The Annals of Statistics
2018, Vol. 46, No. 6B, 3805–3837
https://doi.org/10.1214/17-AOS1676

THINK GLOBALLY, FIT LOCALLY UNDER THE MANIFOLD SETUP: ASYMPTOTIC ANALYSIS OF LOCALLY LINEAR EMBEDDING

BY HAU-TIENG WU[*,†,1] AND NAN WU[‡]

Duke University[*], National Center for Theoretical Sciences[†] and University of Toronto[‡]

Since its introduction in 2000, Locally Linear Embedding (LLE) has been widely applied in data science. We provide an asymptotical analysis of LLE under the manifold setup. We show that for a general manifold, asymptotically we may not obtain the Laplace–Beltrami operator, and the result may depend on nonuniform sampling unless a correct regularization is chosen. We also derive the corresponding kernel function, which indicates that LLE is not a Markov process. A comparison with other commonly applied nonlinear algorithms, particularly a diffusion map, is provided and its relationship with locally linear regression is also discussed.

1. Introduction. Dimension reduction is a fundamental step in data analysis. In past decades, due to the high demand for analyzing the large scale, massive and complicated datasets accompanying technological advances, there have been many efforts to solve this problem from different angles. The resulting algorithms can be roughly classified into two types: linear and nonlinear. Linear methods include Principal Component Analysis (PCA), multidimensional scaling and others. Nonlinear methods include ISOMAP [29], Locally Linear Embedding (LLE) [22] and its variations such as Hessian LLE [12] and modified LLE [35], eigenmap [1], Diffusion Map (DM) [8], local tangent space alignment [36], vector diffusion map [24, 26], horizontal diffusion map [16], maximal variance unfolding [33] and t-distributed stochastic neighbor embedding [30] to name a few.

The subject of this paper, LLE, was published in *Science* in 2000 [22]. According to the Google Scholar, it had been cited almost 10,000 times as of mid-January, 2017. The algorithm is designed to be intuitive and simple. It has also been found to be efficient and practical. It contains two main parts. The first part is to determine the nearest neighbors of each data point, and catch the local geometric structure of the dataset through finding the barycenter coordinate for those neighboring points using a regularization. This is the "fit locally" part of LLE. Second, by viewing the

Received March 2017; revised December 2017.

[1]Supported in part by Sloan Research Fellow FR-2015-65363. He acknowledges the continuous support from the Department of Mathematics, University of Toronto.

MSC2010 subject classifications. 60K35.

Key words and phrases. Locally linear embedding, diffusion maps, dimension reduction, locally linear regression, measurement error.

barycenter coordinates as the "weights" for the neighboring points, the eigenvectors and eigenvalues of the associated "affinity matrix" are evaluated to organize the data points. This is the "think globally" part of LLE. However, unlike the fruitful theoretical results from discussing a diffusion-based approach like DM [2, 3, 8, 14, 17–19, 23, 26, 27, 31, 32], to the best of our knowledge, no systematic analysis of LLE has been undertaken, except an ad hoc argument shown in [1] based on some conditions.

The main contribution of this paper is to analyze the "fit locally" part of LLE. Based on a careful analysis of the barycentric coordinate by the covariance analysis, we provide an asymptotic *pointwise convergence* analysis of LLE under the manifold setup. Although it is widely believed that under the manifold setup, asymptotically LLE should lead to the Laplace–Beltrami operator; in this paper, we show that this might not always be the case. It fundamentally depends on the geometric structure of the data set. Specifically, under the assumption that the point cloud is (non)uniformly sampled from a low dimensional manifold isometrically embedded in the Euclidean space, we show that the asymptotical behavior of LLE depends on its regularization. If the regularization is chosen properly, we obtain the Laplace–Baltrami operator, even if the sampling is nonuniform. If the regularization is not chosen properly, the acquired information will be contaminated by the extrinsic information (the second fundamental form), and we even obtain a fourth-order differential operator in some extreme cases. To catch this dependence on the extrinsic information, we carefully analyze the "local covariance structure" of the dataset up to a higher order term. One key step toward the analysis is to establish the kernel function associated with LLE that comes from the barycentric coordinate estimation. Via the established kernel function, we have a direct comparison of LLE and other relevant nonlinear machine learning algorithms, such as eigenmap and DM. Unlike eigenmap or DM, LLE in general is not a diffusion process on the dataset, and the convergence rate might be different, depending on the regularization used. In the end, we link LLE back to the widely applied kernel regression technique, Locally Linear Regression (LLR) and the measurement error problem. While it is not explored in this paper, we mention that based on the established pointwise convergence, we could further understand the "think globally" part of LLE from the spectral geometry viewpoint [4, 5].

The paper is organized as follows: In Section 2, we review LLE. In Section 3, we provide the asymptotical analysis of LLE under the manifold setup. In Section 4, we provide numerical simulations to support our theoretical findings. The relationship between two common nearest neighbor search schemes is discussed in Section 5. The relationship between LLE, LLR and the shrinkage scheme for the high dimensional covariance matrix are discussed in Section 6. The discussion is shown in Section 7. The technical proofs of the theorems are included in the online supplementary information (SI) [34]. The perturbation argument of the eigenvalues and eigenvectors of a symmetric matrix is summarized in Section SI.1. The statement of technical lemmas for the proof is given in Section SI.2. The covariance

TABLE 1
Commonly used notation in this paper

Symbol	Meaning
p	Dimension of the ambient space
d	Dimension of the low dimensional Riemannian manifold
(M, g)	d-dimensional smooth Riemannian manifold
dV	Riemannian volume form of (M, g)
$\exp_x$	Exponential map at x
$T_x M$	Tangent space of M at x
$\mathtt{Ric}_x$	Ricci curvature tensor of (M, g) at x
ι, ι_*	Isometric embedding of M into $\mathbb{R}^p$ and its differential
$\mathbb{I}_x$	Second fundamental form of the embedding ι at x
P	Probability density function on $\iota(M)$
$n \in \mathbb{N}$	Number of data points sampled from M
$\mathcal{X} = \{z_i\}_{i=1}^n$	Point cloud sampled from $\iota(M) \subset \mathbb{R}^p$
$w_{z_k} \in \mathbb{R}^N$	Barycentric coordinates of z_k with respect to data points in the ε-neighborhood

structure analysis is provided in Section SI.3. The proofs of the main theorems are given in Appendices SI.4 and SI.5. The technical lemmas for the theorems are given in Section SI.6.

Here, we fix the notation used in this paper. For $d \in \mathbb{N}$, $I_{d\times d}$ means the identity matrix of size $d \times d$. For $n \in \mathbb{N}$, denote $\mathbf{1}_n$ to be the n-dim vector with all entries 1. For $\varepsilon \geq 0$, denote $B_\varepsilon^{\mathbb{R}^p}(x) := \{y \in \mathbb{R}^p | \|x - y\|_{\mathbb{R}^p} \leq \varepsilon\}$. Denote $e_i = [0, \ldots, 1, \ldots 0]^\top \in \mathbb{R}^p$ to be the unit p-dim vector with 1 in the ith entry. For $p, r \in \mathbb{N}$ so that $r \leq p$, denote $J_{p,r} \in \mathbb{R}^{p\times r}$ so that the (i, i) entry is 1 for $i = 1, \ldots, r$, and zeros elsewhere and denote $\bar{J}_{p,r} \in \mathbb{R}^{p\times r}$ so that the $(p - r + i, i)$ entry is 1 for $i = 1, \ldots, r$, and zeros elsewhere. $I_{p,r} := J_{p,r} J_{p,r}^\top$ is a $p \times p$ matrix so that the (i, i)th entry is 1 for $i = 1, \ldots, r$ and 0 elsewhere; and $\bar{I}_{p,r} := \bar{J}_{p,r} \bar{J}_{p,r}^\top$ is a $p \times p$ matrix so that the (i, i)th entry is 1 for $i = p - r + 1, \ldots, p$ and 0 elsewhere. Denote $S(p)$ to be the set of a real symmetric matrix of size $p \times p$, $O(p)$ to be the orthogonal group in dimension p, and $\mathfrak{o}(p)$ to be the set of anti-symmetric matrix of size $p \times p$. For $M \in \mathbb{R}^{p\times p}$, denote $M^\top$ to be the transpose of M and $M^\dagger$ to be the Moore–Penrose pseudo-inverse of M. For $a, b \in \mathbb{R}$, we use $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$ to simplify the notation. We summarize the commonly used notation for the asymptotical analysis in Table 1 for the convenience of the readers.

2. Review of locally linear embedding. We start by summarizing LLE. Suppose $\mathcal{X} = \{z_i\}_{i=1}^n \subset \mathbb{R}^p$ is the provided dataset, or the point cloud:

1. Fix $\varepsilon > 0$. For each $z_k \in \mathcal{X}$, denote $\mathcal{N}_{z_k} := B_\varepsilon^{\mathbb{R}^p}(z_k) \cap (\mathcal{X} \setminus \{z_k\}) = \{z_{k,j}\}_{j=1}^{n_k}$, where $n_k \in \mathbb{N}$ is the number of points in $\mathcal{N}_{z_k}$. $\mathcal{N}_{z_k}$ is called the ε-radius neighborhood of z_k. Alternatively, we can also fix a number K, and choose the K nearest

points of z_k. This is called the K-nearest neighbors (KNN) scheme. While the ε-radius neighborhood scheme and the KNN scheme are closely related, they are different. In this paper, we study LLE with the ε-radius neighborhood scheme, and postpone the discussion of the relationship between these two schemes to Section 5.

2. For each $z_k \in \mathcal{X}$, find its barycentric coordinate associated with $\mathcal{N}_{z_k}$ by

$$w_{z_k} = \operatorname*{argmin}_{w \in \mathbb{R}^{n_k}, w^\top \mathbf{1}_{n_k} = 1} \left\| z_k - \sum_{j=1}^{n_k} w(j) z_{k,j} \right\|^2 \in \mathbb{R}^{n_k}. \tag{2.1}$$

Notice that w_{z_k} satisfies $w_{z_k}^\top \mathbf{1}_{n_k} = \sum_{j=1}^{n_k} w_{z_k}(j) = 1$.

3. Define a $n \times n$ matrix W, called the *LLE matrix*, by

$$W_{k,l} = \begin{cases} w_{z_k}(j) & \text{if } z_l = z_{k,j} \in \mathcal{N}_{z_k}; \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

4. To reduce the dimension of $\mathcal{X}$, it is suggested in [22] to embed $\mathcal{X}$ into a low dimension Euclidean space

$$z_k \mapsto Y_k = \big[v_1(k), \ldots, v_\ell(k)\big]^\top \in \mathbb{R}^\ell, \tag{2.3}$$

for each $z_k \in \mathcal{X}$, where ℓ is the dimension of the embedded points chosen by the user, and $v_1, \ldots, v_\ell \in \mathbb{R}^n$ are eigenvectors of $(I - W)^\top (I - W)$ corresponding to the ℓ smallest eigenvalues. Note that this is equivalent to minimizing the cost function $\sum_{k=1}^n \|Y_k - \sum_{l=1}^n W_{k,l} Y_l\|^2$, where $Y = [Y_1, \ldots, Y_n] \in \mathbb{R}^{\ell \times n}$, subject to the constraint $YY^\top = I_{\ell \times \ell}$.

Although the algorithm looks relatively simple, there are actually several details that should be discussed prior to the asymptotical analysis. To simplify the discussion, we focus on one point $z_k \in \mathcal{X}$ and assume that there are N data points in $\mathcal{N}_{z_k} = \{z_{k,1}, \ldots, z_{k,N}\}$. To find the barycentric coordinate of z_k, we define the *local data matrix* associated with $\mathcal{N}_{z_k}$:

$$G_n := \begin{bmatrix} | & & | \\ z_{k,1} - z_k & \cdots & z_{k,N} - z_k \\ | & & | \end{bmatrix} \in \mathbb{R}^{p \times N}. \tag{2.4}$$

It is important to note that G_n depends not only on n, but also ε and z_k. However, we only keep n to make the notation easier. The other notation in this section are simplified in the same way. Minimizing (2.1) is equivalent to minimizing the functional $w^\top G_n^\top G_n w$ over $w \in \mathbb{R}^N$ under the constraint $w^\top \mathbf{1}_N = 1$. Here, $G_n^\top G_n$ is the Gramian matrix associated with the dataset $\{z_{k,1} - z_k, \ldots, z_{k,N} - z_k\}$. In general, $G_n^\top G_n$ might be singular, and it is suggested in [22] to stabilize the algorithm through regularizing the equation by

$$\big(G_n^\top G_n + c I_{N \times N}\big) y = \mathbf{1}_N, \tag{2.5}$$

where $c > 0$ is the regularizer chosen by the user. For example, in [22], c is suggested to be $\frac{\delta}{N}$, where $0 < \delta < \|G_n\|_F^2$ is chosen by the user and $\|G_n\|_F$ is the Frobenius norm of G_n. It has been observed that LLE is sensitive to the choice of the regularizer (see, e.g., [35]). We will later quantify this dependence under the manifold setup. Using the Lagrange multiplier method, the minimizer is

$$w_n = \frac{y_n}{y_n^\top \mathbf{1}_N}, \tag{2.6}$$

where y_n is the solution of (2.5). We will consider the regularized equation (2.5) in the following discussion.

Next, we explicitly express w_n, which is the essential step toward the asymptotical analysis. Suppose $\mathtt{rank}(G_n^\top G_n) = r_n$. Note that $r_n = \mathtt{rank}(G_n G_n^\top) = \mathtt{rank}(G_n) \le p$, so $G_n^\top G_n$ is singular when $p < N$. Moreover, $G_n^\top G_n$ is positive semidefinite. Denote the eigendecomposition of $G_n^\top G_n$ as $V_n \Lambda_n V_n^\top$, where

$$\Lambda_n = \mathtt{diag}(\lambda_{n,1}, \lambda_{n,2}, \ldots, \lambda_{n,N}), \tag{2.7}$$

$\lambda_{n,1} \ge \lambda_{n,2} \ge \cdots \ge \lambda_{n,r_n} > \lambda_{n,r_n+1} = \cdots = \lambda_{n,N} = 0$, and

$$V_n = \begin{bmatrix} | & & | \\ v_{n,1} & \ldots & v_{n,N} \\ | & & | \end{bmatrix} \in O(N). \tag{2.8}$$

Clearly, $\{v_{n,i}\}_{i=r_n+1}^N$ forms an orthonormal basis of the null space of $\mathtt{Null}(G_n^\top G_n)$, which is equivalent to $\mathtt{Null}(G_n)$. Then (2.5) is equivalent to solving

$$V_n(\Lambda_n + cI_{N\times N})V_n^\top y = \mathbf{1}_N, \tag{2.9}$$

and the solution is

$$\begin{aligned} y_n &= V_n(\Lambda_n + cI_{N\times N})^{-1} V_n^\top \mathbf{1}_N \\ &= c^{-1}\mathbf{1}_N + V_n\big[(\Lambda_n + cI_{N\times N})^{-1} - c^{-1}I_{N\times N}\big]V_n^\top \mathbf{1}_N. \end{aligned} \tag{2.10}$$

Therefore,

$$w_n^\top = \frac{\mathbf{1}_N^\top + \mathbf{1}_N^\top V_n[c(\Lambda_n + cI_{N\times N})^{-1} - I_{N\times N}]V_n^\top}{N + \mathbf{1}_N^\top V_n[c(\Lambda_n + cI_{N\times N})^{-1} - I_{N\times N}]V_n^\top \mathbf{1}_N}. \tag{2.11}$$

Without recasting (2.11) into a proper form, it is not clear how to capture the geometric information contained in (2.11). Observe that while $G_n^\top G_n$ is the Gramian matrix, $G_n G_n^\top$ is related to the sample covariance matrix associated with $\mathcal{N}_{z_k}$. We call $\frac{1}{n}G_n G_n^\top$ the *local sample covariance matrix*. Note that this local sample covariance matrix is different from the usual sample covariance matrix associated with $\mathcal{N}_{z_k}$, which is defined as $\frac{1}{n-1}\sum_{j=1}^N (z_{k,j} - \mu_k)(z_{k,j} - \mu_k)^\top$, where $\mu_k = \frac{1}{n}\sum_{j=1}^N z_{k,j}$. Clearly, $r_n \le p$ and $G_n G_n^\top$ and $G_n^\top G_n$ share the same positive

eigenvalues, $\lambda_{n,1} \cdots \lambda_{n,r_n}$. Denote the eigendecomposition of $G_n G_n^\top$ as $U_n \bar{\Lambda}_n U_n^\top$, where $U_n \in O(p)$ and $\bar{\Lambda}_n$ is a $p \times p$ diagonal matrix. By direct calculation, the first r_n columns of V_n are related to U_n by

$$V_n J_{N,r_n} = G_n^\top U_n (\bar{\Lambda}_n^\dagger)^{1/2} J_{p,r_n}, \tag{2.12}$$

where $V_n = [V_n J_{N,r_n} | V_n \bar{J}_{N,N-r_n}]$. Since $(\Lambda_n + cI_{N\times N})^{-1} - c^{-1} I_{N\times N}$ has only r_n nonzero diagonal entries, based on (2.10), we have

$$\begin{aligned} y_n^\top &= c^{-1}\mathbf{1}_N^\top + \mathbf{1}_N^\top V_n \big[(\Lambda_n + cI_{N\times N})^{-1} - c^{-1} I_{N\times N}\big] V_n^\top \\ &= c^{-1}\mathbf{1}_N^\top + \mathbf{1}_N^\top G_n^\top U_n (\bar{\Lambda}_n^\dagger)^{1/2} J_{p,r_n} J_{p,r_n}^\top \big[(\bar{\Lambda}_n + cI_{p\times p})^{-1} - c^{-1} I_{p\times p}\big] \\ &\quad \times J_{p,r_n} J_{p,r_n}^\top (\bar{\Lambda}_n^\dagger)^{1/2} U_n^\top G_n. \end{aligned}$$

Note that we have

$$\begin{aligned} &U_n (\bar{\Lambda}_n^\dagger)^{1/2} J_{p,r_n} J_{p,r_n}^\top \big[(\bar{\Lambda}_n + cI_{p\times p})^{-1} - c^{-1} I_{p\times p}\big] J_{p,r_n} J_{p,r_n}^\top (\bar{\Lambda}_n^\dagger)^{1/2} U_n^\top \\ &\qquad = -c^{-1} U_n J_{p,r_n} J_{p,r_n}^\top (\bar{\Lambda}_n + cI_{p\times p})^{-1} J_{p,r_n} J_{p,r_n}^\top U_n^\top, \end{aligned} \tag{2.13}$$

which could be understood as a "regularized pseudo-inverse." Specifically, when c is small, we have

$$U_n J_{p,r_n} J_{p,r_n}^\top (\bar{\Lambda}_n + cI_{p\times p})^{-1} J_{p,r_n} J_{p,r_n}^\top U_n^\top \approx (G_n G_n^\top)^\dagger. \tag{2.14}$$

We mention that $-c^{-1} U_n J_{p,r_n} J_{p,r_n}^\top (\bar{\Lambda}_n + cI_{p\times p})^{-1} J_{p,r_n} J_{p,r_n}^\top U_n^\top$ can be simplified to $-c^{-1} U_n (\bar{\Lambda}_n + cI_{p\times})^{-1} I_{p,r_n} U_n^\top$. Denote

$$\mathcal{I}_c(G_n G_n^\top) := U_n J_{p,r_n} J_{p,r_n}^\top (\bar{\Lambda}_n + cI_{p\times p})^{-1} J_{p,r_n} J_{p,r_n}^\top U_n^\top. \tag{2.15}$$

Hence, we can recast (2.10) and (2.11) into

$$y_n^\top = c^{-1}\mathbf{1}_N^\top - c^{-1}\mathbf{1}_N^\top G_n^\top \mathcal{I}_c(G_n G_n^\top) G_n \tag{2.16}$$

and

$$w_n^\top = \frac{\mathbf{1}_N^\top - \mathbf{1}_N^\top G_n^\top \mathcal{I}_c(G_n G_n^\top) G_n}{N - \mathbf{1}_N^\top G_n^\top \mathcal{I}_c(G_n G_n^\top) G_n \mathbf{1}_N} = \frac{\mathbf{1}_N^\top - \mathbf{T}_{n,z_k}^\top G_n}{N - \mathbf{T}_{n,z_k}^\top G_n \mathbf{1}_N}, \tag{2.17}$$

where

$$\mathbf{T}_{n,z_k} := \mathcal{I}_c(G_n G_n^\top) G_n \mathbf{1}_N \tag{2.18}$$

is chosen in order to have a better geometric insight into LLE. We now summarize the expansion of the barycentric coordinate.

PROPOSITION 2.1. *Take a data set $\mathcal{X} = \{z_i\}_{i=1}^n \subset \mathbb{R}^p$. Suppose there are N data points in the ε neighborhood of z_k, namely $\{z_{k,1}, \ldots, z_{k,N}\} \subset B_\varepsilon^{\mathbb{R}^p}(z_k) \cap (\mathcal{X} \setminus \{z_k\})$. Assume $p < N$. Let $G_n^\top G_n$ be the Gramian matrix associated with*

$\{z_{k,1} - z_k, \ldots, z_{k,N} - z_k\}$ and let $\{\lambda_{n,i}\}_{i=1}^r$ and $\{u_{n,i}\}_{i=1}^r$, where $r \le p$ is the rank of $G_n^\top G_n$, be the nonzero eigenvalues and the corresponding orthonormal eigenvectors of $G_n G_n^\top$ satisfying (2.12). With $\mathbf{T}_{n,z_k}$ defined in (2.18), the barycentric coordinates of z_k coming from the regularized equation (2.5) are

$$w_n^\top = \frac{\mathbf{1}_N^\top - \mathbf{T}_{n,z_k}^\top G_n}{N - \mathbf{T}_{n,z_k}^\top G_n \mathbf{1}_N}. \tag{2.19}$$

REMARK 2.1. The denominator $N - \mathbf{T}_{n,z_k}^\top G_n \mathbf{1}_N$ is the sum of all entries of the numerator $\mathbf{1}_N^\top - \mathbf{T}_{n,z_k}^\top G_n$. We could thus view the LLE matrix defined in (2.2) as a "normalized kernel" defined on the point cloud. However, while all entries of w_n are summed to 1, the vector $\mathbf{1}_N^\top - \mathbf{T}_{n,z_k}^\top G_n$ might have negative entries, depending on the vector $\mathbf{T}_{n,z_k}^\top$. Hence, in general, W is not a transition matrix.

How LLE achieves the nonlinear dimension reduction and captures the geometric structure of the point cloud could thus be understood by understanding $\mathbf{T}_{n,z_k}$. In the next section, we will show that under the manifold assumption, $\mathbf{T}_{n,z_k}$ is intimately related to the "normal bundle" associated with the manifold, and see how the selection of c influences the convergence behavior.

3. Asymptotic behavior of LLE. In this section, we focus on the asymptotic analysis of LLE assuming an underlying manifold setup. We start by introducing the manifold setup and assumptions for the analysis.

3.1. *Manifold setup.* Let X be a p-dimensional random vector. Assume that the range of X is supported on a d-dimensional compact, smooth Riemannian manifold (M, g) isometrically embedded in $\mathbb{R}^p$ via $\iota : M \hookrightarrow \mathbb{R}^p$, where we assume that M is boundary-free to simplify the discussion. Denote $d(\cdot, \cdot)$ to be the geodesic distance associated with g. For the tangent space $T_y M$ on $y \in M$, denote $\iota_* T_y M$ to be the embedded tangent space in $\mathbb{R}^p$. Denote $\exp_y : T_y M \to M$ to be the exponential map at y. Denote Ric to be the Ricci curvature, ∇ to be the covariant derivative and Δ to be the Laplace–Beltrami operator. Unless otherwise stated, in this paper we will carry out the calculations with the normal coordinate [10].

Let $z = \iota(y)$. Denote $\mathbb{I}_y$ to be the second fundamental form of ι at y. Denote the normal space at z as $(\iota_* T_y M)^\perp$, which could be viewed as $\mathbb{R}^{p-d}$. Recall that the second fundamental form at y is a symmetric bilinear map from $T_y M \times T_y M$ to $(\iota_* T_y M)^\perp$. If S^{d-1} is the $(d-1)$-dim unit sphere in $T_y M$ and $\theta = (\theta^1, \ldots, \theta^d) \in S^{d-1}$, then for a fixed $e_k \in (\iota_* T_y M)^\perp$, we can expand $e_k^\top \mathbb{I}_y(\theta, \theta)$ as $\sum_{i,j=1}^d p_{ij}^k \theta^i \theta^j$, where $p_{ij}^k \in \mathbb{R}$. The eigenvalues of the matrix $A^{(k)} \in \mathbb{R}^{d \times d}$, where $A_{ij}^{(k)} = p_{ij}^k$ for $i, j = 1, \ldots, d$, are the *principal curvatures* at z in the direction e_k.

We now quickly summarize how the probability density function (p.d.f.) associated with X is defined [7]. The random vector $X : \Omega \to \mathbb{R}^p$ is a measurable

function with respect to the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where $\mathcal{P}$ is the probability measure defined on the sigma algebra $\mathcal{F}$ in Ω. By assumption, the range of X is supported on $\iota(M)$. Let $\tilde{\mathcal{B}}$ be the Borel sigma algebra of $\iota(M)$, and denote by $\tilde{\mathcal{P}}_X$ the probability measure defined on $\tilde{\mathcal{B}}$ that is induced from $\mathcal{P}$. If $\tilde{\mathcal{P}}_X$ is absolutely continuous with respect to the volume density on $\iota(M)$ by the Radon–Nikodym theorem, $d\tilde{\mathcal{P}}_X(z) = P(z)\iota_* dV(z)$, where dV is the volume form associated with the metric g, $\iota_* dV(z)$ is the induced measure on $\iota(M)$ via ι and P is a nonnegative measurable function defined on $\iota(M)$. We call P *the p.d.f. of* X *on* M. When P is constant, we call X a *uniform* random sampling scheme; otherwise it is *nonuniform*.

To facilitate the discussion and the upcoming analysis, we make the following assumption about the random vector X and the regularity of the associated p.d.f.

ASSUMPTION 3.1. Assume $\tilde{\mathcal{P}}_X$ is absolutely continuous with respect to the volume density on $\iota(M)$ so that $d\tilde{\mathcal{P}}_X = P\iota_* dV$, where P is a measurable function. We further assume that $P \in C^5(\iota(M))$ and there exist $P_m > 0$ and $P_M \geq P_m$ so that $P_m \leq P(x) \leq P_M < \infty$ for all $x \in \iota(M)$.

Let $\mathcal{X} = \{\iota(x_i)\}_{i=1}^n \subset \iota(M) \subset \mathbb{R}^p$ denote a set of identical and independent (i.i.d.) random samples from X, where $x_i \in M$. We could then run LLE on $\mathcal{X}$. For $\iota(x_k) \in \mathcal{X}$ and $\varepsilon > 0$, we have $\mathcal{N}_{\iota(x_k)} := \{\iota(x_{k,1}), \ldots, \iota(x_{k,N})\} \subset B_\varepsilon^{\mathbb{R}^p}(\iota(x_k)) \cap (\mathcal{X} \setminus \{\iota(x_k)\})$. Take $G_n \in \mathbb{R}^{p\times N}$ to be the local data matrix associated with $\mathcal{N}_{\iota(x_k)}$ and evaluate the barycentric coordinate $w_n = [w_{n,1}, \ldots, w_{n,N}]^\top \in \mathbb{R}^N$. Again, although G_n and w_n depend on ε, n and x_k, to ease the notation, we only keep n to indicate that we have finite sampling points.

3.2. *Local covariance structure and local PCA*. We call

$$C_x := \mathbb{E}\big[(X - \iota(x))(X - \iota(x))^\top \chi_{B_\varepsilon^{\mathbb{R}^p}(\iota(x))}(X)\big] \in \mathbb{R}^{p\times p} \tag{3.1}$$

the *local covariance matrix* at $\iota(x) \in \iota(M)$, which is the covariance matrix associated with local PCA [7, 24]. In the proof of LLE under the manifold setup, the eigenstructure of C_x plays an essential role due to its relationship with the barycentric coordinate. Geometrically, for a d-dim manifold, the first d eigenvectors of C_x corresponding to the largest d eigenvalues provide an estimated basis for the embedded tangent space $\iota_* T_x M$, and the remaining eigenvectors form an estimated basis for the normal space at $\iota(x)$. To be more precise, a smooth manifold can be well approximated locally by an affine subspace. However, this approximation cannot be perfect, in case of nonvanishing curvature. It is well known that the contribution of curvature is of high order. For the purpose of fitting the manifold, we can ignore its contribution. For example, in [7, 24] local PCA is applied to estimate the tangent space. However, in LLE, the curvature plays an essential role and a careful analysis is needed to understand its role. In Lemma SI.5, we show a

generalization of the result shown in [7, 24] by expanding the C_x up to the third order for the sake of capturing the LLE behavior. The third-order term is needed for analyzing the regularization step shown in (2.5).

ASSUMPTION 3.2. Since the barycentric coordinate is rotational and translational invariant, without loss of generality, we assume that the manifold is translated and rotated properly, so that $\iota_* T_x M$ is spanned by $e_1, \ldots, e_d$.

PROPOSITION 3.1. *Fix $x \in M$ and suppose Assumption 3.2 holds. When ε is sufficiently small, we have*

$$C_x = \frac{|S^{d-1}|P(x)}{d(d+2)}\varepsilon^{d+2}\left(\begin{bmatrix} I_{d\times d} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} M_{11}^{(2)} & M_{12}^{(2)} \\ M_{21}^{(2)} & M_{22}^{(2)} \end{bmatrix}\varepsilon^2 + \begin{bmatrix} M_{11}^{(4)} & M_{12}^{(4)} \\ M_{21}^{(4)} & M_{22}^{(4)} \end{bmatrix}\varepsilon^4 + O(\varepsilon^6)\right),$$

where $M_{11}^{(2)}, M_{11}^{(4)} \in S(d)$, $M_{22}^{(2)}, M_{22}^{(4)} \in S(p-d)$, $M_{12}^{(2)}, M_{12}^{(4)} \in \mathbb{R}^{d\times(p-d)}$, $M_{12}^{(2)} = M_{21}^{(2)\top}$ and $M_{12}^{(4)} = M_{21}^{(4)\top}$. These matrices are defined in (SI.4), (SI.6), (SI.8) *and* (SI.9), *and $S(d)$ and $S(p-d)$ are defined in the end of Section 1. $M_{22}^{(2)}$ depends on $\mathbb{I}_x$ but does not depend on the p.d.f. P, and $M_{22}^{(4)}$ depends on the $\mathbb{I}_x$ and its derivatives, the Ricci curvature and P.*

The proof of Proposition 3.1 is postponed to Section SI.3. Since P is bounded by P_m from below, when ε is sufficiently small, the ε^{d+2} term is dominant and the largest d eigenvalues of C_x are of order ε^{d+2}. The other eigenvalues of C_x are of higher order and depend on the ε^{d+4} term or even the ε^{d+6} term. The behavior of eigenvectors is more complicated, due to the possible multiplicity of the corresponding eigenvalues.

To precisely calculate the eigenvalues and the corresponding eigenvectors of C_x, we apply the perturbation technique. We summarize the key steps here. Proposition 3.1 provides a Taylor expansion of C_x in terms of ε up to the third order, and we could view C_x as a function depending on ε around 0. Consider the eigendecomposition of C_x as

$$C_x U_x = U_x \Lambda_x, \tag{3.2}$$

where Λ_x is diagonal and $U_x \in O(p)$. Λ_x and U_x satisfy $\Lambda_x = \Lambda_x(0)\varepsilon^{d+2} + \Lambda_x'(0)\varepsilon^{d+4} + O(\varepsilon^{d+6})$ and $U_x = U_x(0)\varepsilon^{d+2} + U_x'(0)\varepsilon^{d+4} + O(\varepsilon^{d+6})$. Therefore, we obtain U_x and Λ_x if we find $\Lambda_x(0)$, $\Lambda_x'(0)$, $U_x(0)$ and $U_x'(0)$. To achieve this goal, we differentiate (3.2), and compare terms with the same order of ε. This technique fails to uniquely determine U_x when the eigenvalue repeats, and we

need higher order terms in C_x to determine the eigenvectors. The details can be found in Appendix SI.1.

To simplify the statement of the eigenstructure, following Assumption 3.2, we make one more assumption.

ASSUMPTION 3.3. Following Assumption 3.2, without loss of generality, we assume that the manifold is translated and rotated properly, so that $e_{d+1}, \dots, e_p$ "diagonalize" the second fundamental form; that is, $M_{22}^{(2)}$ in Proposition 3.1 is diagonalized to $\Lambda_2^{(2)} = \operatorname{diag}(\lambda_{d+1}^{(2)}, \dots, \lambda_p^{(2)})$.

The eigenstructure of the local covariance matrix is summarized in the following proposition. The detailed proof of the proposition is postponed to Section SI.3.

PROPOSITION 3.2. *Fix $x \in M$. Suppose ε is sufficiently small and Assumptions 3.2 and 3.3 hold. The eigendecomposition of $C_x = U_x \Lambda_x U_x^\top$, where $U_x \in O(p)$ and $\Lambda_x \in \mathbb{R}^{p\times p}$ is a diagonal matrix, is summarized below.*

Case 1: *When all diagonal entries of $\Lambda_2^{(2)}$ are nonzero, we have*

$$\Lambda_x = \frac{|S^{d-1}|P(x)\varepsilon^{d+2}}{d(d+2)} \begin{bmatrix} I_{d\times d} + \varepsilon^2 \Lambda_1^{(2)} + \varepsilon^4 \Lambda_1^{(4)} & 0 \\ 0 & \varepsilon^2 \Lambda_2^{(2)} + \varepsilon^4 \Lambda_2^{(4)} \end{bmatrix} + O(\varepsilon^6),$$

$$U_x = U_x(0)\big(I_{p\times p} + \varepsilon^2 \mathsf{S}\big) + O(\varepsilon^4) \in O(p),$$

where $\Lambda_1^{(2)}, \Lambda_1^{(4)} \in \mathbb{R}^{d\times d}$ and $\Lambda_2^{(4)} \in \mathbb{R}^{(p-d)\times(p-d)}$ are diagonal matrices with diagonal entries of order 1, $U_x(0) = \left[\begin{smallmatrix} X_1 & 0 \\ 0 & X_2 \end{smallmatrix}\right] \in O(p)$, $X_1 \in O(d)$, $X_2 \in O(p-d)$ and $\mathsf{S} \in \mathfrak{o}(p)$. The explicit expression of these matrices are listed in (SI.11)–(SI.18).

Case 2: *When l diagonal entries for $\Lambda_2^{(2)}$ are 0, where $1 \le l \le p-d$, we have the following eigendecomposition under some conditions. Divide C_x into blocks corresponding to the multiplicity l as*

$$C_x = \frac{|S^{d-1}|P(x)}{d(d+2)}\varepsilon^{d+2}\left(\begin{bmatrix} I_{d\times d} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} M_{11}^{(2)} & M_{12,1}^{(2)} & M_{12,2}^{(2)} \\ M_{21,1}^{(2)} & \Lambda_{2,1}^{(2)} & 0 \\ M_{21,2}^{(2)} & 0 & 0 \end{bmatrix}\varepsilon^2 + \begin{bmatrix} M_{11}^{(4)} & M_{12,1}^{(4)} & M_{12,2}^{(4)} \\ M_{21,1}^{(4)} & M_{22,11}^{(4)} & M_{22,12}^{(4)} \\ M_{21,2}^{(4)} & M_{22,21}^{(4)} & M_{22,22}^{(4)} \end{bmatrix}\varepsilon^4 + O(\varepsilon^6)\right), \tag{3.3}$$

where $M_{12,1}^{(2)}, M_{12,1}^{(4)} \in \mathbb{R}^{d\times(p-d-l)}$, $M_{12,2}^{(2)}, M_{12,2}^{(4)} \in \mathbb{R}^{d\times l}$, $M_{12,1}^{(2)} = M_{21,1}^{(2)\top}$, $M_{12,1}^{(4)} = M_{21,1}^{(4)\top}$, $M_{12,2}^{(2)} = M_{21,2}^{(4)\top}$, $M_{12,2}^{(2)} = M_{21,2}^{(4)\top}$, $M_{22,11}^{(4)} \in S(p-d-l)$, $M_{22,22}^{(4)} \in S(l)$, $M_{22,12}^{(4)} \in \mathbb{R}^{(p-d-l)\times l}$, and $M_{22,21}^{(4)} = M_{22,12}^{(4)\top}$.

Denote the eigendecomposition of the matrix $M_{22,22}^{(4)} - 2M_{21,2}^{(2)}M_{12,2}^{(2)}$ *as*

$$M_{22,22}^{(4)} - 2M_{21,2}^{(2)}M_{12,2}^{(2)} = U_{2,2}\Lambda_{2,2}^{(4)}U_{2,2}^{\top}, \tag{3.4}$$

where $U_{2,2} \in O(l)$ *and* $\Lambda_{2,2}^{(4)} = \mathtt{diag}[\lambda_{p-l+1}^{(4)}, \dots, \lambda_p^{(4)}]$ *is a diagonal matrix. If we further assume that all diagonal entries of* $\Lambda_{2,2}^{(4)}$ *are nonzero, we have*

$$\Lambda_x = \frac{|S^{d-1}|P(x)\varepsilon^{d+2}}{d(d+2)}\begin{bmatrix} I_{d\times d} + \varepsilon^2\Lambda_1^{(2)} + \varepsilon^4\Lambda_1^{(4)} & 0 & 0 \\ 0 & \varepsilon^2\Lambda_{2,1}^{(2)} + \varepsilon^4\Lambda_{2,1}^{(4)} & 0 \\ 0 & 0 & \varepsilon^4\Lambda_{2,2}^{(4)} \end{bmatrix} + O(\varepsilon^6),$$

$$U_x = U_x(0)\big(I_{p\times p} + \varepsilon^2\mathsf{S}\big) + O(\varepsilon^4) \in O(p),$$

where $\Lambda_1^{(4)}$ *and* $\Lambda_{2,1}^{(4)}$ *are diagonal matrices,*

$$U_x(0) = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & X_{2,1} & 0 \\ 0 & 0 & X_{2,2} \end{bmatrix} \in O(p),$$

$X_1 \in O(d)$, $X_{2,1} \in O(p-d-l)$, $X_{2,2} \in O(l)$, *and* $\mathsf{S} \in \mathfrak{o}(p)$. *The explicit formulae for these matrices are listed in* (SI.19)–(SI.21).

In general, the eigenstructure of C_x may be more complicated than the two cases considered in Proposition 3.2. In this general case, we could apply the same perturbation theory to evaluate the eigenvalues. Since the proof is similar but there is extensive notational loading, and it does not bring further insight to LLE, we skip details of these more general situations.

3.3. *Variance analysis of LLE.* We now study the asymptotic behavior of LLE. Under the manifold setup, from now on, we fix

$$c = n\varepsilon^{d+\rho}, \tag{3.5}$$

and we call ρ the *regularization order*. By (2.19), for $\boldsymbol{v} \in \mathbb{R}^N$, we have

$$\sum_{j=1}^{N} w_k(j)\boldsymbol{v}(j) = \frac{\mathbf{1}_N^{\top}\boldsymbol{v} - \mathbf{1}_N^{\top}G_n^{\top}\mathcal{I}_{n\varepsilon^{d+\rho}}(G_nG_n^{\top})G_n\boldsymbol{v}}{N - \mathbf{1}_N^{\top}G_n^{\top}\mathcal{I}_{n\varepsilon^{d+\rho}}(G_nG_n^{\top})G_n\mathbf{1}_N}. \tag{3.6}$$

Before proceeding, we provide *a* geometric interpretation of this formula. By the eigendecomposition $G_nG_n^{\top} = U_n\bar{\Lambda}_nU_n^{\top}$ and the fact that

$$\begin{aligned}\mathcal{I}_{n\varepsilon^{d+\rho}}(G_nG_n^{\top}) &= U_nJ_{p,r_n}J_{p,r_n}^{\top}\big(\bar{\Lambda}_n + n\varepsilon^{d+\rho}I_{p\times p}\big)^{-1}J_{p,r_n}J_{p,r_n}^{\top}U_n^{\top} \\ &= U_n\mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{\Lambda}_n)U_n^{\top}\end{aligned}$$

by the definition of $\mathcal{I}_\rho$ in (2.15), we have

$$\mathbf{1}_N^\top G_n^\top \mathcal{I}_{n\varepsilon^{d+\rho}}(G_n G_n^\top) G_n \boldsymbol{v} = \mathbf{1}_N^\top G_n^\top U_n \mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{\Lambda}_n) U_n^\top G_n \boldsymbol{v}$$

and

$$\mathbf{1}_N^\top G_n^\top \mathcal{I}_{n\varepsilon^{d+\rho}}(G_n G_n^\top) G_n \mathbf{1} = \mathbf{1}_N^\top G_n^\top U_n \mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{\Lambda}_n) U_n^\top G_n \mathbf{1}_N.$$

By the discussion of local PCA in Section 3.2, $U_n^\top G_n$ means evaluating the coordinates of all neighboring points of $\iota(x_k)$ with the basis composed of the column vectors of U_n, $U_n^\top G_n \mathbf{1}$ means the mean coordinate of all neighboring points, $\mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{\Lambda}_n)$ means a regularized weighting of the coordinates that helps to enhance the nonlinear geometry of the point cloud, and $G_n^\top U_n \mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{\Lambda}_n) U_n^\top G_n$ is a quadratic form of the averaged coordinates of all neighboring points. We could thus view the "kernel" part, $\mathbf{1}_N^\top G_n^\top U_n \mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{\Lambda}_n) U_n^\top G_n$, as preserving the geometry of the point cloud, by evaluating how strongly the weighted coordinates of neighboring points are related to the mean coordinate of all neighboring points by the inner product.

Asymptotically, by the law of large numbers, when conditional on $\iota(x_k)$,

$$\frac{1}{n} G_n \mathbf{1}_N = \frac{1}{n} \sum_{j=1}^{N} \big(\iota(x_{k,j}) - \iota(x_k)\big) \xrightarrow{n\to\infty} \mathbb{E}\big[\big(X - \iota(x_k)\big) \chi_{B_\varepsilon^{\mathbb{R}^p}(\iota(x_k))}(X)\big]$$

and we "expect" the following holds:

$$n\mathcal{I}_{n\varepsilon^{d+\rho}}(G_n G_n^\top) = \mathcal{I}_{\varepsilon^{d+\rho}}\left(\frac{1}{n} G_n G_n^\top\right) \xrightarrow{n\to\infty} \mathcal{I}_{\varepsilon^{d+\rho}}(C_{x_k}).$$

Also, we would "expect" to have

$$n\mathcal{I}_{n\varepsilon^{d+\rho}}(G_n G_n^\top) \frac{1}{n} G_n \mathbf{1}_N \xrightarrow{n\to\infty} \mathcal{I}_{\varepsilon^{d+\rho}}(C_{x_k}) \big[\mathbb{E}\big(X - \iota(x_k)\big) \chi_{B_\varepsilon^{\mathbb{R}^p}(x_k)}\big] =: \mathbf{T}_{\iota(x_k)}.$$

Hence, for $f \in C(\iota(M))$, for $\iota(x_k)$ and its corresponding $\mathcal{N}_{\iota(x_k)}$, we would "expect" to have

$$\begin{aligned} &\sum_{j=1}^{N} w_n(j) f(x_{k,j}) \\ &\xrightarrow{n\to\infty} \frac{\mathbb{E}[\chi_{B_\varepsilon^{\mathbb{R}^p}(x_k)}(X) f(X)] - \mathbf{T}_{\iota(x_k)}^\top \mathbb{E}[(X - \iota(x_k)) \chi_{B_\varepsilon^{\mathbb{R}^p}(x_k)}(X) f(X)]}{\mathbb{E}[\chi_{B_\varepsilon^{\mathbb{R}^p}(x_k)}(X)] - \mathbf{T}_{\iota(x_k)}^\top \mathbb{E}[(X - \iota(x_k)) \chi_{B_\varepsilon^{\mathbb{R}^p}(x_k)}(X)]} \\ &= \frac{\mathbb{E}[f(X)(1 - \mathbf{T}_{\iota(x)}^\top (X - \iota(x))) \chi_{B_\varepsilon^{\mathbb{R}^p}(x)}(X)]}{\mathbb{E}[(1 - \mathbf{T}_{\iota(x)}^\top (X - \iota(x))) \chi_{B_\varepsilon^{\mathbb{R}^p}(x)}(X)]}. \end{aligned} \tag{3.7}$$

However, it is not possible to directly see how the convergence happens, due to the dependence among different terms and how the regularized pseudo-inverse

converges. The dependence on the regularization order is also unclear. A careful theoretical analysis is needed.

To proceed with the proof, we need to discuss a critical observation. Note that the term C_x might be ill-conditioned for the pseudo-inverse procedure, and the regularized pseudo inverse depends on how the regularization penalty ρ is chosen. As we will see later, the choice of ρ is critical for the outcome. The ill condition depends on the manifold geometry, and can be complicated. In this paper, we focus on the following three cases.

CONDITION 3.1. Follow the notation used in Proposition 3.2. For the local covariance matrix C_x with the rank r, without loss of generality, we consider the following three cases:

- Case 0: $r = d$;
- Case 1: $r = p > d$, and $\lambda_{d+1}^{(2)}, \ldots, \lambda_p^{(2)}$ are nonzero;
- Case 2: $r = p > d$, $\lambda_{d+1}^{(2)}, \ldots, \lambda_{p-l}^{(2)}$, are nonzero, where $1 \le l \le p-d$, $\lambda_{p-l+1}^{(2)} = \cdots = \lambda_p^{(2)} = 0$, and $\lambda_{p-l+1}^{(4)}, \ldots, \lambda_p^{(4)}$ are nonzero.

At first glance, it is limited to assume that when $r > d$, we have $r = p$ in Cases 1 and 2. However, it is general enough in the following sense. In Cases 1 and 2, if C_x is degenerate, that is, $d < r < p$, it means that locally the manifold only occupies a lower dimensional affine subspace. Therefore, the sampled data are constrained to this affined subspace, and hence the rank of the local sample covariance matrix satisfies $r_n \le r$. As a result, the analysis can be carried out only on this affine subspace without changing the outcome. More general situations could be studied by the same analysis techniques shown below, but they will not provide more insights about our understanding of the algorithm and will introduce additional notational burdens. For $f \in C(\iota(M))$, define

$$Qf(x) := \frac{\mathbb{E}[f(X)(1 - \mathbf{T}_{\iota(x)}^\top (X - \iota(x)))\chi_{B_\varepsilon^{\mathbb{R}^p}(x)}(X)]}{\mathbb{E}[(1 - \mathbf{T}_{\iota(x)}^\top (X - \iota(x)))\chi_{B_\varepsilon^{\mathbb{R}^p}(x)}(X)]}, \tag{3.8}$$

The following theorem summarizes the relationship between LLE and Qf under these three cases.

THEOREM 3.1. *Fix $f \in C(\iota(M))$. Suppose the regularization order is $\rho \in \mathbb{R}$, $\varepsilon = \varepsilon(n)$ so that $\frac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2+1}} \to 0$ and $\varepsilon \to 0$ as $n \to \infty$. With probability greater than $1 - n^{-2}$, for all $x_k \in \mathcal{X}$, under different conditions listed in Condition 3.1, we*

have

$$\sum_{j=1}^{N} w_k(j) f(x_{k,j}) - f(x_k)$$

$$= \begin{cases} Qf(x_k) - f(x_k) + O\left(\dfrac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2-1}}\right) \\ \qquad \textit{in Case } 0, \\ Qf(x_k) - f(x_k) + O\left(\dfrac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2+[(-1)\vee(0\wedge(\rho-4))]}}\right) \\ \qquad \textit{in Cases } 1, 2. \end{cases} \tag{3.9}$$

Particularly, when $\rho \le 3$, *with probability greater than* $1 - n^{-2}$, *for all* $x_k \in \mathcal{X}$, *for all cases listed in Condition* 3.1, *we have*

$$\sum_{j=1}^{N} w_k(j) f(x_{k,j}) - f(x_k) = Qf(x_k) - f(x_k) + O\left(\frac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2-1}}\right). \tag{3.10}$$

The proof of Theorem 3.1 is postponed to Appendix SI.5. Note that the convergence rate of Case 0 is fast, no matter what regularization order ρ is chosen, while the convergence rate of Cases 1 and 2 depends on ρ. This theorem echoes several practical findings of LLE that the choice of regularization is critical in the performance, and it suggests that we should choose $\rho = 3$.

REMARK 3.1. We should compare the convergence rate of LLE with that of DM. The convergence rate of Case 0 is the same as that of eigenmap or DM without any normalization [26], while the convergence rate of Case 1 and Case 2 is the same as that of the α-normalized DM [8, 26] when $\rho \ge 4$. Note that the main convergence rate bottleneck for the α-normalized DM comes from the p.d.f. estimation, while the convergence bottleneck for LLE is the regularized pseudo-inverse.

3.4. *The kernel function corresponding to LLE.* Theorem 3.1 describes how LLE could be viewed as a "diffusion process" on the dataset. Note that

$$\begin{aligned} &\mathbb{E}\big[f(X)\big(1 - \mathbf{T}_{\iota(x)}^{\top}\big(X - \iota(x)\big)\big)\chi_{B_\varepsilon^{\mathbb{R}^p}(x)}(X)\big] \\ &\quad = \int_M \big(1 - \mathbf{T}_{\iota(x_k)}^{\top}\big(\iota(y) - \iota(x_k)\big)\big)\chi_{B_\varepsilon^{\mathbb{R}^p}(x_k)}\big(\iota(y)\big) f\big(\iota(y)\big) P(y)\, dV(y). \end{aligned} \tag{3.11}$$

Therefore, we can view w_n as a "zero-one" kernel supported on $B_\varepsilon^{\mathbb{R}^p}(x_k) \cap \iota(M)$ with the correction depending on $\mathbf{T}_{\iota(x_k)}$. Note that after the correction, the whole operator may no longer be a diffusion.

COROLLARY 3.1. *The integral kernel associated with LLE when the regularization order is $\rho \in \mathbb{R}$ is*

$$K_{\mathrm{LLE}}(x, y) = \big[1 - \mathbf{T}_{\iota(x)}^{\top}\big(\iota(y) - \iota(x)\big)\big]\chi_{B_{\varepsilon}^{\mathbb{R}^p}(\iota(x))\cap\iota(M)}\big(\iota(y)\big), \tag{3.12}$$

where $x, y \in M$ and

$$\mathbf{T}_{\iota(x)} := \mathcal{I}_{\varepsilon^{d+\rho}}(C_x)\big[\mathbb{E}\big(X - \iota(x)\big)\chi_{B_{\varepsilon}^{\mathbb{R}^p}(x)}\big] \in \mathbb{R}^p. \tag{3.13}$$

Note that K_{LLE} depends on ε, the geometry of the manifold near x, and ρ via $\mathbf{T}_{\iota(x)}$. We provide some properties of the kernel function K_{LLE}. By a direct expansion, we have $\mathbf{T}_{\iota(x)}^{\top} = \sum_{i=1}^{r} \frac{u_i^{\top}\mathbb{E}[(X-x_k)\chi_{B_{\varepsilon}^{\mathbb{R}^p}(x_k)}(X)]}{\lambda_i+\varepsilon^{d+\rho}} u_i^{\top}$, where u_i and λ_i are the ith eigenpair of C_x. Since $|\mathbb{E}(X - \iota(x_k))\chi_{B_{\varepsilon}^{\mathbb{R}^p}(x_k)}(X)|$ is bounded above by $\mathrm{vol}(M)\varepsilon$, $\lambda_i + \varepsilon^{d+\rho}$ is bounded below by $\varepsilon^{d+\rho}$ and each u_i is a unit vector, $|\mathbf{T}_{x_k}|$ is bounded above by $\sum_{i=1}^{r} \frac{\varepsilon\,\mathrm{vol}(M)}{\lambda_i+\varepsilon^{d+\rho}}$. Consequently, we have the following proposition.

PROPOSITION 3.3. *The kernel K_{LLE} is compactly supported and is in $L^2(M \times M)$. Thus, the linear operator $A : L^2(M, PdV) \to L^2(M, PdV)$ defined by*

$$Af(x) := \mathbb{E}\big[f(X)\big(1 - \mathbf{T}_{\iota(x)}^{\top}\big(X - \iota(x)\big)\big)\chi_{B_{\varepsilon}^{\mathbb{R}^p}(x)}(X)\big] \tag{3.14}$$

is Hilbert–Schmidt.

Note that the kernel function $K_{\mathrm{LLE}}(x, \cdot)$ depends on x, and hence the manifold, and the kernel are dominated by normal bundle information, due to the regularized pseudo-inverse procedure. For example, if M is an affine subspace of $\mathbb{R}^p$ and the data is uniformly sampled, then $\mathbb{E}[(X - x)\chi_{B_{\varepsilon}^{\mathbb{R}^p}(x)}(X)] = 0$, $\mathbf{T}_x = 0$ and $K(x, y) = 1$. If M is S^{p-1}, a unit sphere centered at origin embedded in $\mathbb{R}^p$ and the data is uniformly sampled, the first dominant $p - 1$ eigenvectors are perpendicular to x and the last eigenvector is parallel to x. By a direct calculation, $\mathbb{E}[(X - x)\chi_{B_{\varepsilon}^{\mathbb{R}^p}(x)}(X)]$ is parallel to x, and hence $K(x, y)$ behaves like a quadratic function $1 - cu_p^{\top}(y - x) = 1 - cx^{\top}(y - x)$, where c is the constant depending on the eigenvalues.

3.5. *Bias analysis.* For $f \in C(\iota(M))$, by the definition of A, we have

$$Qf(x) = \frac{(Af)(x)}{(A1)(x)}, \tag{3.15}$$

where 1 means the constant function. We now provide an *approximation of identity* expansion of the Q operator. By direct expansion, we have

$$Af(x) = \int_M K_{\mathrm{LLE}}(x, y) f\big(\iota(y)\big) P(y)\, dV(y). \tag{3.16}$$

While the formula of the Q operator looks like the diffusion process commonly encountered in the graph Laplacian based approach, like DM [8], the proof and the result are essentially different. To ease the notation, define

$$\begin{aligned}\mathfrak{N}_0(x) &:= \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \mathbb{I}_x(\theta,\theta)\,d\theta,\\ \mathfrak{M}_2(x) &:= \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \mathbb{I}_x(\theta,\theta)\theta\theta^\top\,d\theta,\\ \mathfrak{H}_f(x) &:= \mathtt{tr}\big(\mathfrak{M}_2(x)\nabla^2 f(x)\big),\end{aligned} \tag{3.17}$$

where $f \in C^3(\iota(M))$.

THEOREM 3.2. *Suppose $f \in C^3(\iota(M))$ and $P \in C^5(\iota(M))$ and fix $x \in M$. Assume that Assumptions 3.2 and 3.3 hold and the regularization order is $\rho \in \mathbb{R}$. Following the same notation used in Proposition 3.2, we have the following result*:

$$Qf(x) - f(x) = \big(\mathfrak{C}_1(x) + \mathfrak{C}_2(x)\big)\varepsilon^2 + O\big(\varepsilon^3\big), \tag{3.18}$$

where $\mathfrak{C}_1(x)$ and $\mathfrak{C}_2(x)$ depend on different cases stated in Condition 3.1.

• Case 0. *In this case*,

$$\mathfrak{C}_1(x) = \frac{1}{d+2}\left[\frac{1}{2}\Delta f(x) + \frac{\nabla f(x)\cdot\nabla P(x)}{P(x)} - \frac{\nabla f(x)\cdot\nabla P(x)}{P(x) + \frac{d(d+2)}{|S^{d-1}|}\varepsilon^{\rho-2}}\right], \tag{3.19}$$

$$\mathfrak{C}_2(x) = 0. \tag{3.20}$$

• Case 1. *In this case*,

$$\mathfrak{C}_1(x) = \frac{\frac{1}{d+2}[\frac{1}{2}\Delta f(x) + \frac{\nabla f(x)\cdot\nabla P(x)}{P(x)} - \frac{\nabla f(x)\cdot\nabla P(x)}{P(x)+\frac{d(d+2)}{|S^{d-1}|}\varepsilon^{\rho-2}}]}{1 - \frac{d}{2(d+2)}\sum_{i=d+1}^{p}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\frac{2}{d}\lambda_i^{(2)} + \frac{2(d+2)}{P(x)|S^{d-1}|}\varepsilon^{\rho-4}}}, \tag{3.21}$$

$$\mathfrak{C}_2(x) = -\frac{\frac{1}{4(d+4)}\sum_{i=d+1}^{p}\frac{(\mathfrak{N}_0^\top(x)e_i)(\mathfrak{H}_f^\top(x)e_i)}{\frac{2}{d}\lambda_i^{(2)} + \frac{2(d+2)}{P(x)|S^{d-1}|}\varepsilon^{\rho-4}}}{\frac{1}{d} - \frac{1}{2(d+2)}\sum_{i=d+1}^{p}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\frac{2}{d}\lambda_i^{(2)} + \frac{2(d+2)}{P(x)|S^{d-1}|}\varepsilon^{\rho-4}}}. \tag{3.22}$$

• Case 2. *In this case*,

$$\mathfrak{C}_1(x) = \frac{\frac{1}{d+2}[\frac{1}{2}\Delta f(x) + \frac{\nabla f(x)\cdot\nabla P(x)}{P(x)} - \frac{\nabla f(x)\cdot\nabla P(x)}{P(x)+\frac{d(d+2)}{|S^{d-1}|}\varepsilon^{\rho-2}}]}{1 - \frac{d}{2(d+2)}\sum_{i=d+1}^{p-l}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\frac{2}{d}\lambda_i^{(2)} + \frac{2(d+2)}{P(x)|S^{d-1}|}\varepsilon^{\rho-4}}}, \tag{3.23}$$

$$\mathfrak{C}_2(x) = -\frac{\frac{1}{4(d+4)}\sum_{i=d+1}^{p-l}\frac{(\mathfrak{N}_0^\top(x)e_i)(\mathfrak{H}_f^\top(x)e_i)}{\frac{2}{d}\lambda_i^{(2)}+\frac{2(d+2)}{P(x)|S^{d-1}|}\varepsilon^{\rho-4}}}{\frac{1}{d}-\frac{1}{2(d+2)}\sum_{i=d+1}^{p-l}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\frac{2}{d}\lambda_i^{(2)}+\frac{2(d+2)}{P(x)|S^{d-1}|}\varepsilon^{\rho-4}}}. \tag{3.24}$$

The proof of this long theorem is postponed to Appendix SI.4. Intuitively, based on the approximation of the identity, the kernel representation of the Q operator suggests that asymptotically we get the function value back, with the second-order derivative popping out in the second-order error term. In the GL setup, it has been well known that the second-order derivative term is the Laplace–Beltrami operator when the p.d.f. is constant [8]. However, due to the interaction between the geometric structure and the barycentric coordinate, LLE usually does not lead to the Laplace–Beltrami operator, unless under special situations. Note that while we could still see the Laplace–Beltrami operator in $\mathfrak{C}_1$, it is contaminated by other quantities, including $\mathfrak{N}_0(x)$, $\mathfrak{H}_f(x)$ and $\lambda_i^{(2)}$. These terms all depend on the second fundamental form. When $\rho > 4$, a curvature term appears in the ε^2 order term.

This theorem states that the asymptotic behavior of LLE is sensitive to the choice of ρ. We discuss each case based on different choices of ρ. If $\rho < 2$, for all cases,

$$\mathfrak{C}_1(x) = \frac{1}{(d+2)}\left[\frac{1}{2}\Delta f(x) + \frac{\nabla f(x)\cdot\nabla P(x)}{P(x)}\right] \quad \text{and} \quad \mathfrak{C}_2(x) = 0, \tag{3.25}$$

which comes from the fact that when ε^ρ is large, $\mathbf{T}_{\iota(x)}$ is small, and hence K_{LLE} is dominated by 1. Note that not only the Laplacian–Beltrami operator but also the p.d.f are involved, if the sampling is nonuniform. Therefore, when the choice of ρ is too small, the resulting asymptotic operator is the Laplace–Beltrami operator, only when the sampling is uniform. If $\rho = 3$, for all cases we have

$$\mathfrak{C}_1(x) = \frac{1}{2(d+2)}\Delta f(x) \quad \text{and} \quad \mathfrak{C}_2(x) = 0. \tag{3.26}$$

In this case, we recover the Laplacian–Beltrami operator, and the asymptotic result of LLE is independent of the nonuniform p.d.f.. This theoretical finding partially explains why such regularization could lead to a good result. If $\rho > 4$, since $\varepsilon^{d+\rho}$ is smaller than all eigenvalues of the local covariance matrix, asymptotically $\varepsilon^{d+\rho}$ is negligible and the result depends on different cases considered in Condition 3.1: for Case 0, we have

$$\mathfrak{C}_1(x) = \frac{1}{2(d+2)}\Delta f(x) \quad \text{and} \quad \mathfrak{C}_2(x) = 0,$$

for Case 1, we have

$$\mathfrak{C}_1(x) = \frac{\frac{1}{2(d+2)}\Delta f(x)}{1 - \frac{d^2}{4(d+2)}\sum_{i=d+1}^{p}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\lambda_i^{(2)}}},$$

$$\mathfrak{C}_2(x) = -\frac{\frac{d}{8(d+4)}\sum_{i=d+1}^{p}\frac{(\mathfrak{N}_0^\top(x)e_i)(\mathfrak{H}_f^\top(x)e_i)}{\lambda_i^{(2)}}}{\frac{1}{d} - \frac{d}{4(d+2)}\sum_{i=d+1}^{p}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\lambda_i^{(2)}}},$$

and for Case 2, we have

$$\mathfrak{C}_1(x) = \frac{\frac{1}{2(d+2)}\Delta f(x)}{1 - \frac{d^2}{4(d+2)}\sum_{i=d+1}^{p-l}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\lambda_i^{(2)}}},$$

$$\mathfrak{C}_2(x) = -\frac{\frac{d}{8(d+4)}\sum_{i=d+1}^{p-l}\frac{(\mathfrak{N}_0^\top(x)e_i)(\mathfrak{H}_f^\top(x)e_i)}{\lambda_i^{(2)}}}{\frac{1}{d} - \frac{d}{4(d+2)}\sum_{i=d+1}^{p-l}\frac{(\mathfrak{N}_0^\top(x)e_i)^2}{\lambda_i^{(2)}}}.$$

Note that when $\rho > 4$, we do not get the Laplace–Beltrami operator asymptotically in Cases 1 and 2. Furthermore, the behavior of LLE is dominated by the curvature and is independent of the p.d.f.

It is worth mentioning a specific situation when $\rho > 4$. Suppose the principal curvatures are equal to $\mathfrak{p} \in \mathbb{R}$ in the direction e_i, where $i = d+1, \ldots, p$, and vanish in the other directions. Then there is a choice of basis $e_1, \ldots, e_d$ so that $\mathbb{I}_x(\theta,\theta)\cdot e_i = \sum_{j=1}^{d}\mathfrak{p}\theta_j^2 = \mathfrak{p}$, where $\theta = (\theta_1, \ldots, \theta_d) \in S^{d-1}$. Under this specific situation, by a direct expansion, we have a simplification that

$$\frac{d}{8(d+4)}\big(\mathfrak{N}_0^\top(x)e_i\big)\big(\mathfrak{H}_f^\top(x)e_i\big) = \frac{1}{2(d+2)}\Delta f(x),$$

which leads to $\mathfrak{C}_1(x) + \mathfrak{C}_2(x) = 0$, and hence we obtain a fourth-order term.

The relationship between ε and the intrinsic geometry of the manifold requires further discussion, in order to better understand how the curvature plays a role in the whole analysis. We mention that the statement "suppose ε is sufficiently small" in Proposition 3.1, Proposition 3.2 and Theorem 3.2 is a technical condition needed in the proof of Lemma SI.3, which describes how well we could estimate the local geodesic distance by the ambient space metric. This technical condition depends on the fact that the exponential map is a diffeomorphism only if it is restricted to a subset of $\iota_* T_x M$ that is bounded by the injectivity radius of the manifold. That is, ε needs to be less than the injectivity radius. For any closed (compact without boundary) and smooth manifold, it is clear that different kinds of curvatures are bounded and the injectivity radius is strictly positive,

so there exists $\varepsilon_0 > 0$ less than the injectivity radius, so that for all $\varepsilon \le \varepsilon_0$, the statement "suppose ε is sufficiently small" is satisfied. The relationship between the curvature and ε_0 could be further elaborated by quoting the well-known result in [6]: for a closed Riemannian manifold of dimension d with the sectional curvature bounded by K, where $K \ge 0$, and with the volume lower bound v, where $v > 0$, the injectivity radius is bounded below by $i(d, K, v) > 0$, where $i(d, K, v)$ can be expressed explicitly in terms of d, K and v. Hence, ε_0 needs to satisfy $\varepsilon_0 < i(d, K, v)$.

3.6. *Convergence of LLE.* By combining the variation analysis and the bias analysis shown above, we conclude the following *pointwise* convergence theorem for LLE, when we have a proper choice of ρ.

THEOREM 3.3. *Take* $f \in C^3(\iota(M))$ *and* $P \in C^5(\iota(M))$, $\rho = 3$ *and* $\varepsilon = \varepsilon(n)$ *so that* $\frac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2+1}} \to 0$ *and* $\varepsilon \to 0$ *as* $n \to \infty$. *With probability greater than* $1 - n^{-2}$, *for all* $x_k \in \mathcal{X}$,

$$\frac{1}{\varepsilon^2}\left[\sum_{j=1}^{N} w_k(j) f(x_{k,j}) - f(x_k)\right] = \frac{1}{2(d+2)}\Delta f(x) + O(\varepsilon) + O\left(\frac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2+1}}\right).$$

Based on the Borel–Cantelli lemma, it is clear that asymptotically LLE converges almost surely. For practical purposes, we need to discuss the bandwidth choice when $\rho = 3$. Based on the assumption about the relationship between n and ε, we have $\frac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2+1}} \to 0$ as $n \to \infty$, but the convergence rate of $\frac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2+1}}$ might be slower than $\varepsilon \to 0$. Suppose we call a bandwidth "optimal," if it balances the standard deviation and the bias for all cases in Condition 3.1; that is, $\frac{\sqrt{\log(n)}}{n^{1/2}\varepsilon^{d/2+1}} \asymp \varepsilon$. We then have $\frac{n}{\log(n)} \asymp \frac{1}{\varepsilon^{d+4}}$, and we can estimate the optimal bandwidth from n.

4. Numerical examples. We adapt the LLE code provided in https://www.cs.nyu.edu/~roweis/lle/code.html to implement LLE with the ε-radius neighborhood. The Matlab code for the figures can be found in https://sites.google.com/site/hautiengwu/home/download.

4.1. *Sphere.* Suppose that $S^{p-1} \in \mathbb{R}^p$ is the unit sphere in $\mathbb{R}^p$. Denote H_k to be the space of homogeneous polynomials in $\mathbb{R}^p$ restricted on S^{p-1}. We have that the space H_k is the eigenspace of the Laplace–Beltrami operator on S^{p-1} corresponding to eigenvalue $-k(k+p-2)$, and the dimension of H_k is $\binom{p+k-1}{p-1} - \binom{p+k-3}{p-1}$ [28]. In this example, we show that if we choose a $\varepsilon^{d+\rho}$ that is too small, then we are not going to get the Laplace–Beltrami operator. When $\rho = 8$, which is

much greater than 3, by Theorem 3.2, we have

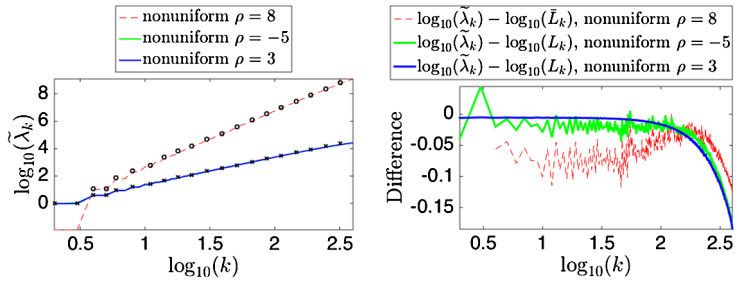
$$
\begin{aligned}
Qf(x_k) - f(x_k) = \Bigg(& \frac{-(p-1)}{8(p+3)(p+5)} \sum_{i=1}^{p-1} \partial_i^4 f(x_k) \\
& - \frac{(p-1)}{24(p+3)(p+5)} \sum_{i \neq j} \partial_i^2 \partial_j^2 f(x_k) \\
& - \frac{p+1}{24(p+3)(p+5)} \sum_{i=1}^{p-1} \partial_i^2 f(x_k) \Bigg) \varepsilon^4 + O\big(\varepsilon^6\big).
\end{aligned} \tag{4.1}
$$

A detailed calculation is shown in Section SI.7 (a calculation for the torus case is also provided). It is obvious that asymptotically, we get a fourth-order differential operator, instead of the Laplace–Beltrami operator. Specifically, when $p = 2$, or S^1,

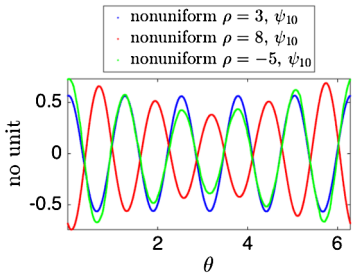
$$
Qf(x_k) - f(x_k) = -\frac{1}{280}\big(f''''(x_k) + f''(x_k)\big)\varepsilon^4 + O\big(\varepsilon^6\big). \tag{4.2}
$$

We mention that if the data set $\{x_i\}_{i=1}^n$ is nonuniformly sampled based on the p.d.f. P from S^1, then for any x_k we have $Qf(x_k) - f(x_k) = C\varepsilon^4 + O(\varepsilon^6)$, where C depends on the first four order differentiation of f at x_k and the first three order differentiations of P at x_k.

We now numerically show the relationship between the nonuniform sampling scheme and the regularization term. Fix $n = 30{,}000$. Take nonuniform sampling points $\theta_i := 2\pi U_i + 0.3\sin(2\pi i/n)$ on $(0, 2\pi]$, where $i = 1, \ldots, n$ and U_i is the uniform distribution on $[0, 1]$, and construct $\mathcal{X}_2 = \{(\cos(\theta_i), \sin(\theta_i))^\top\}_{i=1}^n \subset \mathbb{R}^2$. Run LLE with $\varepsilon = 0.0002$ and different ρ's, and evaluate the first 400 eigenvalues. Based on the theory, we know that when $\rho < 3$, the asymptotics depends on the nonuniform p.d.f.; when $\rho = 3$, we recover the Laplace–Beltrami operator in the ε^2 order; when $\rho > 3$, we get a fourth-order differential operator in the ε^4, which depends on the p.d.f. See Figure 1 for a comparison of the estimated eigenvalues and the predicted eigenvalues under different setups. We clearly see that the eigenvalues are well predicted under different ρ. When $\rho = 8$, we get a fourth-order term that depends on the nonuniform p.d.f.; when $\rho = 3$, LLE is independent of the nonuniform p.d.f. and we recover the spectrum of the Laplace–Beltrami operator in the second-order term, as is predicted by the developed theory; when $\rho = -5$, the nonuniform p.d.f. comes into play, and the eigenvalues are slightly shifted. To enhance the visualization, the difference between the estimated eigenvalues of S^1 and the theoretical values are shown on the middle subplot. The eigenfunctions provide more information. When $\rho = -5$ and $\rho = 8$, the dependence of the eigenfunctions on the p.d.f. could be clearly seen. For the nonuniform sampling scheme and $\rho = 8$, theoretically the first three eigenvalues come from the six-order term and depend on the p.d.f. Thus, numerically the first three eigenvalues are nonzero.



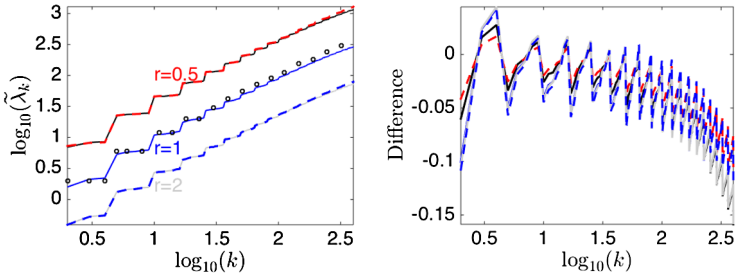
(a) S^1 Eigenvalues (b) S^1 Eigenvalues error



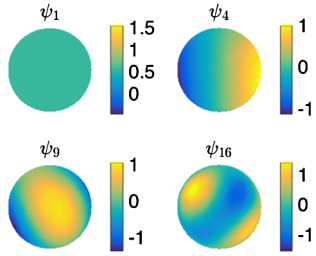
(c) S^1 Eigenfunctions

FIG. 1. *The first* 400 *eigenvalues of LLE on* 30,000 *points sampled from* S^1 *under a nonuniform sampling scheme with* $\rho = -5, 3, 8$. $\tilde{\lambda}_k$ *and* ψ_k *are the* k*th largest eigenvalue and the corresponding eigenfunction of LLE under different situations. The theoretical value,* $L_k := \lceil \frac{k-1}{2} \rceil^2$ *for the Laplace–Beltrami operator and* $\bar{L}_k := \lceil \frac{k-1}{2} \rceil^4 - \lceil \frac{k-1}{2} \rceil^2$ *for the fourth-order differential operator* $f'''' + f''$, *where* $\lceil x \rceil$ *means the smallest integer greater than or equal to* x, *are provided for a comparison. The eigenvalues and the theoretical values under different setups are shown in* (a), *with* L_k *shown as the black crosses and* $\bar{L}_k$ *as the black circles. The deviation of the evaluated eigenvalues from the theoretical values under different setups are shown in* (b). *The tenth eigenfunctions of LLE under different setups are shown in* (c).

Next, we show the results on S^2 with different radii under the nonuniform sampling scheme with $\rho = 3$ and different ε's. Fix $n = 30{,}000$. Take uniform sampling points $x_i = (x_{i1}, x_{i2}, x_{i3})^\top \in S^2 \subset \mathbb{R}^3$, where $i = 1, \ldots, n$, randomly choose $n/10$ points, randomly perturb those $n/10$ points by setting $\bar{x}_{i3} := x_{i3} + 1 - \cos(2\pi U_i)$, where U_i is the uniform distribution on $[0, 1]$ and $y_i := \frac{(x_{i1}, x_{i2}, \bar{x}_{i3})^\top}{\|(x_{i1}, x_{i2}, \bar{x}_{i3})^\top\|}$. As a result, $\mathcal{Y} := \{y_i\}_{i=1}^n \subset S^2$ is nonuniformly distributed on S^2. Denote $r\mathcal{Y}$ to be the scaled sampling points on the sphere with radius $r > 0$. Run LLE on $r\mathcal{Y}$ with different ε's, and evaluate the first 400 eigenvalues. We consider $r = 0.5, 1, 2$. For $r = 1$, consider $\varepsilon = 0.02$; for $r = 0.5$, consider $\varepsilon = 0.02/4$ and $0.02/6$; for $r = 2$, consider $\varepsilon = 0.02 \times 4$ and 0.02×3. Based on the theory, when $\rho = 3$, LLE is independent of the p.d.f. and we obtain the eigenvalues of the Laplace–Beltrami operator in all cases. See Figure 2 for the results under different setups. Theoretically, the eigenvalues of S^2 without counting multiplicities are $\nu_i = -i(i+1)$,



(a) S^2 Eigenvalues (b) S^2 Eigenvalues error



(c) S^2 Eigenfunctions

FIG. 2. (a) *The first* 400 *eigenvalues of LLE with* $\rho = 3$ *but different* ε, *over a* $n = 30{,}000$ *nonuniform sampling points on* S^2 *with different radii* $r > 0$. $\tilde{\lambda}_k$ *is the* k*th smallest eigenvalue of the Laplace–Beltrami operator estimated by LLE under different situations. When* $r = 0.5$ (*resp.*, $r = 1$ *and* $r = 2$), $\tilde{\lambda}_k$ *are shown in the black* (*resp.*, *blue and gray*) *curve. The results with different* ε *are shown as the red dash* (*resp.*, *blue dash*) *when* $r = 0.5$ (*resp.*, $r = 2$). *The theoretical eigenvalues for the canonical* S^2 (*with the radius* 1), *denoted as* L_k, $k = 1, \ldots,$ *are provided for a comparison* (*superimposed as black circles*). (b) *To enhance the visualization, the difference between the theoretical values and numerical values,* $\log_{10}(\tilde{\lambda}_k) - \log_{10}(L_k)$, *are shown with the same color and line properties as those shown on* (a). *Some eigenfunctions evaluated when* $r = 0.5$ *are shown on* (c).

where $i = 0, 1, \ldots$. The multiplicity of ν_i is $2i + 1$. When the radius is $r > 0$, the eigenvalues are scaled by r^{-2}. The eigenvalues, as is shown in Figure 2, can be well estimated by LLE, and the gap between the eigenvalues of spheres with different radii is predicted. The sawtooth behavior of the error comes from the spectral convergence behavior of eigenvalues with multiplicities. Note that there are 19 eigenvalues with multiplicity greater than 1 in the first 400 eigenvalues, which match the 19 oscillations found in Figure 2(b). The eigenfunctions are shown in Figure 2(c). As is predicted, the first eigenfunction is constant, as is shown in ψ_1. The eigenspace of ν_1 is spanned by three linear functions x, y and z, restricted on S^2. Therefore, ψ_4 is linear. The eigenspace of ν_ℓ is spanned by spherical harmonics of order ℓ, and its oscillation is illustrated in ψ_9 associated with ν_2 and ψ_{16} associated with ν_3.

4.2. *Examining the kernel.* We now show the numerical simulations of the corresponding kernel on the unit circle S^1 embedded in $\mathbb{R}^2$. We take a uni-

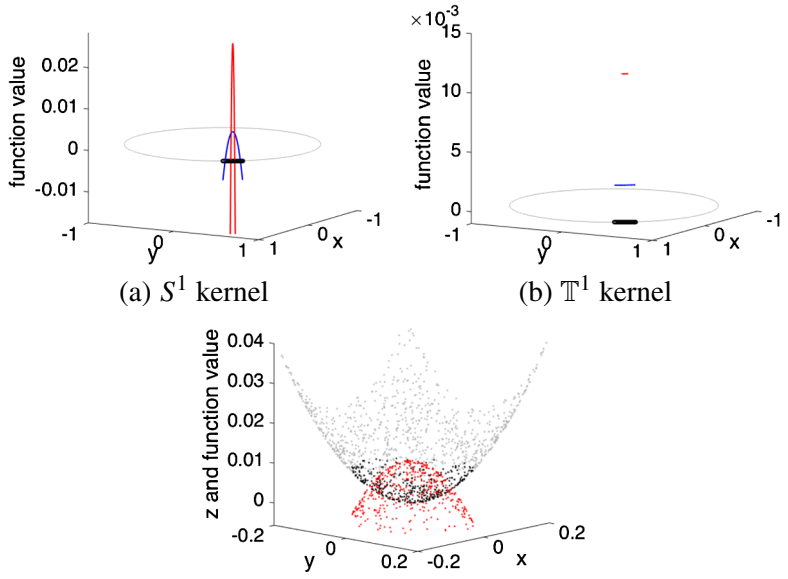


FIG. 3. (a) *The sampled* S^1 *is illustrated as the gray circle embedded in the* (x, y)*-plane. The black thick line indicates the first* 320 *neighbors of the central point* x_{1000}*. The red* (*resp.*, *blue*) *line is the corresponding normalized kernel when* $K = 80$ (*resp.*, $K = 320$). (b) *A surrogate of the sampled flat* 1*-dim torus* $\mathbb{T}^1$ *is illustrated as the gray circle embedded in the* (x, y)*-plane. The black thick line indicates the first* 320 *neighbors of the central point* x_{1000}*. The red* (*resp.*, *blue*) *line is the corresponding normalized kernel when* $K = 80$ (*resp.*, $K = 320$). (c) *A surrogate of the uniformly sampled* $\mathbb{S}^2$*. Only the first* 10,000 *nearest points of the chosen* $x = (0, 0, 0)$ *are plotted as the gray points. Note that the scale of the* x *and* y *axes and the* z *axis are different. The black points indicate the first* 400 *neighbors of* x*. The red points are the corresponding normalized kernel values when* $K = 400$.

form grid $\theta_i := 2\pi i/n$ on $(0, 2\pi]$, where $n \in \mathbb{N}$ and $i = 1, \ldots, n$, and construct $\mathcal{X} = \{x_i := (\cos(\theta_i), \sin(\theta_i))^\top\}_{i=1}^n \subset \mathbb{R}^2$, which could be viewed as a uniform sampled set from the unit circle. We fix $n = 10{,}000$. We then run LLE with $\varepsilon = [(\cos(\theta_{K/2}) - 1)^2 + \sin(\theta_{K/2})^2]^{1/2}$, where $K \in \mathbb{N}$. See Figure 3 for an example of the corresponding kernels when $K = 80$ and $K = 320$. Note that the constructed normalized kernel, $\frac{K_{\mathrm{LLE}}(x_{1000}, y)}{\int K_{\mathrm{LLE}}(x_{1000}, y)\, dV(y)}$, is nonpositive.

Next, we show the numerical simulations of the corresponding kernel on the 1-dim flat torus $\mathbb{T}^1 \sim \mathbb{R}/\mathbb{Z}$ with the induced metric from the canonical metric on $\mathbb{R}^1$. We take a uniform grid on $\mathbb{T}^1$ as $\{\theta_i = 2\pi i/n\}_{i=1}^n$, and take $\mathcal{X} = \{x_i := (\cos(\theta_i), \sin(\theta_i))^\top\}_{i=1}^n \subset \mathbb{R}^2$ to illustrate the flat torus. Fix $n = 10{,}000$ and run LLE with $\varepsilon = |\theta_{K/2}|$, where $K \in \mathbb{N}$. See Figure 3 for an example of the corresponding kernels when $K = 80$ and $K = 320$. The constructed normalized kernel, as the theory predicts, is constant. Note that in this case, we can view the flat 1-dim flat torus as the unit circle, when we have access to the geodesic distance.

Finally, we take a look at the unit sphere S^2 embedded in $\mathbb{R}^3$ with the center at $(0, 0, 1)$, and its corresponding kernel. We uniformly sample n points, $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^3$, from S^2. Fix $n = 10{,}000$ and run LLE with 400 nearest neighbors. See Figure 3 for the corresponding kernel. Note that the normalized kernel is not positive. These examples show that even for simple manifolds, the corresponding kernels might be complicated.

4.3. *Two-dimensional random tomography example.* To further examine the capability of LLE from the viewpoint of nonlinear dimension reduction, we consider the two-dimensional random tomography problem [25]. It is chosen due to its well-known and complicated geometrical structure.

We briefly describe the dataset and refer the interested reader to [25]. The classical two-dimensional transmission computerized tomography problem is to recover the function $f : \mathbb{R}^2 \to \mathbb{R}$ from its Radon transform. In the parallel beam model, the Radon transform of f is given by the line integral $R_\theta f(s) = \int_{x\cdot\theta=s} f(x)\,dx$, where $\theta \in S^1$ is perpendicular to the beaming direction $\theta^\perp \in S^1$, where S^1 is the unit circle, and $s \in \mathbb{R}$. We call θ the *projection direction* and $R_\theta f$ the *projected image*. There are cases, however, in which we only have the projected images and the projection directions are unknown. In such cases, the problem at hand is to estimate f from these projected images without knowing their corresponding projection directions. To better study this random projection problem, we need the following facts and assumptions. First, we know that for $f \in L^2(\mathbb{R}^2)$ with a compact support within $B_1(0)$, the map $R_\cdot f : \theta \in S^1 \mapsto L^2([-1, 1])$ is continuous [25]. To simplify the discussion, we assume that there is no symmetry in f; that is, $R_{\theta_1} f$ and $R_{\theta_2} f$ are different for all pairs of $\theta_1 \neq \theta_2$. Next, take $S := \{s_i\}_{i=1}^p$ to be the chosen set of sampling points on $[-1, 1]$, where $p \in \mathbb{N}$. In this example, we assume that S is a uniform grid on $[-1, 1]$; that is, $s_i = -1 + 2(i-1)/(p-1)$. For $\theta \in S^1$, denote the discretization of the projection image $R_\theta f$ as $D_S : L^2([-1, 1]) \to \mathbb{R}^p$, which is defined by $D_S : R_\theta f \mapsto (R_\theta f \star h_\varepsilon(s_1), R_\theta f \star h_\varepsilon(s_2), \ldots, R_\theta f \star h_\varepsilon(s_p))^\top \in \mathbb{R}^p$, where $h_\varepsilon(x) := \frac{1}{\varepsilon} h(\frac{x}{\varepsilon})$, h is a Schwartz function and h_ε converges weakly to the Dirac delta measure at 0 as $\varepsilon \to 0$. Note that, in general, $R_\theta f$ is a L^2 function when f is a L^2 function. Therefore, we need a convolution to model the sampling step. We assume that the discretization D_S is dense enough, so that $M^1 := \{D_p \circ R_\theta f\}_{\theta \in S^1}$ is also simple. In other words, we assume that p is large enough so that M^1 is a one-dimensional closed simple curve embedded in $\mathbb{R}^p$ and M^1 is diffeomorphic to S^1. Finally, we sample finite points from S^1 uniformly and obtain the simulation.

With the above facts and assumptions, we sample the Radon transform $\mathcal{X} := \{x_i := D_S \circ R_{\theta_i} f\}_{i=1}^n \subset \mathbb{R}^p$ with finite projection directions $\{\theta_i\}_{i=1}^n$, where $\{\theta_i\}_{i=1}^n$ is a finite uniform grid on S^1; that is, $\mathcal{X}$ is sampled from the one-dimensional manifold M^1. For the simulations with the Shepp–Logan phantom, we take $n = 4096$, and the number of discretization points was $p = 128$. It has been shown

in [25] that DM could recover the M^1 up to diffeomorphism, that is, we could achieve the nonlinear dimensional reduction. In order to avoid distractions, we do not consider any noise as is considered in [25], and focus our analysis on the clean dataset. The Shepp–Logan image, some examples of the projections and the results of PCA, DM and LLE, are shown in Figure 4. As is shown in [25], PCA fails to

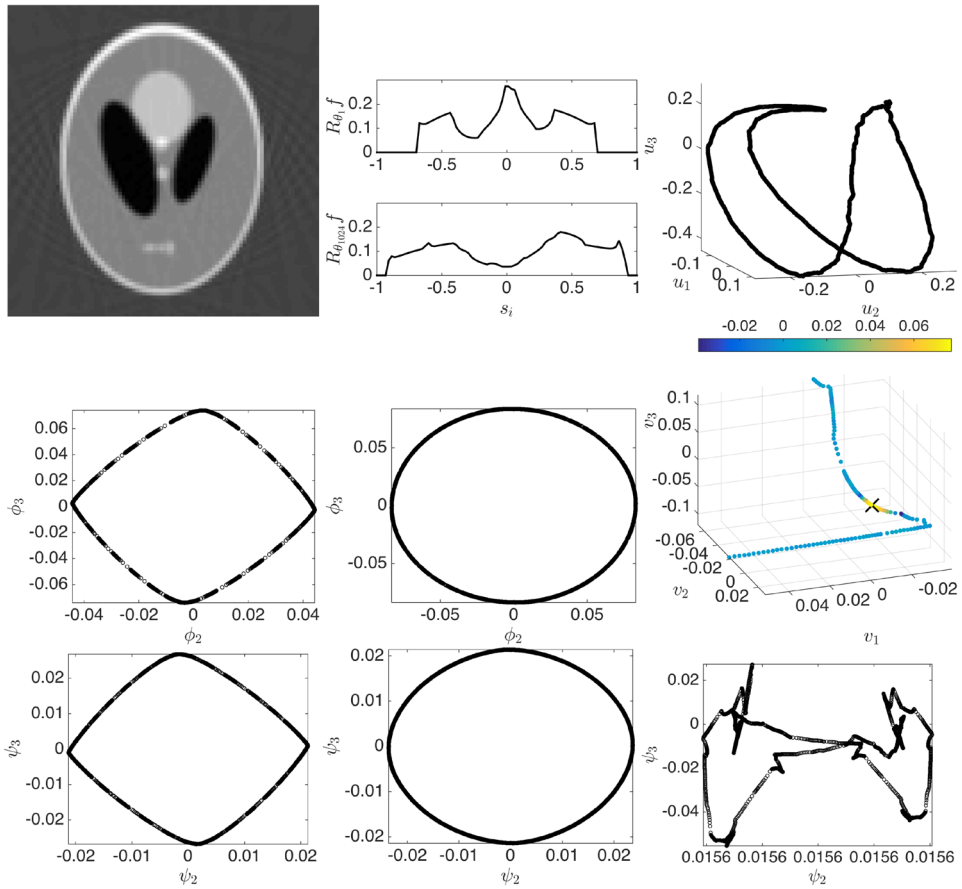


FIG. 4. *Top row*: *the left panel is the Shepp–Logan phantom, the middle panel shows two projection images from two different projection directions and the right panel shows the linear dimension reduction of the dataset by the first three principal components,* u_1, u_2 *and* u_3. *Middle row*: *the left panel shows DM without the* α*-normalization technique* [8], *where the embedding is done by choosing the first two nontrivial eigenvectors of the graph Laplacian,* ϕ_2 *and* ϕ_3, *the middle panel shows DM with the* α*-normalization technique when* $\alpha = 1$, *and the right panel shows that the sign of the kernel corresponding to LLE is indeterminate, where the black cross indicates* x_{3555}, *and the kernel values on its neighbors are encoded by color* (*the neighbors are visualized by the top three principal components,* v_1, v_2 *and* v_3). *Bottom row*: *the embedding using the second and third eigenvectors of LLE,* ψ_2 *and* ψ_3, *under different setups is shown. The left, middle and right panels show the result with* $\rho = -5$, $\rho = 3$ *and* $\rho = 8$, *respectively.*

embed $\mathcal{X}$ with only the first three principal components, while DM succeeded. There is more discussion for DM, particularly its robustness to the noise and metric design [25], so they are not discussed here. For LLE, we take $\varepsilon = 0.004$. The embedding results of LLE with different regularization orders, $\rho = 8, 3, -5$, are shown. Due to the complicated geometrical structure, we encounter difficulty even to recover the topology of M^1 by LLE, if the regularization is not chosen properly.

To examine whether the sign of the kernel corresponding to LLE is indeterminate in this database, we fix $x_{3555} \in \mathcal{X}$, and apply PCA to visualize its $K = 150$ neighbors. The kernel function is shown in Figure 4 as the color encoded on the embedded points. The sign of the kernel is indeterminate, as is predicted by the above theory due to the existence of curvature. In summary, we should be careful when we apply LLE to a complicated real database.

5. ε-radius neighborhood v.s. K nearest neighborhood. In the original article [22], the KNN scheme was proposed for LLE. However, the analysis in this paper has been based on the ε-radius neighborhood scheme. These two schemes are closely related asymptotically from the viewpoint of p.d.f. estimation [21]. The following argument shows that the developed theorems are actually transferrable to the KNN scheme under the manifold setup.

We follow the notation in Section 3.1. For $\iota(x_k) \in \mathcal{X}$, take K nearest neighbors of $\iota(x_k)$, namely $\iota(x_{k,1}), \ldots, \iota(x_{k,K})$, with respect to the Euclidean distance. Intuitively, K is closely related to the volume of the minimal ball centered at x_k with the radius $\varepsilon(x_k)$ containing the K nearest neighbors of x_k, where $\varepsilon(x_k)$ depends on K and the p.d.f.; that is, we expect to have

$$n P(x_k) \operatorname{vol}(D_{x_k}) \approx K, \tag{5.1}$$

where $D_x := B^{\mathbb{R}^p}_{\varepsilon(x)}(\iota(x)) \cap \iota(M)$ is the minimal ball centered at $x \in M$ with the radius $\varepsilon(x) > 0$ so that D_x contains the K nearest neighbors of x. Under the smoothness assumption of the p.d.f. and the manifold setup, we claim that asymptotically when $n \to \infty$, this relationship holds uniformly over the manifold a.s., if $K = K(n)$, $K/\log(n) \to \infty$ and $K/n \to 0$ as $n \to \infty$. This claim could be achieved by slightly modifying the argument for the Theorem in [9] to obtain the large deviation bound for (5.1) when n is finite. To bound $\Pr\{\sup_{x \in M} |\frac{K}{n \operatorname{vol}(D_x)} - P(x)| > \alpha\}$, where $\alpha > 0$, it is sufficient to bound the two terms on the right-hand side of [9], equation (10). By a straightforward calculation of the equations on page 539 in [9], we achieve the bound $\Pr\{\sup_{x \in M} |\frac{K}{n \operatorname{vol}(D_x)} - P(x)| > \alpha\} \le \operatorname{poly}(n) e^{-cK\alpha^3}$, where c is a constant depending on d and the upper bounds of $P(x)$ on M, and $\operatorname{poly}(n) = 3(1 + 2^{p+3} n^{p+3})$. This can be observed by combining [9], equations (6), (7), (9) and (10)—the second term on the right-hand side of [9], equation (10), is dominated by the first term. To bound the first term, we can substitute $\delta = \frac{K\beta}{4n(P_M + \beta)}$ and $M = \frac{4kP_M}{n\beta}$ into the fourth unlabeled equation on page 539 in [9], where P_M is the upper bound of p.d.f. In the fourth unlabeled equation, α is

the upper bound of the volume ratio of $B^{\mathbb{R}^p}_{2\varepsilon(x)}(\iota(x)) \cap \iota(M)$ and $B^{\mathbb{R}^p}_{\varepsilon(x)}(\iota(x)) \cap \iota(M)$, which can be chosen as 3^d when $\varepsilon(x)$ is sufficiently small. To conclude the bound for $\Pr\{\sup_{x\in M} |\frac{K}{n\,\mathrm{vol}(D_x)} - P(x)| > \alpha\}$, we use the fact that when β is small, the equation follows. Therefore, if we choose $\alpha = (\frac{2p+10}{c})^{1/3}(\frac{\log n}{K})^{1/3}$, with probability greater than $1 - n^{-2}$, we have uniformly $\frac{K}{n\,\mathrm{vol}(D_x)} = P(x) + O(\alpha)$. Note that by assumption, $\alpha \to 0$ as $n \to \infty$. We conclude that with probability greater than $1 - n^{-2}$,

$$\varepsilon(x) = \left(\frac{d}{|S^{d-1}|}\right)^{1/d} \left(\frac{K}{nP(x)}\right)^{1/d} \left(1 + O\left(\left(\frac{\log n}{K}\right)^{1/3}\right)\right), \tag{5.2}$$

where we use the fact that $\mathrm{vol}(D_x) = \frac{|S^{d-1}|}{d}\varepsilon(x)^d + O(\varepsilon(x)^{d+1})$ when $\varepsilon(x)$ is sufficiently small. It is transparent that $\varepsilon(x)$ depends on n and $\varepsilon(x) \to 0$ a.s. as $n \to \infty$ since $K(n)/n \to 0$ by assumption. In other words, ε is not a constant value. It is a function depending on the p.d.f.. If we require $K = K(n)$ to additionally satisfy $\frac{K(n)}{n}\frac{K(n)^{d/2}}{\log(n)^{d/2}} \to \infty$, then $\varepsilon(x_k)$ satisfies $\frac{\sqrt{n}}{n^{1/2}\varepsilon(x)^{d/2+1}} \to 0$ a.s. On the other hand, notice that the statement of Theorem 3.3 is pointwise. Therefore, its proof could be directly employed to the case when ε is chosen pointwisely, and hence the KNN scheme. As a result, if we take $\rho = 3$ and is $K/n \to 0$, $K/\log(n) \to \infty$, and $(K/n)(K/\log(n))^{d/2} \to \infty$ when $n \to \infty$, by plugging (5.2) into Theorem 3.3, when n is sufficiently large, the following convergence holds for all x_k with probability greater than $1 - 2n^{-2}$:

$$\begin{aligned}
&\sum_{j=1}^{K} w_k(j) f(x_{k,j}) - f(x_k) \\
&\quad = \frac{(\frac{d}{|S^{d-1}|})^{1/d}}{2(d+2)} \frac{\Delta f(x_k)}{P(x_k)^{2/d}} \left(\frac{K}{n}\right)^{2/d} \\
&\qquad + O\left(\left(\frac{\log(n)}{K}\right)^{1/3} \left(\frac{K}{n}\right)^{2/d}\right) + O\left(\left(\frac{\log(n)}{K}\right)^{1/2} \left(\frac{K}{n}\right)^{1/d}\right).
\end{aligned} \tag{5.3}$$

In summary, unless the sampling is uniform, we do not obtain the Laplace–Beltrami operator with the KNN scheme. Based on the expansion (5.3), to obtain the Laplace–Beltrami operator with the KNN scheme, we could numerically consider a "normalized LLE matrix"; that is, find the eigenstructure of $\tilde{L} := \mathcal{E}^{-1}(W - I)$, where W is the ordinary LLE matrix, and $\mathcal{E} \in \mathbb{R}^{n\times n}$ is a diagonal matrix so that $\mathcal{E}_{ii} = \varepsilon(x_i)^2$. Since the analysis of the pointwise convergence of $\tilde{L}$ is similar to that of Theorem 3.3, we skip the details here.

6. Relationship with two statistical topics.

6.1. *Locally linear regression.* Based on the above theoretical study under the manifold setup, we could link LLE to LLR [7, 15]. Recall that in LLR, we locally fit a linear function to the response, and the associated kernel depends on the inverse of a variation of the covariance matrix. We summarize how LLR is operated. Consider the following regression model:

$$Y = m(X) + \sigma(X)\xi, \tag{6.1}$$

where ξ is a random error independent of X with $\mathbb{E}(\xi) = 0$ and $\operatorname{Var}(\xi) = 1$, and both the regression function m and the conditional variance function σ^2 are defined on $\mathbb{R}^d$. Let $\{(X_l, Y_l)\}_{l=1}^n$ denote a random sample observed from model (6.1) with $\mathcal{X} := \{X_l\}_{l=1}^n$ being sampled from X. Given $\{(X_l, Y_l)\}_{l=1}^n$ and $x \in \mathbb{R}^d$, the problem is then to estimate $m(x)$ assuming enough smoothness of m. Choose a smooth kernel function with fast decay $K : [0, \infty] \to \mathbb{R}$ and a bandwidth $\varepsilon > 0$. The LLR estimator for $m(x)$ is defined as $e_1^\top \hat{\boldsymbol{\beta}}_x$, where

$$\begin{aligned} &\hat{\boldsymbol{\beta}}_x = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} (\mathbf{Y} - \mathbf{X}_x\boldsymbol{\beta})^\top \mathbf{W}_x (\mathbf{Y} - \mathbf{X}_x\boldsymbol{\beta}), \\ &\mathbf{Y} = (Y_1, \ldots, Y_n)^\top, \qquad \mathbf{X}_x = \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix}^\top \in \mathbb{R}^{n \times (d+1)}, \\ &\mathbf{W}_x = \mathtt{diag}\big(K_\varepsilon(X_1, x), \ldots, K_\varepsilon(X_n, x)\big) \in \mathbb{R}^{n \times n}, \end{aligned} \tag{6.2}$$

and $K_\varepsilon(X_l, x) := \varepsilon^{-d} K(\|X_l - x\|_{\mathbb{R}^d}/\varepsilon)$. By a direct expansion, (6.2) becomes

$$\hat{\boldsymbol{\beta}}_x = \big(\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\big)^{-1} \mathbf{X}_x^\top \mathbf{W}_x \mathbf{Y} \tag{6.3}$$

if $(\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x)^{-1}$ exists. We have $\mathbf{X}_x = \left[\begin{smallmatrix} \mathbf{1}_n^\top \\ \mathbf{G}_x \end{smallmatrix}\right]$, where $\mathbf{G}_x$ is the data matrix associated with $\{X_i\}_{i=1}^n$ centered at x. By yet another direct expansion by the block inversion,

$$e_1^\top \hat{\boldsymbol{\beta}}_x = {w_x^{(\mathrm{LLR})}}^\top \mathbf{Y}, \tag{6.4}$$

where $w_x^{(\mathrm{LLR})}$ is called the "smoothing kernel" and satisfies

$$w_x^{(\mathrm{LLR})} := \frac{\mathbf{1}_n^\top \mathbf{W}_x - \mathbf{1}_n^\top \mathbf{W}_x \mathbf{G}_x^\top (\mathbf{G}_x \mathbf{W}_x \mathbf{G}_x^\top)^{-1} \mathbf{G}_x \mathbf{W}_x}{\mathbf{1}_n^\top \mathbf{W}_x \mathbf{1}_n - \mathbf{1}_n^\top \mathbf{W}_x \mathbf{G}_x^\top (\mathbf{G}_x \mathbf{W}_x \mathbf{G}_x^\top)^{-1} \mathbf{G}_x \mathbf{W}_x \mathbf{1}_n}. \tag{6.5}$$

Through a direct comparison, we see that the vector $w_x^{(\mathrm{LLR})}$ is almost the same as the weight matrix in LLE shown in (2.17), except the weighting by the chosen kernel—in LLE, the kernel function and its support are both determined by the data, while in LLR the kernel is selected in the beginning and the data points are

weighted by the chosen kernel like $\mathbf{G}_x\mathbf{W}_x$. If we choose the kernel to be a zero-one kernel with the support on the ball centered at x with the radius ε, then we "recover" (2.17).

Under the low dimensional manifold setup, $\mathbf{G}_x\mathbf{W}_x\mathbf{G}_x^\top$ might not be of full rank. Note that the term $\mathbf{G}_x\mathbf{W}_x\mathbf{G}_x^\top$ is the weighted local covariance matrix, which is considered in [24] to estimate the tangent space. Unlike the regularized pseudo-inverse (2.15) in LLE, to handle this degeneracy issue, in LLR the data matrix $\mathbf{G}_x$ is constructed by projecting the point cloud to the estimated tangent plane. This projection step could be understood as taking the Moore–Penrose pseudo-inverse approach to handle the degeneracy. We mention that in [7], Section 6, the relationship between LLR and manifold learning under the manifold setup is established. It is shown that asymptotically, the smooth matrix from the kernel $w_x^{(\mathrm{LLR})}$ leads to the Laplace–Beltrami operator, which is parallel to the reported result in this paper.

These relationships between LLE and LLR suggest the possibility of fitting the data locally by taking the locally polynomial regression into account, and generalizing the barycentric coordinate by fitting a polynomial function locally. By this generalization, it is possible to catch more delicate structure of the manifold in a different adaptive way. Since this direction is outside the scope of this paper, the study of this possibility is left to future studies.

6.2. *Measurement errors*. In this work, we analyze LLE under the assumption that the dataset is randomly sampled directly from a manifold, without any influence of noise. However, noise is inevitable and a further study is needed. By the analysis, we observe that LLE takes care of the measurement error (or "error in variable") challenge "in some sense."

Suppose the dataset is $\{y_i\}_{i=1}^n \subset \mathbb{R}^p$, where $y_i = z_i + \xi_i$, z_i is supported on a manifold and ξ_i is an i.i.d. noise with good properties. The question of interest is: how much information LLE could recover from $\{z_i\}_{i=1}^n$. A parallel problem for the GL, or the more general graph connection Laplacian (GCL), has been studied in [13, 14]. It shows that the spectral properties of the GL and GCL are robust to noise. For LLE, while a similar analysis could be applied, if we view LLE as a kernel method and show a similar result, we mention that we might benefit by taking the special algorithmic structure of LLE into account.

When the dimension of the dataset is high, the noise might have a nontrivial behavior. For example, when the dimension of the database $p = p(n)$ satisfies $p(n)/n \to \gamma > 0$ when $n \to \infty$ (known as the *large p and large n setup*), it is problematic to even estimate the covariance matrix. Note that the covariance matrix is directly related to LLE since the covariance matrix appears in the regularized pseudo inverse, $\mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{G}_n\bar{G}_n^\top)$, where $\bar{G}_n$ is the local data matrix associated with y_k determined from the noisy database $\{y_i\}_{i=1}^n$, and $\bar{G}_n\bar{G}_n^\top$ is the covariance matrix. Under the large p and large n setup, the eigenvalues and eigenvectors of the covariance matrix will both be biased, depending on the "signal-to-noise ratio"

and γ [20]. A careful manipulation of the noise, or a modification of the covariance matrix estimator, is needed in order to address these introduced biases. For example, the "shrinkage technique" was introduced to correct the eigenvalue bias with a theoretical guarantee [11, 25]. The covariance matrix estimator based on the shrinkage technique is $\tilde{C}_n := \sum_{l=1}^{p} f(\lambda_l) u_l u_l^\top$, where u_l and λ_l form the lth eigenpair of $\bar{G}_n \bar{G}_n^\top$ and f is the designed shrinkage function.

A direct comparison shows that the regularized pseudo inverse in LLE behaves like a shrinkage technique. Recall that $\mathcal{I}_{n\varepsilon^{d+\rho}}(\bar{G}_n \bar{G}_n^\top) = \sum_{l=1}^{r_n} \frac{1}{\lambda_l + n\varepsilon^{d+\rho}} u_l u_l^\top$ (2.15), where r_n is the rank of $\bar{G}_n \bar{G}_n^\top$, the shrinkage function is $f(x) = \frac{1}{x+n\varepsilon^{d+\rho}} \chi_{(0,\infty)}(x)$ and χ is the indicator function. Although how f corrects the noise impact is outside the scope of this paper, it would be possible to carefully improve the regularized pseudo inverse by taking the shrinkage technique into account. In other words, by modifying the barycentric coordinate evaluation and applying the technique discussed in [13, 14], it is possible to improve LLE. An extensive study of the topic will be reported in upcoming research.

7. Conclusions and discussion. We provide an asymptotical analysis of LLE under the manifold setup. The theoretical results indicate that asymptotically, LLE generally may not give the expected Laplace–Beltrami operator, unless the regularization is chosen properly. From the integral operator viewpoint, the corresponding kernel of LLE in general is not positive. Therefore, LLE in general is not a diffusion operator. Some direct calculations of the operator associated with LLE over simple manifolds, like the sphere, indicate that asymptotically a fourth-order differential operator might pop out as the dominant term, if the regularization chosen is too small. The numerical results support the theoretical findings. In addition, we also discuss the relationship between LLE and two statistical topics—LLR and the measurement error problem, and point out potential future work.

There are more important topics we do not explore in this paper. First, note that the pointwise convergence result established in this paper comes from a careful analysis of the "fit locally" part of LLE. However, it is not sufficient to fully understand the "think globally" part of LLE. Recall that we evaluate the eigendecomposition of the LLE matrix for the embedding in the last step of LLE. The theoretical and numerical results suggest that the eigenstructure of the LLE matrix provides an approximation of the eigenstructure of the Laplace–Beltrami operator. The embedding in the last step could therefore be understood from the point of view of the spectral embedding theory [4, 5]. The eigenstructure of the LLE matrix integrates the local information. As a result, we catch the "think globally" part. However, pointwise convergence is not strong enough to guarantee spectral convergence. In other words, we need to show that asymptotically, the eigen-decomposition provides a proper approximation of the eigenstructure of the Laplace–Beltrami operator. While a proof similar to that in [26] could be slightly modified to achieve the spectral convergence of LLE, however, more may

be needed, such as the spectral convergence rate, from the statistical viewpoint. Recently, there have been some relevant works for the GL under the manifold model in this direction [17, 32]. Based on the special structure of LLE, like the regularization, the optimal convergence rate of LLE could be different and additional exploration is needed. The result will be reported in our future work.

Another important topic is the appearance of a fourth-order differential operator in LLE, when the manifold has a special structure and the regularization is improperly chosen. Although it would be a by-product, it would be interesting to ask if it is possible to take this fourth-order differential operator into account in the data analysis and which kind of information could be extracted from the dataset. It would also be interesting to ask if it is possible to directly obtain a fourth-order differential operator for more general manifolds with a slight modification of LLE. A direct benefit of this possibility is linked back to the regression problem, such as LLR. If we could directly eliminate the second-order term, the regression result could be more accurate. We leave this study direction to our future work.

Acknowledgment. The authors acknowledge anonymous reviewers' valuable comments to improve the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Think globally, fit locally under the manifold setup: Asymptotic analysis of locally linear embedding" (DOI: 10.1214/17-AOS1676SUPP; .pdf). Proof of main theorems and technical details.

REFERENCES

[1] BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput*. **15** 1373–1396.

[2] BELKIN, M. and NIYOGI, P. (2005). Towards a theoretical foundation for Laplacian-based manifold methods. In *Learning Theory*. *Lecture Notes in Computer Science* **3559** 486–500. Springer, Berlin. MR2203282

[3] BELKIN, M. and NIYOGI, P. (2007). Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems* 19 (*NIPS* 2006) 129–136. MIT Press, Cambridge, MA.

[4] BÉRARD, P., BESSON, G. and GALLOT, S. (1994). Embedding Riemannian manifolds by their heat kernel. *Geom. Funct. Anal*. **4** 373–398.

[5] BÉRARD, P. H. (1986). *Spectral Geometry*: *Direct and Inverse Problems*. *Lecture Notes in Math*. **1207**. Springer, Berlin. MR0861271

[6] CHEEGER, J., GROMOV, M. and TAYLOR, M. (1982). Finite propagation speed, kernel estimates for functions of the Laplace operator, and the geometry of complete Riemannian manifolds. *J. Differential Geom*. **17** 15–53.

[7] CHENG, M.-Y. and WU, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc*. **108** 1421–1434.

[8] COIFMAN, R. R. and LAFON, S. (2006). Diffusion maps. *Appl. Comput. Harmon. Anal*. **21** 5–30. MR2238665

[9] DEVROYE, L. P. and WAGNER, T. J. (1977). The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.* **5** 536–540. MR436442

[10] DO CARMO, M. P. and FLAHERTY, F. (1992). *Riemannian Geometry*. Birkhäuser, Boston, MA.

[11] DONOHO, D. L., GAVISH, M. and JOHNSTONE, I. M. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. Available at arXiv:1311.0851.

[12] DONOHO, D. L. and GRIMES, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* **100** 5591–5596.

[13] EL KAROUI, N. (2010). On information plus noise kernel random matrices. *Ann. Statist.* **38** 3191–3216. MR2722468

[14] EL KAROUI, N. and WU, H.-T. (2016). Graph connection Laplacian methods can be made robust to noise. *Ann. Statist.* **44** 346–372. MR3449771

[15] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC, Boca Raton, FL.

[16] GAO, T. (2016). The diffusion geometry of fibre bundles. Available at arXiv:1602.02330.

[17] GARCIA TRILLOS, N. and SLEPCEV, D. (2018). A variational approach to the consistency of spectral clustering. *Appl. Comput. Harmon. Anal.* To appear.

[18] GINÉ, E. and KOLTCHINSKII, V. (2006). Empirical graph Laplacian approximation of Laplace–Beltrami operators: Large sample results. In *High Dimensional Probability. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **51** 238–259. IMS, Beachwood, OH. MR2387773

[19] HEIN, M., AUDIBERT, J.-Y. and VON LUXBURG, U. (2005). From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians. In *Learning Theory. Lecture Notes in Computer Science* **3559** 470–485. Springer, Berlin. MR2203281

[20] JOHNSTONE, I. M. (2006). High dimensional statistical inference and random matrices. Available at arXiv:math/0611589v1.

[21] MOORE, D. S. and YACKEL, J. W. (1977). Consistency properties of nearest neighbor density function estimators. *Ann. Statist.* **5** 143–154. MR426275

[22] ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

[23] SINGER, A. (2006). From graph to manifold Laplacian: The convergence rate. *Appl. Comput. Harmon. Anal.* **21** 128–134.

[24] SINGER, A. and WU, H.-T. (2012). Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.* **65** 1067–1144.

[25] SINGER, A. and WU, H.-T. (2013). 2-D tomography from noisy projections taken at unknown random directions. *SIAM J. Imaging Sci.* **6** 136–175.

[26] SINGER, A. and WU, H.-T. (2017). Spectral convergence of the connection Laplacian from random samples. *Inf. Inference* **6** 58–123. MR3636868

[27] SMOLYANOV, O., WEIZSACKER, H. V. and WITTICH, O. (2007). Chernoff's theorem and discrete time approximations of Brownian motion on manifolds. *Potential Anal.* **26** 1–29.

[28] STEIN, E. M. and WEISS, G. (2016). *Introduction to Fourier Analysis on Euclidean Spaces. PMS* **32**. Princeton Univ. Press, Princeton, NJ.

[29] TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[30] VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.

[31] VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586. MR2396807

[32] WANG, X. (2015). Spectral convergence rate of graph Laplacian. Available at arXiv:1510.08110.

[33] WEINBERGER, K. Q. and SAUL, L. K. (2006). An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI* 1683–1686.

[34] WU, H.-T. and WU, N. (2018). Supplement to "Think globally, fit locally under the manifold setup: Asymptotic analysis of locally linear embedding." DOI:10.1214/17-AOS1676SUPP.

[35] ZHANG, Z. and WANG, J. (2006). MLLE: Modified locally linear embedding using multiple weights. In *Advances in Neural Information Processing Systems* 19 (*NIPS* 2006) 1593–1600. MIT Press, Cambridge, MA.

[36] ZHANG, Z. and ZHA, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.* **26** 313–338.

DEPARTMENT OF MATHEMATICS
AND DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708
USA
AND
MATHEMATICS DIVISION
NATIONAL CENTER FOR THEORETICAL SCIENCES
TAIPEI
TAIWAN
E-MAIL: hauwu@math.duke.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO M5S 2E4
CANADA
E-MAIL: n.wu@mail.utoronto.ca