilon\rho(y)}\right) = k_0\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right) + k_0'\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right)\delta_{r_\epsilon}(x, y) + \frac{1}{2}k_0''\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right)\delta_{r_\epsilon}(x, y)^2 \\ + \frac{1}{6}k_0^{(3)}\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right)\delta_{r_\epsilon}(x, y)^3 + \frac{1}{24}k_0^{(4)}(\xi(x, y))\delta_{r_\epsilon}(x, y)^4 \end{aligned}$$

where $\xi(x, y)$ is between $\frac{\|x-y\|^2}{\epsilon\rho(y)}$ and $\frac{\|x-y\|^2}{\epsilon\rho(x)}$. Thus,

$$\begin{aligned} G_\epsilon^{(\rho)} f = \epsilon^{-d/2}\Bigg\{ & \int_{\mathcal{M}} k_0\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right) f(y)\,\mathrm{d}V(y) + \int_{\mathcal{M}} k_0'\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right)\delta_{r_\epsilon}(x, y) f(y)\,\mathrm{d}V(y) \\ & + \cdots + \frac{1}{24}\int_{\mathcal{M}} k_0^{(4)}(\xi(x, y))\delta_{r_\epsilon}(x, y)^4 f(y)\,\mathrm{d}V(y)\Bigg\} \\ := {} & ① + ② + ③ + ④ + ⑤. \end{aligned}$$

We first bound $|⑤|$. Because $\rho(x) < \rho_{max}$ uniformly on $\mathcal{M}$,

$$\frac{\|x-y\|^2}{\epsilon\rho(y)}, \frac{\|x-y\|^2}{\epsilon\rho(x)} \geqslant \frac{\|x-y\|^2}{\epsilon\rho_{max}}.$$

Thus,

$$|k_0^{(4)}(\xi(x, y))| \leqslant a_4 e^{-a\xi} \leqslant a_4 e^{-\frac{a}{\rho_{max}}\frac{\|x-y\|^2}{\epsilon}} = \bar{k}_4\left(\frac{\|x-y\|^2}{\epsilon}\right),$$

where we define

$$\bar{k}_4(r) := a_4 e^{-\frac{a}{\rho_{max}}r} \geqslant 0.$$

Note that $\bar{k}_4$ satisfies Assumption A.3, and $m_0[\bar{k}_4]$ are constant depending on ρ_{max}. Then

$$\begin{aligned}
24|⑤| &\leqslant \epsilon^{-d/2}\int_{\mathcal{M}} |k_0^{(4)}(\xi(x,y))||f(y)|\delta r_\epsilon(x,y)^4\,\mathrm{d}V(y)\\
&\leqslant \epsilon^{-d/2}\int_{\mathcal{M}} \bar{k}_4\left(\frac{\|x-y\|^2}{\epsilon}\right)|f(y)|\delta r_\epsilon(x,y)^4\,\mathrm{d}V(y)\\
&\leqslant \|f\|_\infty \epsilon^{-d/2}\int_{\mathcal{M}} \bar{k}_4\left(\frac{\|x-y\|^2}{\epsilon}\right)\left(\frac{\|x-y\|^2}{\epsilon}(\frac{1}{\rho(y)}-\frac{1}{\rho(x)})\right)^4\,\mathrm{d}V(y),
\end{aligned}$$

where we define $\tilde{k}(r) := \bar{k}_4(r)r^4$. By Lemma A.5,

$$\epsilon^{-d/2}\int_{\mathcal{M}} \tilde{k}\left(\frac{\|x-y\|^2}{\epsilon}\right)\left(\frac{1}{\rho(y)}-\frac{1}{\rho(x)}\right)^4\,\mathrm{d}V(y) = O^{[g^{(\leqslant 4)}]}(\epsilon^2),$$

where we denote $g(y) = \left(\frac{1}{\rho(y)}-\frac{1}{\rho(x)}\right)^4$, and then there is $c_5^\rho = c_5^\rho(\rho_{min},\rho_{max})$, such that $\sum_{l=0}^4 \|g^{(l)}\|_\infty \leqslant c_5^\rho(1+\sum_{l=1}^4 \|\mathbf{D}^{(l)}\rho^{-1}\|_\infty)$. This proves that

$$|⑤| \leqslant \epsilon^2\|f\|_\infty c_5^\rho\left(1+\sum_{l=1}^4 \|\mathbf{D}^{(l)}\rho^{-1}\|_\infty\right). \tag{A.58}$$

The other four terms involve fixed bandwidth $\epsilon\rho(x)$ where x is fixed. Applying Lemma A.5 gives the following. First,

$$\begin{aligned}
① &= \rho(x)^{\frac{d}{2}} G_{\epsilon\rho(x)} f(x) = \rho^{\frac{d}{2}}\left(m_0 f + \epsilon\rho\frac{m_2}{2}(\omega f + \Delta f) + O^{[f^{(\leqslant 4)}]}(\epsilon^2)\rho^2\right)\\
&=: ①_1 + O^{[f^{(\leqslant 4)}]}(\epsilon^2)\rho^{\frac{d}{2}+2}.
\end{aligned}$$

Define $k_1(r) := k_0'(r)r$ and $g_1(y) := (\frac{\rho(x)}{\rho(y)}-1)f(y)$. We have $g_1(x) = 0$, and

$$\begin{aligned}
② &= \epsilon^{-d/2}\int_{\mathcal{M}} k_0'\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right)\frac{\|x-y\|^2}{\epsilon\rho(x)}\left(\frac{\rho(x)}{\rho(y)}-1\right)f(y)\,\mathrm{d}V(y)\\
&= \rho(x)^{\frac{d}{2}} G_{\epsilon\rho(x)}[k_1](g_1)(x)\\
&= \rho^{\frac{d}{2}}\left(m_0[k_1]g_1 + \epsilon\rho\frac{m_2[k_1]}{2}(\omega g_1 + \Delta g_1) + O^{[g_1^{(\leqslant 4)}]}(\epsilon^2)\rho^2\right)\\
&= \rho^{\frac{d}{2}}\left(\epsilon\rho\frac{m_2[k_1]}{2}(\Delta g_1) + O^{[g_1^{(\leqslant 4)}]}(\epsilon^2)\rho^2\right)\\
&=: ②_1 + O^{[g_1^{(\leqslant 4)}]}(\epsilon^2)\rho^{\frac{d}{2}+2}.
\end{aligned}$$

Define $k_2(r) := k_0''(r)r^2$ and $g_2(y) := (\frac{\rho(x)}{\rho(y)} - 1)^2 f(y)$. We have that $g_2(x) = 0$ and

$$\begin{aligned}
③ &= \frac{1}{2}\epsilon^{-d/2}\int_{\mathcal{M}} k_0''\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right)\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}(\frac{\rho(x)}{\rho(y)}-1)\right)^2 f(y)\,\mathrm{d}V(y)\\
&= \frac{1}{2}\rho(x)^{\frac{d}{2}} G_{\epsilon\rho(x)}[k_2](g_2)(x)\\
&= \frac{1}{2}\rho^{\frac{d}{2}}\left(m_0[k_2]g_2 + \epsilon\rho\frac{m_2[k_2]}{2}(\omega g_2 + \Delta g_2) + O^{[g_2^{(\leqslant 4)}]}(\epsilon^2)\rho^2\right)\\
&= \frac{1}{2}\rho^{\frac{d}{2}}\left(\epsilon\rho\frac{m_2[k_2]}{2}(\Delta g_2) + O^{[g_2^{(\leqslant 4)}]}(\epsilon^2)\rho^2\right)\\
&=: ③_1 + O^{[g_2^{(\leqslant 4)}]}(\epsilon^2)\rho^{\frac{d}{2}+2}.
\end{aligned}$$

Define $k_3(r) = k_0^{(3)}(r)r^3$ and $g_3(y) = (\frac{\rho(x)}{\rho(y)} - 1)^3 f(y)$. Then, we have $g_3(x) = 0$, $\Delta g_3(x) = 0$, and

$$\begin{aligned}
④ &= \epsilon^{-d/2}\int_{\mathcal{M}} k_0^{(3)}\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}\right)\left(\frac{\|x-y\|^2}{\epsilon\rho(x)}(\frac{\rho(x)}{\rho(y)}-1)\right)^3 f(y)\,\mathrm{d}V(y)\\
&= \rho(x)^{\frac{d}{2}} G_{\epsilon\rho(x)}[k_3](g_3)(x)\\
&= \rho^{\frac{d}{2}}\left(m_0[k_3]g_3 + \epsilon\rho\frac{m_2[k_3]}{2}(\omega g_3 + \Delta g_3) + O^{[g_3^{(\leqslant 4)}]}(\epsilon^2)\rho^2\right)\\
&= O^{[g_3^{(\leqslant 4)}]}(\epsilon^2)\rho^{\frac{d}{2}+2}.
\end{aligned}$$

Collecting the leading terms, we have

$$①_1 + ②_1 + ③_1 = \rho^{\frac{d}{2}}\left(m_0 f + \epsilon\rho\frac{m_2}{2}(\omega f + \Delta f)) + \epsilon\rho\frac{m_2[k_1]}{2}(\Delta g_1) + \frac{1}{2}\epsilon\rho\frac{m_2[k_2]}{2}(\Delta g_2)\right).$$

Note that

$$\begin{aligned}
m_2[k_1] &= \frac{1}{d}\int_{\mathbb{R}^d} k_0'(|u|^2)|u|^4\mathrm{d}u = -\frac{d+2}{2}m_2[k_0],\\
m_2[k_2] &= \frac{1}{d}\int_{\mathbb{R}^d} k_0''(|u|^2)|u|^6\mathrm{d}u = -\frac{d+4}{2}m_2[k_1] = \frac{d+4}{2}\frac{d+2}{2}m_2[k_0],
\end{aligned}$$

$$\begin{aligned}
\Delta g_1 &= \rho f\Delta\frac{1}{\rho} + 2\rho\nabla f\cdot\nabla\frac{1}{\rho} = 2f\rho^{-2}|\nabla\rho|^2 - f\rho^{-1}\Delta\rho - 2\rho^{-1}\nabla f\cdot\nabla\rho,\\
\Delta g_2 &= 2f\rho^2|\nabla\frac{1}{\rho}|^2 = 2f\rho^{-2}|\nabla\rho|^2,
\end{aligned}$$

then we have

$$\begin{aligned}
&\frac{m_2}{2}\Delta f + \frac{m_2[k_1]}{2}\Delta g_1 + \frac{1}{2}\frac{m_2[k_2]}{2}\Delta g_2 \\
&= \frac{m_2}{2}\left(\Delta f - \frac{d+2}{2}(2f\rho^{-2}|\nabla\rho|^2 - f\rho^{-1}\Delta\rho - 2\rho^{-1}\nabla f\cdot\nabla\rho) + \frac{d+4}{2}\frac{d+2}{2}f\rho^{-2}|\nabla\rho|^2\right) \\
&= \frac{m_2}{2}\left(\Delta f + (\frac{d}{2}+1)(f\rho^{-1}\Delta\rho + 2\rho^{-1}\nabla f\cdot\nabla\rho) + \frac{d}{2}(\frac{d}{2}+1)f\rho^{-2}|\nabla\rho|^2\right) \\
&= \frac{m_2}{2}\rho^{-1-d/2}\Delta(f\rho^{1+d/2}),
\end{aligned}$$

and this proves that $①_1 + ②_1 + ③_1$ equals the leading term in (3.7).

To prove the lemma, it remains to specify the constants in

$$\rho^{\frac{d}{2}+2}\left(O^{[f^{(\leqslant 4)}]}(\epsilon^2) + O^{[g_1^{(\leqslant 4)}]}(\epsilon^2) + O^{[g_2^{(\leqslant 4)}]}(\epsilon^2) + O^{[g_3^{(\leqslant 4)}]}(\epsilon^2)\right) + |⑤|\,. \tag{A.59}$$

Observe the bound of $r_\epsilon^{(2)}$ in (3.7). By definition of g_1, g_2, g_3, there is $c_j^\rho(\rho_{min}, \rho_{max}) > 0, j = 1,2,3$, such that

$$\sum_{l=0}^{4}\|\mathbf{D}^{(l)}g_s\|_\infty \leqslant c_j^\rho\left(1 + \sum_{l=0}^{4}\|D^{(l)}f\|_\infty\right)\left(1 + \sum_{l=0}^{4}\|D^{(l)}\rho^{-1}\|_\infty\right), \quad j = 1,2,3,$$

and together with (A.58), the constant in front of ϵ^2 in (A.59) is bounded by

$$\begin{aligned}
&\rho_{max}^{d/2+2}\left\{\sum_{l=0}^{4}\|\mathbf{D}^{(l)}f\|_\infty + \left(\sum_{j=1}^{3}c_j^\rho + c_5^\rho\right)\left(1 + \sum_{l=0}^{4}\|D^{(l)}f\|_\infty\right)\left(1 + \sum_{l=0}^{4}\|D^{(l)}\rho^{-1}\|_\infty\right)\right\} \\
&= c^\rho\left(1 + \sum_{l=0}^{4}\|D^{(l)}f\|_\infty\right)\left(1 + \sum_{l=0}^{4}\|D^{(l)}\rho^{-1}\|_\infty\right),
\end{aligned}$$

where constant c^ρ equals a finite summation of certain powers and ratios of ρ_{min} and ρ_{max}. Thus, the bound in (3.7) holds. □

Proof of Lemma 3.2. Under the condition,

$$0.9\rho_{min} < 0.9\rho(x) < \tilde{\rho}(x) < 1.1\rho(x) < 1.1\rho_{max}, \quad \forall x \in \mathcal{M}. \tag{A.60}$$

By definition,

$$G_\epsilon^{(\tilde{\rho})}f(x) - G_\epsilon^{(\rho)}f(x) = \epsilon^{-d/2}\int_{\mathcal{M}}\left(k_0\left(\frac{\|x-y\|^2}{\epsilon\tilde{\rho}(y)}\right) - k_0\left(\frac{\|x-y\|^2}{\epsilon\rho(y)}\right)\right)f(y)\,\mathrm{d}V(y),$$

and

$$k_0\left(\frac{\|x-y\|^2}{\epsilon\tilde{\rho}(y)}\right) - k_0\left(\frac{\|x-y\|^2}{\epsilon\rho(y)}\right) = k_0'(\xi)\frac{\|x-y\|^2}{\epsilon\rho(y)}\left(\frac{\rho(y)}{\tilde{\rho}(y)} - 1\right),$$

where ξ is between $\frac{\|x-y\|^2}{\epsilon\tilde{\rho}(y)}$ and $\frac{\|x-y\|^2}{\epsilon\rho(y)}$. Then, by (A.60), $\xi \geqslant \frac{\|x-y\|^2}{\epsilon 1.1\rho(y)}$, and then by Assumption 3.1(C2),

$$|k_0'(\xi)| \leqslant a_1 e^{-a\xi} \leqslant a_1 e^{-\frac{a}{1.1}\frac{\|x-y\|^2}{\epsilon\rho(y)}}.$$

Thus, also by that

$$\left|\frac{\rho(y)}{\tilde{\rho}(y)} - 1\right| \leqslant \frac{\varepsilon\rho(y)}{0.9\rho(y)} = \frac{\varepsilon}{0.9},$$

we have that

$$|G_\epsilon^{(\tilde{\rho})}f(x) - G_\epsilon^{(\rho)}f(x)| \leqslant \epsilon^{-d/2}\int_{\mathcal{M}} |k_0'(\xi)|\frac{\|x-y\|^2}{\epsilon\rho(y)}\left|\frac{\rho(y)}{\tilde{\rho}(y)} - 1\right| |f(y)|\,\mathrm{d}V(y)$$

$$\leqslant \frac{\varepsilon}{0.9}\|f\|_\infty \epsilon^{-d/2}\int_{\mathcal{M}} a_1 e^{-\frac{a}{1.1}\frac{\|x-y\|^2}{\epsilon\rho(y)}}\frac{\|x-y\|^2}{\epsilon\rho(y)}\,\mathrm{d}V(y) = \frac{\varepsilon}{0.9}\|f\|_\infty G_\epsilon^{(\rho)}[k_1]\mathbf{1}(x),$$

where we define

$$k_1(r) := a_1 e^{-\frac{a}{1.1}r}r, \quad r \geqslant 0,$$

and k_1 satisfies Assumption 3.1. Next, applying Lemma 3.1 gives that

$$G_\epsilon^{(\rho)}[k_1]\mathbf{1}(x) = m_0[k_1]\rho^{\frac{d}{2}} + O^{[\rho]}(\epsilon).$$

Thus, with sufficiently small ϵ, uniformly for all x,

$$|G_\epsilon^{(\tilde{\rho})}f(x) - G_\epsilon^{(\rho)}f(x)| \leqslant \frac{\varepsilon}{0.9}\|f\|_\infty(m_0[k_1]\rho^{\frac{d}{2}} + O^{[\rho]}(\epsilon)) \leqslant c_\rho'\|f\|_\infty\varepsilon,$$

where c_ρ' is $O(1)$ constant depending on k_1 multiplied by certain powers of ρ_{max} or ρ_{min}. □

A.3.2 *Proofs of Lemmas A.1 and A.2*

Proof of Lemma A.1. Define

$$① := \epsilon^{-\frac{d}{2}}\int_{\mathcal{M}}\int_{\mathcal{M}} f(x)^2 k_0\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\hat{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\,\mathrm{d}V(x)\,\mathrm{d}V(y),$$

and

$$② := \epsilon^{-\frac{d}{2}}\int_{\mathcal{M}}\int_{\mathcal{M}} f(x)^2 k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\bar{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\,\mathrm{d}V(x)\,\mathrm{d}V(y).$$

Then, same as in the analysis of (A.16), applying Lemma 3.1 only to the $O(\epsilon)$ term gives that

$$② = \int (p\hat{\rho}^{\frac{d}{2}-\alpha})(y)G_{\epsilon\hat{\rho}(y)}^{(\bar{\rho})}(\frac{f^2 p}{\bar{\rho}^\alpha})(y)\,\mathrm{d}V(y)$$

$$= \int (p\hat{\rho}^{\frac{d}{2}-\alpha})\left(m_0[k_0]f^2 p\bar{\rho}^{\frac{d}{2}-\alpha} + \hat{\rho}r_1^{(1)}\right), \quad \|r_1^{(1)}\|_\infty = O^{[f,p]}(\epsilon).$$

By (A.14) and (A.15),

$$\left|\int p(\hat{\rho}^{\frac{d}{2}-\alpha} - \bar{\rho}^{\frac{d}{2}-\alpha})(m_0[k_0]f^2 p\bar{\rho}^{\frac{d}{2}-\alpha})\right| \leqslant O^{[f,p]}(\varepsilon_\rho),$$

and then

$$\left|\int p\hat{\rho}^{\frac{d}{2}-\alpha+1}r_1^{(1)}\right| \leqslant O^{[f,p]}(\epsilon)\int p\bar{\rho}^{\frac{d}{2}-\alpha+1}\max\{0.9^{\frac{d}{2}-\alpha+1}, 1.1^{\frac{d}{2}-\alpha+1}\} = O^{[f,p]}(\epsilon),$$

which gives that

$$② = m_0[k_0]\int p^2f^2\bar{\rho}^{d-2\alpha} + O^{[f,p]}(\epsilon, \varepsilon_\rho).$$

To bound $|② - ①|$, introduce

$$③ := \epsilon^{-\frac{d}{2}}\int_{\mathcal{M}}\int_{\mathcal{M}} f(x)^2k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\hat{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\,\mathrm{d}V(x)\,\mathrm{d}V(y)\,.$$

Then,

$$③ - ② = \epsilon^{-\frac{d}{2}}\int_{\mathcal{M}}\int_{\mathcal{M}} f(x)^2k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right)\frac{p(x)p(y)}{\bar{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\left(\frac{\bar{\rho}(x)^\alpha}{\hat{\rho}(x)^\alpha} - 1\right)\mathrm{d}V(x)\,\mathrm{d}V(y).$$

By non-negativity of k_0, p, $\bar{\rho}$, $\hat{\rho}$ and f^2, and ② $\geqslant 0$ and is $O^{[f,p]}(1)$, similar as in (A.20), we have

$$|③ - ②| \leqslant O^{[p]}(\varepsilon_\rho)② = O^{[f,p]}(\varepsilon_\rho).$$

This gives that $③ = m_0[k_0]\int p^2f^2\bar{\rho}^{d-2\alpha} + O^{[f,p]}(\epsilon, \varepsilon_\rho) = O^{[f,p]}(1)$. Meanwhile, ③ $\geqslant 0$ by definition. Finally, we have

$$③ - ① = \epsilon^{-\frac{d}{2}}\int_{\mathcal{M}}\int_{\mathcal{M}} f(x)^2\left(k_0\left(\frac{\|x-y\|^2}{\epsilon\bar{\rho}(x)\hat{\rho}(y)}\right) - k_0\left(\frac{\|x-y\|^2}{\epsilon\hat{\rho}(x)\hat{\rho}(y)}\right)\right)\frac{p(x)p(y)}{\hat{\rho}(x)^\alpha\hat{\rho}(y)^\alpha}\,\mathrm{d}V(x)\,\mathrm{d}V(y)\,.$$

And same as in (A.23), by introducing $k_1(r)$ and using the non-negativity of f^2, p and $\hat{\rho}$, one can show that

$$|③ - ①| \leqslant O^{[p]}(\varepsilon_\rho)\cdot(③ \text{ with } k_1) = O^{[f,p]}(\varepsilon_\rho).$$

Putting together, we have $|② - ①| = O^{[f,p]}(\varepsilon_\rho)$, and this proves the lemma. □

Proof of Lemma A.2. By definition,

$$G_\epsilon^{(\tilde{\rho})}\left(\frac{f}{\tilde{\rho}^\alpha} - \frac{f}{\rho^\alpha}\right)(x) = \epsilon^{-d/2}\int_{\mathcal{M}} k_0\left(\frac{\|x-y\|^2}{\epsilon\tilde{\rho}(y)}\right)\frac{f(y)}{\rho(y)^\alpha}\left(\frac{\rho(y)^\alpha}{\tilde{\rho}(y)^\alpha} - 1\right)\mathrm{d}V(y).$$

To proceed, by that $\sup_{x\in\mathcal{M}}\frac{|\tilde{\rho}(x)-\rho(x)|}{\rho(x)} < \varepsilon < 0.1$ and (A.60), for some $c_{\rho,1}$ and $c_{\rho,2}$ equaling certain powers of ρ_{max} or ρ_{min} multiplied by $\Theta^{[\alpha]}(1)$ constant, we have

$$\left|\frac{\rho(x)^\alpha}{\tilde{\rho}(x)^\alpha} - 1\right| \leqslant c_{\rho,1}\varepsilon, \quad \frac{1}{\rho(x)^\alpha} \leqslant c_{\rho,2}, \quad \forall x \in \mathcal{M}.$$

Then,

$$\left|G_\epsilon^{(\tilde{\rho})}\left(\frac{f}{\tilde{\rho}^\alpha} - \frac{f}{\rho^\alpha}\right)(x)\right| \leqslant c_{\rho,1}\varepsilon\cdot c_{\rho,2}\|f\|_\infty\left(\epsilon^{-d/2}\int_{\mathcal{M}} k_0\left(\frac{\|x-y\|^2}{\epsilon\tilde{\rho}(y)}\right)\mathrm{d}V(y)\right)$$
$$= c_{\rho,1}\varepsilon\cdot c_{\rho,2}\|f\|_\infty G_\epsilon^{(\tilde{\rho})}\mathbf{1}(x).$$

By Lemmas 3.1 and 3.2,

$$|G_\epsilon^{(\tilde{\rho})}\mathbf{1}(x) - m_0\rho^{\frac{d}{2}}(x)| \leqslant O^{[\rho]}(\epsilon) + c'_\rho\varepsilon,$$

this proves the lemma with $c''_\rho = \Theta^{[1]}(c_{\rho,1}c_{\rho,2}(m_0\rho_{max}^{d/2} + 0.1c'_\rho))$ when ϵ gets sufficiently small. □

A.4 *Point-wise convergence to* $\mathcal{L}_{\hat{\rho}}^{(\alpha)}$

In parallel to Theorems 3.5 and 3.6, we show the point-wise convergence of $L_N f(x)$ to another limiting operator involving $\mathcal{L}_{\hat{\rho}}^{(\alpha)}$, in Theorems A.5 and A.6. The analysis is by adopting the approach in [**5**] after conditioning on a fixed $\hat{\rho}$, yet the difficulty is to handle the a.s. differentiability of the kNN-estimated $\hat{\rho}$.

As pointed out by Section 2.3, $\hat{\rho}(x)$ at any point of differentiability equals $\Theta((k/N_y)^{-1/d})$ which diverges to ∞ asymptotically. We first derive a lemma to bound the derivatives of $\hat{\rho}(x)$ by certain inverse powers of $\hat{R}(x)$. The proof of Lemma 2.1 shows that, when Y has distinct points, the estimated $\hat{\rho}$ from Y is piecewise C^∞ on $\mathbb{R}^D$, and it has the structure that on each of the finitely many polygon $\mathbf{p}$, $\hat{\rho}(x) = (\frac{1}{m_0[h]}\frac{k}{N_y})^{-1/d}\|x - y_{\mathbf{p}}\|$, for a some point $y_{\mathbf{p}}$ outside $\mathbf{p}$. We then can upper bound the derivatives of $\hat{R}$ as below.

LEMMA A.7 Under the condition of Lemma 2.1, for any $x \in \mathbb{R}^D \backslash E$,

$$|\mathbf{D}^{(l)}\hat{R}(x)| \leqslant (l!)\hat{R}(x)^{-l+1}, \quad l = 0, 1, \cdots, 4,$$

where the lth derivative $\mathbf{D}^{(l)}\hat{R}(x)$ is an l-way tensor, and for any l-way tensor $\mathbf{T} : \mathbb{R}^D \times \cdots \times \mathbb{R}^D \to \mathbb{R}$,

$$|\mathbf{T}| = \sup_{v\in\mathbb{R}^D,\, \|v\|_2\leqslant 1} |T(v, \cdots, v)|.$$

The claim extends to higher-order derivatives $l > 4$, and we only need up to the fourth derivative in the diffusion kernel analysis. The piecewise architecture of $\hat{R}$ also allows us to construct smooth uniform approximators to $\hat{\rho}$ without enlarging the derivatives.

LEMMA A.8 Under the condition of Lemma 2.1, for any $s > 0$, $\exists\hat{\rho}_s \in C^\infty(\mathcal{M})$ s.t. $\sup_{x\in\mathcal{M}} |\hat{\rho}_s(x) - \hat{\rho}(x)| < s$, and

$$\sup_{x\in\mathcal{M}} |D_{\mathcal{M}}^{(l)}\hat{\rho}_s(x)| \leqslant \sup_{x\in\mathcal{M}\backslash E} |D_{\mathcal{M}}^{(l)}\hat{\rho}(x)|, \quad l = 0, 1, \cdots, 4. \tag{A.61}$$

Combined with Lemma A.7 and (A.14), we have the following:

PROPOSITION A.4 There is a constant $C_p > 0$ depending on $(\mathcal{M}, p)$ such that when $\sup_{x\in\mathcal{M}} |\hat{\rho} - \bar{\rho}|/\bar{\rho} < \varepsilon_\rho < 0.1$, for any $s > 0$, $\exists\hat{\rho}_s \in C^\infty(\mathcal{M})$ s.t. $\sup_{x\in\mathcal{M}} |\hat{\rho}_s(x) - \hat{\rho}(x)| < s$, and

$$\|D_{\mathcal{M}}^{(l)}\hat{\rho}_s\|_{\infty,\mathcal{M}} \leqslant C_p\left(\left(\frac{k}{N_y}\right)^{-l/d}\right), \quad l = 0, \cdots 4. \tag{A.62}$$

Because we can make s arbitrarily small, it is equivalent to prove the graph Laplacian convergence with $\hat{\rho}_s$, which satisfies (A.62) by the proposition and also (A.14) by the uniform approximation Theorem 2.3. Below, we write $\hat{\rho}_s$ as $\hat{\rho}$. We then have the following:

THEOREM A.5 Suppose Theorem 2.3 holds and as $N_y \to \infty$ and $N_x \to \infty$,

$$\epsilon = o(1), \quad \epsilon^{d/2+1} N_x = \Omega(\log N_x), \quad \epsilon = o\left((\frac{k}{N_y})^{4/d} \right).$$

Then, for any $f \in C^\infty(\mathcal{M})$, for sufficiently large N_x and N_y, w.p. higher than $1 - 4N_x^{-10} - 2N_y^{-10}$,

$$L^{(\alpha)}_{rw'} f(x) = \mathcal{L}^{(\alpha)}_{\hat{\rho}} f(x) + O^{[f,p]}\left(\left(\frac{k_y}{N_y} \right)^{-4/d} \epsilon \right) + O^{[1]}\left(|\nabla f(x)| p(x)^{1/d} \sqrt{\frac{\log N}{N\epsilon^{d/2+1}}} \right).$$

THEOREM A.6 Under the same setting as in Theorem (A.5) and in the same sense of w.h.p,

$$L^{(\alpha)}_{un} f(x) = (p\hat{\rho}^{d+2-2\alpha})(x) \mathcal{L}^{(\alpha)}_{\hat{\rho}} f(x) + O^{[f,p]}\left(\left(\frac{k_y}{N_y} \right)^{-4/d} \epsilon \right) + O^{[1]}\left(|\nabla f(x)| p(x)^{\frac{\alpha-1}{d}} \sqrt{\frac{\log N}{N\epsilon^{d/2+1}}} \right).$$

In both theorems, the error rates can be worse than those in Theorems 3.5 and 3.6. It also gives different optimal scaling when choosing ϵ and k so as to balance the bias and variance errors there. The reason is due to that the bounds of magnitudes of derivatives of $\hat{\rho}$ are scaled with inverse powers of $(k/N)^{1/d}$.

The proofs of Theorem A.5 and A.6 are basically the same as those of Theorems 3.5 and 3.6. The difference is replacing the usage of Lemma 3.2 by a vanilla application of Lemma 3.1 with ρ being $\hat{\rho}$ (which is $\hat{\rho}_s$), and details are omitted.

Proof of Lemma A.7. Note that for any $x \in \mathbb{R}^D \backslash E$, as shown in the proof of Lemma 2.1, x is in a polygon $\mathbf{p}$, and $\hat{R}(x) = \|x - y_{\mathbf{p}}\|$ for some $y_{\mathbf{p}}$ outside $\mathbf{p}$. For $l = 0$, the claim is identity. For $l = 1$, $\nabla r(x) = \frac{x}{\|x\|}$ and $|\nabla r(x)| = 1$. When $l = 2$,

$$\mathbf{D}^{(2)} r(x) = \frac{\|x\|^2 I_d - xx^T}{\|x\|^3}.$$

Thus, for any $v \in \mathbb{R}^D$, $\|v\| = 1$,

$$|\mathbf{D}^{(2)} r(x)(v,v)| = \frac{|\|v\|^2 \|x\|^2 - (v^T x)^2|}{\|x\|^3} \leqslant \frac{\|v\|^2 \|x\|^2}{\|x\|^3} = \frac{1}{\|x\|},$$

and hence $|\mathbf{D}^{(2)} r(x)| \leqslant \frac{2}{\|x\|}$. When $l = 3$ and 4, one can verify by definition that

$$|\mathbf{D}^{(3)} r(x)| < \frac{6}{\|x\|^2}, \quad |\mathbf{D}^{(4)} r(x)| < \frac{6 \times 4}{\|x\|^3}.$$

This proves that $|\mathbf{D}^{(l)} \hat{R}(x)| \leqslant \frac{l!}{\hat{R}(x)^{l-1}}$, for $l = 1, \cdots, 4$. □

Proof of Lemma A.8. Because $\hat{\rho} = (\frac{1}{m_0[h]} \frac{k}{N_y})^{-1/d} \hat{R}$, we consider the smooth approximation of $\hat{R}$ called $\hat{R}_s$, and let $\hat{\rho}_s = (\frac{1}{m_0[h]} \frac{k}{N_y})^{-1/d} \hat{R}_s$. By Lemma 2.1 and its proof, $\hat{R}$ is continuous on $\mathbb{R}^D$ and C^∞ on $\mathbb{R}^D \backslash E$ which is a finite union of (possibly unbounded) polygons. We consider the restriction of $\hat{R}$ on $\mathcal{M}$, and because $\mathcal{M}$ is C^∞, $\hat{R}$ is C^∞ on $\mathcal{M} \backslash E$. Under the probability assumption on p in Assumption 2.1, the set E intersects with $\mathcal{M}$ over finitely many hypersurfaces of dimensionality $(d-1)$ w.p. 1. Then, there is a finite partition of $\mathcal{M}$ into pieces with piecewise C^∞ boundaries. For $s > 0$, a function $\hat{R}_s \in C^\infty(\mathcal{M})$ can

be constructed to uniformly approximate $\hat{R}$ on $\mathcal{M}$ to within s, and in addition, $D^{(l)}_{\mathcal{M}}\hat{R}_s(x)$ for $l = 0, \cdots, 4$ is smoothly averaged from the values of $D^{(l)}_{\mathcal{M}}\hat{R}(x)$ on a neighborhood of x. This means that at $x \in \mathcal{M} \cap E$, $D^{(l)}_{\mathcal{M}}\hat{R}_s(x)$ is smoothly interpolating between the values of $D^{(l)}_{\mathcal{M}}\hat{R}$ on each sides of the hypersurface (at intersection of multiple hypersurfaces, i.e. 'corners', $D^{(l)}_{\mathcal{M}}\hat{R}_s(x)$ is interpolating among the multiple values). Then whether x is near $\mathcal{M} \cap E$ or not, we have that $|D^{(l)}_{\mathcal{M}}\hat{R}_s(x)| \leqslant \sup_{x\in\mathcal{M}\setminus E} |D^{(l)}_{\mathcal{M}}\hat{R}(x)|$. This can be done, e.g., by convolving $\hat{R}$ on $\mathcal{M}$ using a Gaussian kernel in $\mathbb{R}^d$ with small bandwidth and under the manifold metric. The uniform approximation of $|\hat{R}_s(x) - \hat{R}(x)|$ is then guaranteed by that $\hat{R}$ is Lipschitz-1 on $\mathbb{R}^D$ and thus is globally Lipschitz-1 on $\mathcal{M}$ with respect to the manifold geodesic metric. This proves (A.61). □

Proof of Proposition A.4. Under the good event in Theorem 2.3, (A.14) equivalently gives that

$$\sup_{x\in\mathcal{M}} \frac{|\hat{R}(x) - \bar{R}(x)|}{\bar{R}(x)} < \varepsilon_\rho < 0.1. \tag{A.63}$$

Meanwhile, $\bar{\rho}(x) = p(x)^{-1/d}$ and is uniformly bounded from below and above by ρ_{min} and ρ_{max} which are constants depending on p. We write $m_0[h]$ as m_0 in this proof.

$$\bar{R}(x) = \left(\frac{1}{m_0}\frac{k}{N_y}\right)^{1/d} \bar{\rho}(x) \in \left[c_{p,1}\left(\frac{k}{N_y}\right)^{1/d}, c_{p,2}\left(\frac{k}{N_y}\right)^{1/d}\right]. \tag{A.64}$$

Lemma A.7 gives that, for $l = 0, 1, \cdots, 4$, $\mathbf{D}^{(l)}$ being the derivatives in $\mathbb{R}^D$,

$$|\mathbf{D}^{(l)}\hat{R}(x)| \leqslant \frac{l!}{\hat{R}(x)^{l-1}}, \quad \forall x \in \mathcal{M}\setminus E.$$

The manifold derivatives are determined by ambient space derivatives via

$$D^{(l)}_{\mathcal{M}}\hat{R}(x) = \sum_{m=0}^{l} A_m(x)(\mathbf{D}^{(m)}\hat{R}(x)),$$

where $A_m(x)$ are linear transforms determined by extrinsic manifold coordinates and their derivatives at x, and are in $C^\infty(\mathcal{M})$. Thus,

$$|D^{(l)}_{\mathcal{M}}\hat{R}(x)| \leqslant \sum_{m=0}^{l} |A_m(x)||\mathbf{D}^{(m)}\hat{R}(x)| \leqslant c_{\mathcal{M}} \sum_{m=0}^{l} \frac{m!}{\hat{R}(x)^{m-1}},$$

where $c_{\mathcal{M}}$ is a constant depending on $\mathcal{M}$. This gives that, $\forall x \in \mathcal{M}\backslash E$,

$$|D^{(l)}_{\mathcal{M}}\hat{\rho}(x)| = \left(\frac{1}{m_0}\frac{k}{N_y}\right)^{-1/d} |D^{(l)}_{\mathcal{M}}\hat{R}| \leqslant \left(\frac{1}{m_0}\frac{k}{N_y}\right)^{-1/d} c_{\mathcal{M}} \sum_{m=0}^{l} \frac{m!}{\hat{R}(x)^{m-1}}$$

$$\leqslant c'_{\mathcal{M}} \frac{c_{p,2}}{\bar{R}(x)} \sum_{m=0}^{l} \hat{R}(x)^{-m+1} \quad \text{(by (A.64), } c'_{\mathcal{M}} \text{ depending on } \mathcal{M})$$

$$< c'_{\mathcal{M}} c_{p,2} 1.1 \sum_{m=0}^{l} \hat{R}(x)^{-m} \quad \text{(by (A.63))}$$

$$\leqslant c'_{\mathcal{M}} c_{p,2} 1.1 \sum_{m=0}^{l} (0.9\bar{R}(x))^{-m} \leqslant c'_{\mathcal{M}} c_{p,2} 1.1 \sum_{m=0}^{l} \left(0.9 c_{p,1} (\frac{k}{N_y})^{1/d}\right)^{-m},$$

which means that

$$\sup_{x\in\mathcal{M}\backslash E} |D^{(l)}_{\mathcal{M}}\hat{\rho}(x)| \leqslant C_p \left(\frac{k}{N_y}\right)^{-l/d},$$

where C_p is a constant depending on p, for l up to 4. Finally, Lemma A.8 constructs $\hat{\rho}_s$ satisfying (A.61), and then (A.62) follows. □