SIAM J. MATH. DATA SCI.
Vol. 2, No. 3, pp. 770–808

Reconstruction of a Riemannian Manifold from Noisy Intrinsic Distances*

Charles Fefferman†, Sergei Ivanov‡, Matti Lassas§, and Hariharan Narayanan¶

Dedicated to the memory of Yaroslav Kurylev

Abstract. We consider the reconstruction of a manifold (or, invariant manifold learning), where a smooth Riemannian manifold M is determined from the intrinsic distances (that is, geodesic distances) of points in a discrete subset of M. In the studied problem, the Riemannian manifold (M, g) is considered as an abstract metric space with intrinsic distances, not as an embedded submanifold of an ambient Euclidean space. Let $\{X_1, X_2, \dots, X_N\}$ be a set of N sample points sampled randomly from an unknown Riemannian M manifold. We assume that we are given the numbers $D_{jk} = d_M(X_j, X_k) + \eta_{jk}$, where $j, k \in \{1, 2, \dots, N\}$. Here, $d_M(X_j, X_k)$ are geodesic distances, and η_{jk} are independent, identically distributed random variables such that the exponential moment $\mathbb{E}e^{|\eta_{jk}|}$ is finite. We show that when $N \sim C_0\delta^{-3n}(\log(1/\delta))^5 \log(1/\theta)$, with the probability $1-\theta$, it is possible to construct a manifold that approximates the Riemannian manifold (M, g) with the error δ. Here, C_0 depends on the intrinsic dimension n of M and the bounds for the diameter, sectional curvature, the injectivity radius of (M, g), and the the exponential moment of the noise. This problem is a generalization of the geometric Whitney problem with random measurement errors. We also consider the case when the information on the noisy distance D_{jk} of points X_j and X_k is missing with a certain probability. In particular, we consider the case when we have no information on points that are far away.

Key words. inverse problems, manifold learning, geometric Whitney problem

AMS subject classifications. 60D05, 53C21, 35R30

DOI. 10.1137/19M126829X

1. Introduction.

Let M be a manifold of dimension n, and let g be an intrinsic Riemannian metric on it. Assume that one is given the distances $d_M(X_j, X_k)$ with random measurement errors between points in a randomly sampled set $\{X_1, X_2, \dots, X_N\}$ of points of M. In this paper we ask how one can construct a Riemannian manifold (M^*, g^*) from these data so that the distance (in the Lipschitz sense) of the constructed manifold (M^*, g^*) to the original Riemannian manifold (M, g) can be estimated with a large probability. The

*Received by the editors June 13, 2019; accepted for publication (in revised form) May 12, 2020; published electronically September 2, 2020.

https://doi.org/10.1137/19M126829X

Funding: The first author was partly supported by AFOSR grant DMS-1265524 and NSF grant FA9550-12-1-0425. The second author was partly supported by RFBR grants 14-01-00062 and 17-01-00128-A. The third author was supported by Academy of Finland grants 273979, 284715, and 312110. The fourth author was partly supported by NSF grant DMS-1620102, DAE project 12-R&D-TFR-5.01-0500, and a Ramanujan Fellowship.

†Department of Mathematics, Princeton University, Princeton, NJ 08544 USA (cf@math.princeton.edu).

‡St. Petersburg Department of Steklov Institute of Mathematics, 191023 St. Petersburg, Russia (svivanov@pdmi.ras.ru).

§Department of Mathematics and Statistics, University of Helsinki, Helsinki, FIN-00014, Finland (Matti.Lassas@helsinki.fi).

¶Tata Institute for Fundamental Research, Mumbai 400005, India (harius80@gmail.com).

need to construct the non-Euclidean, intrinsic metric is encountered in many applications; for example, in medical and seismic imaging as discussed in section 5.

In traditional manifold learning, for instance, by using the ISOMAP algorithm introduced in the seminal paper [58], one often aims to map points X_j to points $Y_j = F(X_j)$ in a Euclidean space $\mathbb{R}^m$, where $m \geq n$ is as small as possible so that the Euclidean distances $\|Y_j - Y_k\|_{\mathbb{R}^m}$ are close to the intrinsic distances $d_M(X_j, X_k)$, and find a submanifold $\widetilde{M} \subset \mathbb{R}^m$ that is close to the points Y_j. This method has turned out to be very useful, in particular, in finding the topological manifold structure of the manifold (M, g). It has been shown that when the original manifold (M, g) has a vanishing Riemann curvature and satisfies certain convexity conditions, the manifold reconstructed by the ISOMAP approaches the original manifold as the number of the sample points tends to infinity (see the results in [5, 18, 20] for ISOMAP and [60] for the continuum version of ISOMAP). We note that for a general Riemannian manifold (M, g), the Nash embedding theorem implies that there exists a smooth isometric embedding $F : M \to \mathbb{R}^m$, where m is sufficiently large [45, 46]. For such a map F, the embedded manifold $F(M) = \widetilde{M} \subset \mathbb{R}^m$, with the metric inherited from the metric of $\mathbb{R}^m$, is isometric to the manifold (M, g). However, a numerical construction of the Nash embedding map F is difficult (see [59] on numerical techniques based on the Nash embedding theorem).

We emphasize that the construction of an isometric embedding $f : M \to \mathbb{R}^n$ is outside the context of this paper.

One can overcome the difficulties related to the construction of the Nash embedding by formulating the problem in a coordinate invariant way: Given the geodesic distances of points sampled from a Riemannian manifold (M, g), we construct a manifold M^* with an intrinsic metric tensor g^* so that the Lipschitz distance of (M^*, g^*) to the original manifold (M, g) is small. The construction of abstract manifolds from the distances of sampled data points has also been considered by Coifman and Lafon [13] and Coifman et al. [11, 12] using "Diffusion Maps," and by Belkin and Niyogi [3] using "EigenMaps," where the data points are mapped to the values of the approximate eigenfunctions or diffusion kernels at the sample points. These methods construct a nonisometric embedding of the manifold M into $\mathbb{R}^m$ with a sufficiently large m. This construction is continued in [48] by computing an approximation of the metric tensor g by using finite differences to find the Laplacian of the products of the local coordinate functions. In this paper our aim is to study the problem of constructing a Riemannian manifold when distances are given with random errors and to use metric geometry to construct (M^*, g^*) so that the Lipschitz distance of (M^*, g^*) and (M, g) can be estimated in terms of the number of the sample points with a large probability. More specifically, in this paper we show how the noisy distances in a randomly sampled set of M can be used to find distances with small errors in a sparse subset of sample points. After obtaining this information, it is possible to construct an approximation of the Riemannian manifold in the Lipschitz sense, using interpolation results, as in [24]. We emphasize that we consider M^* as an abstract manifold that is not isometrically embedded into a Euclidean space but where the metric is given by a metric tensor g^* that is constructed from the above data. In this paper, we extend the results of [24] that deal with the question how a smooth manifold that approximates a manifold (M, g) can be constructed when one is given the distances of the points of in a discrete subset X of M with small deterministic errors. In this paper we extend these results to two directions. First,

the discrete set is randomly sampled and the distances have (possibly large) random errors. Second, we consider the case when some distance information is missing.

1.1. The main result. Let $n \geq 2$ be an integer, $\Lambda > 0$, $D > 0$, and $i_0 > 0$. Let (M, g) be a compact Riemannian manifold of dimension n such that

$$\text{(i) } \|\mathrm{Sec}_M\|_{L^\infty(M)} \leq \Lambda^2, \quad \text{(ii) } \mathrm{diam}\,(M) \leq D, \quad \text{(iii) } \mathrm{inj}\,(M) \geq i_0, \tag{1.1}$$

where Sec_M is the sectional curvature of (M, g), $\mathrm{diam}\,(M)$ is the diameter of (M, g), and $\mathrm{inj}\,(M)$ is the injectivity radius of (M, g), that is, the minimal radius of Riemannian normal coordinates. Let $d_M(x, y)$ denote the intrinsic (or geodesic) distance of the points $x, y \in M$, determined by the metric tensor g corresponding to the line element $ds^2 = g_{jk}(x)dx^j dx^k$. Here and below, we use Einstein's summation convention and sum over indexes appearing as super- and subindexes.

Let $(\Omega, \Sigma, \mathbb{P})$ be a complete probability space, let $\mathcal{B}$ be the σ-algebra of Borel sets on M, and let $\nu : \mathcal{B} \to [0, 1]$ be a probability measure on M. Let dV_g be the Riemannian volume on (M, g). Assume that the Radon–Nikodym derivative of ν satisfies

$$0 < \rho_{min} \leq \frac{d\nu}{dV_g} \leq \rho_{max}, \quad \text{where } \rho_{min}, \rho_{max} \in \mathbb{R}_+. \tag{1.2}$$

Definition 1.1. *Let X_j, $j = 1, 2, \ldots, N$ be independent, identically distributed (i.i.d.) random variables having the distribution ν. Let $\mu \in \mathbb{R}$, $\sigma \geq 0$, $\beta \geq 1$, and η_{jk} be random variables satisfying*

$$\mathbb{E}\eta_{jk} = \mu, \quad \mathbb{E}((\eta_{jk} - \mu)^2) = \sigma^2, \quad \mathbb{E}e^{|\eta_{jk} - \mu|} = \beta. \tag{1.3}$$

We assume that all random variables η_{jk} and X_j are independent. Let

$$D_{jk} = d_M(X_j, X_k) + \eta_{jk} \tag{1.4}$$

be the geodesic distances of points X_j and X_k, measured with errors η_{jk}.

Note that the above assumptions are satisfied when $\eta_{jk} \sim N(0, \sigma^2)$ are i.i.d. Gaussian random variables and $\beta \leq 2e^{\sigma^2}$. We are mostly interested in the case when σ is fixed and N is large.

Definition 1.2. *The partial data is given by*

$$\overline{D}_{jk} = D_{jk}^{(\text{partial data})} = \begin{cases} D_{jk} & \text{if } Y_{jk} = 1, \\ \text{"missing"} & \text{if } Y_{jk} = 0, \end{cases}$$

where Y_{jk} are random variables taking values in $\{0, 1\}$ and $j, k \in \{1, 2, \ldots, N\}$. We assume that Y_{jk} are independent of random variables $X_{j'}$ for all $j' \in \{1, 2, \ldots, N\} \setminus \{j, k\}$ and of $\eta_{j''k''}$ for all j'' and k''. Below, to make $\overline{D}_{jk}$ a real-valued random variable, the "missing" value, is replaced by the real value $2D$.

Assume that the conditional probability of the event $\{Y_{jk} = 1\}$, when X_j and X_k are known, is

$$\mathbb{P}(Y_{jk} = 1 \,|\, X_j, X_k) = \Phi(X_j, X_k). \tag{1.5}$$

More precisely, when $\mathcal{B}_{jk} \subset \Sigma$ is the σ-algebra generated by the random variables X_j and X_k above in formula (1.5), we use the notation $\mathbb{P}(Y_{jk} = 1 \,|\, X_j, X_k) = \mathbb{P}(Y_{jk} = 1 \,|\, \mathcal{B}_{jk})$. Here, $\Phi : M \times M \to [0,1]$ is a measurable function such that there is a function $\Phi^1 : [0,\infty) \to [0,1]$ so that $s \mapsto \Phi^1(s)$ is nonincreasing and

(1.6)
$$\Phi^1(0) = \phi_0, \quad \|\Phi^1\|_{C^1(\mathbb{R})} \le H, \quad \lambda_1 \Phi^1(d_M(x,y)) \le \Phi(x,y) \le \lambda_2 \Phi^1(d_M(x,y)), \quad x, y \in M,$$

where $0 < \lambda_1 \le 1 < \lambda_2$, $H \ge \Lambda^{-1}$, and $0 < \phi_0 \le 1$.

The parameters $H, \phi_0, \lambda_2, \lambda_1$, and β (and $\sigma^2 \le 2\beta$) determine the number

$$p_{H,\phi_0,\beta,\lambda_2,\lambda_1} = H^{n^2+2n+1} \phi_0^{-(2n^2+8n+7)} \Big(\frac{\lambda_2}{\lambda_1}\Big)^{2n^2+10n+7} (\log(e+\beta))^4 \tag{1.7}$$

that will be used as a coefficient in the formula for the required sample size. Note that here n is the intrinsic dimension of M that is not very large in typical applications. Also, for $t \in \mathbb{R}$ we denote by $\lfloor t \rfloor$ the largest integer m such that $m \le t$.

The parameters and constants that we use in the text are summarized in the table below:

n	the dimension of M; see (1.1),	$c_0, \widehat{c}_0$	(1.19),
D	the bound for the diameter of M; see (1.1),	r_0, r_1, ϕ_1	(1.20),
Λ^2	the bound for the curvature of M; see (1.1),	a	(3.2),
i_0	the bound for the injectivity radius; see (1.1),	$\widehat{a}, c_2$	(3.7),
μ, σ^2	the expectation and the variance of the noise η; see (1.3),	b, c_1	(3.19),
β	the expectation of $e^{\lvert\eta-\mu\rvert}$; see (1.3),	$L, \varepsilon(L), \varepsilon_2$	(3.29),
ρ_{min}, ρ_{max}	the bounds for the sampling density ν; see (1.2),	u_0, u_1, u_2	(4.1),
$\Phi(X_j, X_k)$	the probability that the distance $d(X_j, X_k)$ is known,	ρ, c_3	(4.1),
λ_1, λ_2	the lower and upper bounds for Φ/Φ^1; see (1.6),	h_0, ε_3	(4.10),
H, ϕ_0	the bound for the derivative of Φ_0, $\phi_0 = \Phi^1(0)$; see (1.6),	C_4	(4.25).

Our main result is the following.

Theorem 1. *Suppose we are given the dimension $n \ge 2$ and the geometric bounds D, Λ, i_0 for the studied manifolds, the density bounds ρ_{min}, ρ_{max} for the sampling density, the parameters μ, σ, β for the statistics of the noise, and the missing data parameters $H \ge \Lambda^{-1}$, $0 < \phi_0 \le 1$, and $\lambda_1, \lambda_2 > 0$.*

Then there are $\delta_0 > 0$ and $C_0 > 0$, depending on $n, D, \Lambda, i_0, \rho_{min}, \rho_{max}$ (but not depending on $H, \phi_0, \lambda_2, \lambda_1, \mu, \sigma$, or β), and there is $C_1 > 0$, depending on n, such that the following holds for $\theta \in (0, \frac{1}{2})$.

Let M be a compact n-dimensional manifold satisfying (1.1), $0 < \delta < \delta_0$, and

$$N = \left\lfloor C_0\, p_{H,\phi_0,\beta,\lambda_2,\lambda_1}\, \delta^{-3n} Q_n(\delta) \left(\log\left(\frac{1}{\theta}\right) + \log\left(\frac{1}{\delta}\right)\right)\right\rfloor, \tag{1.8}$$

where $Q_n(\delta) = 1$ when $n \ge 3$, and $Q_n(\delta) = \log^4(1/\delta)$ when $n = 2$, and $p_{H,\phi_0,\beta,\lambda_2,\lambda_1}$ is given by (1.7). Also, let $\overline{D}_{jk}$, $j,k = 1,2,\dots,N$, be as in Definitions 1.1 and 1.2. Suppose that one is given samples of the random variables $\overline{D}_{jk}$ for $j,k = 1,2,\dots,N$. Then one can construct a compact, smooth n-dimensional Riemannian manifold (M^*, g^*) that with the probability $1-\theta$ approximates the manifold (M, g) in the following way.

(1) *There is a diffeomorphism* $F : M^* \to M$ *satisfying*

$$\frac{1}{L} \le \frac{d_M(F(x), F(y))}{d_{M^*}(x, y)} \le L \quad \text{for all } x, y \in M^*, \tag{1.9}$$

where $L = 1 + C_1\delta$, *that is, the Lipschitz distance of the metric spaces* (M^*, g^*) *and* (M, g) *satisfies* $d_{Lip}((M^*, g^*), (M, g)) \le \log L$.

(2) *The sectional curvature* Sec_{M^*} *of* M^* *satisfies* $|Sec_{M^*}| \le C_1\Lambda^2$.

(3) *The injectivity radius* $inj(M^*)$ *of* M^* *satisfies*

$$inj(M^*) \ge \min\{(C_1\Lambda)^{-1}, (1 - C_1\delta)inj(M)\}.$$

We note that in Theorem 1 the dimension n of the manifold, the bound D for the diameter, and the expectation μ and the variance σ^2 of the noise are assumed to be a priori known, and our reconstruction algorithm requires these parameters. The determination of the parameters n, μ, and σ from the noisy distances $\overline{D}_{jk}$ is discussed in Remark 4.10.

To the best of the authors' knowledge, the results are also new in the case when there is no missing data, that is, $\Phi(x, y) = 1$ for all $x, y \in M$. Observe that the factor $(1/\delta)^{3n}$ in the estimate (1.8) for the number of sampled points is exponential in dimension n. In contrast, for a manifold that is a priori known to be an affine subspace of a Euclidean space, the number of the needed sample points depends only polynomially on the dimension n; cf. [25, Lemma 29].

The proof of Theorem 1 below and the results in [24] give a procedure in which the input is the noisy distances $\overline{D}_{jk}$ and the output is a submanifold $M^* \subset \mathbb{R}^d$ (where d only depends on n, D, Λ, and i_0) and a metric tensor g^* on M^*.

Remark 1.3. Theorem 1 concerns the regime where the noise level σ is a fixed constant, the number of points N is large, and we are interested in the situation where we want the probability θ of a wrong final reconstruction to be very small. This is reflected by the fact that the probability θ of obtaining a wrong reconstruction only appears in the logarithmic term $\log(\theta^{-1})$.

1.2. The idea of the proof and three nets of points on the manifold. Let us assume that $N = N_0 + N_1 + N_2$, where $N_0, N_1, N_2 \in \mathbb{Z}_+$. We are interested in the case when $N_2 > N_1 > N_0$. We consider the subsets of the sample points and call $S_0 = \{X_1, \dots, X_{N_0}\}$ a *sparse net* on M and compute approximate distances between the points in the net S_0 by using two auxiliary sets, called the *intermediate net* $S_1 = \{X_{N_0+1}, \dots, X_{N_0+N_1}\}$ and the *dense net* $S_2 = \{X_{N_0+N_1+1}, \dots, X_{N_0+N_1+N_2}\}$. Roughly speaking, when θ is fixed and $\delta \to 0$, we can choose the sizes of the nets to be $N_0 \sim \delta^{-n/2}\log(1/\delta)$, $N_1 \sim \delta^{-3n}\log(1/\delta)$, and $N_2 \sim \delta^{-3n}(\log(1/\delta))^5$.

Our strategy to prove Theorem 1 is to first prove (see Theorem 2) that with the probability $1 - \theta$ the points X_j, $j = 1, 2, \dots, N_0$, form a $C\delta^{1/2}$-dense subset of M, and that it is possible to construct the distances $d_M(X_j, X_k)$, for $X_j, X_k \in S_0$, with deterministic errors that are smaller than $C\delta^{3/2}$. This is done using the noisy distances between the points in the sets S_0, S_1, and S_2. Once such distances are found, then the possibility to construct a smooth manifold (M^*, g^*) that approximates the original manifold (M, g) with the error $C\delta$ is guaranteed by the

deterministic results given in Corollary 1.10 in [24], which is slightly extended in Proposition A.1 in Appendix A. The main technique in the proof—using the subsets of samples as networks with different sampling densities in order to estimate the distances—is tightly related to the approaches presented by Coifman, Haddad, and Kushnir in [32, 42]. In these works, a reference set is used to estimate the Riemannian metric in a different, but related, context. In [42] one chooses a sparse subset of the sample points, called the reference points, and considers the distances of all sample points to these reference points. These distances are used as auxiliary coordinates of the sample points. In [32] one considers natural images, e.g., photographs, and defines for each pixel a patch that consists of the values in the neighboring pixels. Then, one chooses a sparse subset of pixels, called a reference set, and uses the patches of the reference set as a basis, where all the other patches are approximately represented. The patches of the reference set are used to construct a data-based nonlocal filter that is efficiently used for noise removal. The methods in [42] are similar to those in our paper in the sense that we also consider a sparse subset of the sample points and use the distances of all the other points to the sparse subset to reconstruct the unknown manifold structure. Also, similarly to [32] we use weighted averaging of the data to remove noise.

To prove Theorem 2, we use the noisy distances of the points in the set S_1 to the points in S_2, and the distances of the points in S_1 to the points in S_0. With these, we find distances of the points in the sparse net S_0 with small errors. The random sets S_0, S_1, and S_2 correspond to the index sets

(1.10)
$$I^{(0)} = \{1, 2, \dots, N_0\}, \quad I^{(1)} = \{N_0+1, \dots, N_0+N_1\}, \quad I^{(2)} = \{N_0+N_1+1, \dots, N_0+N_1+N_2\}.$$

Below, we say that a set $Y \subset M$ is δ-dense in M if for all $p \in M$ there is $y \in Y$ such that $d_M(p, y) < \delta$. A δ-dense subset S of M is often called an *δ-net*.

Let us give an overview of the proof of Theorem 1. First, we use the dense net S_2 to compute

$$k_\Phi(y, z) = \int_M |d_M(y, x) - d_M(z, x)|^2 \, \Phi(y, x)\Phi(x, z) \, d\nu(x) \tag{1.11}$$

for y and z in the intermediate net S_1; see Proposition 3.5. This corresponds to taking the average of function $|d_M(y, x) - d_M(z, x)|^2$ over all those sample points $x \in S_2$ for which the distances $d_M(y, x)$ and $d_M(z, x)$ are not missing (see (3.15)).

Note that when the product $\Phi(y, x)\Phi(x, z)$ is small, there are only a small number of sample points $x \in S_2$ for which the value of the function $|d_M(y, x) - d_M(z, x)|^2$ can be computed; then the estimator for the function $k_\Phi(y, z)$ is not reliable. Thus the reliability of the estimator for the function $k_\Phi(y, z)$ is measured by

$$A_\Phi(y, z) = \int_M \Phi(y, x)\Phi(x, z) \, d\nu(x). \tag{1.12}$$

Indeed, when $A_\Phi(y, z)$ is larger than the threshold value $b > 0$, the obtained estimator for the function $k_\Phi(y, z)$ is reliable with a large probability. When we compute an estimator for the function $k_\Phi(y, z)$ using a sample imitating the integral in (1.11), we can also compute an estimator for $A_\Phi(y, z)$ (see (3.16)).

Second, we are going to use the set S_1 to compute the approximate distances $d^{app}(y_1, y_2)$ of the points y_1 and y_2 in the sparse net S_0 using reliable distances $k_\Phi(y,z)^{1/2}$. We do this by computing estimators for the functions (see Definition 4.2)

$$
\begin{aligned}
(1.13)\qquad & Q(y_1,y_2) = \frac{V_\Phi(y_1,y_2)}{W_\Phi(y_1,y_2)}, \quad \text{where} \\
& V_\Phi(y_1,y_2) = \int_M \widetilde{\psi}_1\left(\frac{A_\Phi(y_1,z)}{b}\right)\psi_\rho(k_\Phi(y_1,z))\Phi(z,y_2)d_M(z,y_2)d\nu(z), \\
& W_\Phi(y_1,y_2) = \int_M \widetilde{\psi}_1\left(\frac{A_\Phi(y_1,z)}{b}\right)\psi_\rho(k_\Phi(y_1,z))\Phi(z,y_2)d\nu(z)
\end{aligned}
$$

(see Figure 1 (right)), where $\widetilde{\psi}_1 \in C^\infty(\mathbb{R})$ is a cut-off function such that $\widetilde{\psi}_1(t) = 0$ for $t < 1$ and $\widetilde{\psi}_1(t) = 1$ for $t > 2$, and $\psi_\rho \in C^\infty(\mathbb{R})$ is a cut-off function such that $\psi_\rho(s) = 1$ for $s < \rho^2$ and $\psi_\rho(s) = 0$ for $s > 2\rho^2$. Here, $\widetilde{\psi}_1(A_\Phi(y_1,z)/b)\psi_\rho(k_\Phi(y,z))$ is the smoothened version of the indicator function of the set

$$
(1.14)\qquad D_\Phi(y_1,\rho) = \{z \in M :\ k_\Phi(y,z) < \rho^2,\ A_\Phi(y_1,z) \geq b\}.
$$

The set $D_\Phi(y_1,\rho)$ is a Lipschitz approximation the union of the ball $B_M(y_1,\rho)$.

Then, roughly speaking, we compute an estimator for the function $V_\Phi(y_1,y_2)$, computing the averages of the function $\widetilde{\psi}_1(A_\Phi(y_1,z)/b)d_M(z,y_2)$ over all the sample points z in the medium net S_1 that are in the set $D_\Phi(y_1,\rho)$ and for which data on the distance $d_M(z,y_2)$ is not missing. At the same time, we compute an estimator for the function $W_\Phi(y_1,y_2)$ by computing averages of the function $\widetilde{\psi}_1(A_\Phi(y_1,z)/b)\psi_\rho(k_\Phi(y,z))$ over the same sample points. The idea is that when $W_\Phi(y_1,y_2)$ is larger than the threshold $u > 0$, the estimators computed from random data for the functions $V_\Phi(y_1,y_2)$, $W_\Phi(y_1,y_2)$, and $Q(y_1,y_2)$ are reliable with a large probability. Then, we define for $y_1, y_2 \in S_0$

$$
(1.15)\qquad d^{app}(y_1,y_2) = \begin{cases} Q(y_1,y_2) & \text{if } W_\Phi(y_1,y_2) > u, \\ D & \text{otherwise (see Figure 1 (left)).} \end{cases}
$$

Then there is r_1 such that $d^{app}(y_1,y_2)$ approximates the true distance $d_M(y_1,y_2)$ with a small error ε_1 when $d_M(y_1,y_2) < r_1$, and moreover, if $d_M(y_1,y_2) \geq r_1$, then $d^{app}(y_1,y_2) > r_1 - \varepsilon_1$. In other words, we can construct the distances $d_M(y_1,y_2)$ with a large probability and with small errors for all the points y_1 and y_2 in S_0 that are close to each other.

After the above constructions, we will use Proposition A.1 in Appendix A, concerning a reconstruction of a Riemannian manifold, when we are given distances with small (deterministic) errors. This result is an improved version of the results given in [24, Corollary 1.10].

1.3. Earlier results for the submanifolds of $\mathbb{R}^n$ and graphs of functions. In dimensionality reduction and in traditional manifold learning, the aim is to transform data, consisting of points in a d-dimensional space, that are near an n-dimensional submanifold M, where $d \gg n$ to a set of points in the low-dimensional space $\mathbb{R}^m$ close to an n-dimensional submanifold, where $d > m \geq n$. During transformation, all of them try to preserve some geometric

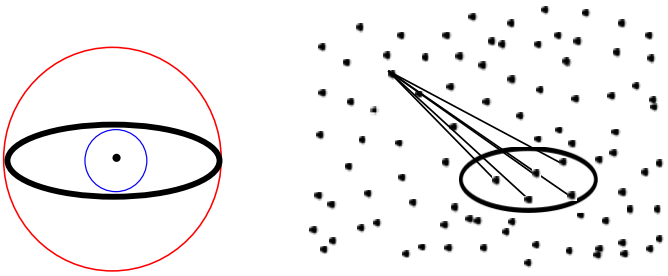

Figure 1. *Left: The set $D_\Phi(y_1,\rho)$ satisfies $B_M(y,\kappa^{-1}\rho)\subset D_\Phi(y_1,\rho)\subset B_M(y,\rho)$, and thus $D_\Phi(y_1,\rho)$ can be considered as an approximate ρ-neighborhood of the the point y. Right: The approximate distance $d^{app}(y_1,y_2)$ in the formula (1.15) is the average of the distances from y_2 to the points in the neighborhood $D_\Phi(y_1,\rho)$. Later, we approximate $d^{app}(y_1,y_2)$ by taking the average distances of y_2 to the points in $S_1\cap D_\Phi(y_1,\rho)$, where S_1 is the intermediate net of the sample points.*

properties, such as appropriately measured distances between the points of the original data set (see [9, 10, 58]). Perhaps the most basic of such methods is principal component analysis (PCA) (see [54] and the references therein), where one projects the data points onto the span of n eigenvectors corresponding to the n largest eigenvalues of the $(d\times d)$ covariance matrix of the data points.

In the case of multidimensional scaling (MDS) [14], the algorithm attempts to preserve the pairwise distances between points. One minimizes a certain stress function which captures the total error in pairwise distances between the data points and between their lower-dimensional counterparts. For instance, given points $(x_j)_{j=1}^N$, $x_j\in\mathbb{R}^d$, one tries to find $(y_j)_{j=1}^N$, $x_j\in\mathbb{R}^m$, which is an (approximate) minimizer of the stress function

$$\min_{y_j\in\mathbb{R}^m}\left(\sum_{i,j=1}^{N}(\|y_i-y_j\|_{\mathbb{R}^m}-d_{ij})^2\right), \tag{1.16}$$

where $d_{ij}=\|x_i-x_j\|_{\mathbb{R}^d}$ are the Euclidean distances of points x_i and x_j.

ISOMAP [58] attempts to improve on MDS by trying to capture geodesic distances between points while projecting. For each data point x_i in the data set $\mathcal{X}=(x_j)_{j=1}^N$, $x_j\in\mathbb{R}^d$, a neighborhood graph is constructed using the K-neighbors of x_i, that is, using the K nearest points of $\mathcal{X}$ to x_i, the edges carrying the length between the points. Now the shortest distance between points is computed in the resulting global graph containing all the neighborhood graphs using a standard graph theoretic algorithm such as Dijkstra's. Let $D^G=[d_{ij}^G]$ be the $N\times N$ matrix of graph distances. Then MDS is used to find $(y_j)_{j=1}^N$, $x_j\in\mathbb{R}^m$ that (approximately) solves the minimization problem (1.16) with distances d_{ij} replaced by d_{ij}^G. If the data set $\mathcal{X}=(x_j)_{j=1}^N$ consists of the δ-dense set points of a submanifold $M\subset\mathbb{R}^d$ with small δ, then ISOMAP tries to find an approximation for isometric embedding, that is, a map $F:M\to\mathbb{R}^m$ for which

$$\|F(x)-F(y)\|_{\mathbb{R}^m}\approx d_M(x,y),\quad x,y\in M, \tag{1.17}$$

where $d_M(x,y)$ is the intrinsic distance of the points x and y of the isometrically embedded manifold $M\subset\mathbb{R}^d$. As noted above, ISOMAP has turned out to be very useful in finding

the topological manifold structure related to the data. The convexity and flatness conditions that guarantee that the ISOMAP algorithm reconstructs the original manifold are studied in [5, 18, 20] in relation to ISOMAP and in [60] in relation to the continuum ISOMAP.

In the seminal papers [11, 12, 13] on Diffusion Maps and [3] on EigenMaps, a complete graph or a neighborhood graph is built on the data points sampled from a manifold (M, g), and each edge is assigned a weight $a(x, y)$ that is a function of the distance of the points x and y. The normalized version of the kernel $a(x, y)$ defines a diffusion or a differential operator on M having the eigenfunctions $\phi_j(x)$. These functions can be used to construct a nonisometric embedding $x \mapsto (\phi_j(x))_{j=1}^m$ of the manifold M into $\mathbb{R}^m$. This construction is continued in [48] by computing an approximation of the metric tensor g by using finite differences to find the Laplacian of the products of the local coordinate functions. Also the practical algorithms performing the construction of the manifold using Diffusion Maps are considered in [11, 12, 13] and with nets having different densities in [32, 42].

Other topological embedding methods for manifolds, based on heat kernels or eigenfunctions, have been developed in [3, 19, 38]. Moreover, locally linear construction methods are studied in [6, 43, 44, 53, 61].

An extensively studied question in manifold learning is the manifold denoising problem. In this problem one aims to find a submanifold M of $\mathbb{R}^d$ when one is given the points $y_j = x_j + \xi_j$, $j = 1, 2, \dots, N$, where $x_j \in M$ are randomly sampled points of the submanifold M, and $\xi_j \in \mathbb{R}^d$ are independent random errors having, e.g., a Gaussian distribution. This problem has been studied in [1, 17, 23, 25, 28, 29, 30, 31, 34, 49, 52]. Such data (y_j) is closely related to the data considered in this paper, but as the errors in the observed distances, that is, $\varepsilon_{jk} = |y_j - y_k| - |x_j - x_j|$, depend on the points x_j and x_k, we see that the data set studied in this paper and those studied in the manifold denoising problem are quite different.

The construction of a surface approximating a set of points in $\mathbb{R}^d$ is closely related to the classical Whitney problem. This problem is the construction of a function $F(x) \in C^k(\overline{S})$, where $S \subset \mathbb{R}^n$ is open, which is equal to a given function $f(x)$ on K, where $K \subset S$. This problem has been studied in different norms in [21, 22, 26] and the interpolation results on the Whitney problem have been applied for a submanifold of $\mathbb{R}^d$ in [23, 27].

1.4. Notations. To simplify the formulas for the constants C_k below, we will assume that $\Lambda, D \geq 1$, $i_0 \leq 1$, and $H \geq 2$. Let vol_g be the Riemannian volume on (M, g), and let $p \in M$ be the exponential map $\exp_p : T_pM \to M$, which defines a smooth surjective map $\exp_p : \{\xi \in T_pM : \ \|\xi\|_g < D + 1\} \to M$. By the Bishop–Gromov inequality [50, Chap. 9, Lem. 1.6], the function $r \mapsto \mathrm{vol}_g(B(x, r))/v(n, -\Lambda^2, r)$ is nonincreasing and bounded by 1, where $\mathrm{vol}_g(B(x, r))$ is the volume of the ball $B(x, r) \subset M$, and $v(n, -\Lambda^2, R)$ is the volume of the ball of radius r in the hyperbolic space of dimension n having the curvature $-\Lambda^2$. Hence

$$\mathrm{vol}_g(M) \leq V_0 = v(n, -\Lambda^2, D) \leq \omega_n \left(\frac{\sinh(\Lambda D)}{\Lambda D} \right)^{n-1} D^n, \tag{1.18}$$

where ω_n is the volume of the unit ball in $\mathbb{R}^n$ (see [50, Chap. 6, Cor. 2.4]). Moreover, as $v(n, -\Lambda^2, \rho) \geq \omega_n \rho^n$, we have

$$\frac{\mathrm{vol}_g(B(x, \rho))}{\mathrm{vol}_g(M)} \geq \frac{v(n, -\Lambda^2, \rho)}{v(n, -\Lambda^2, D)} \geq \widehat{c}_0 \rho^n, \quad \nu(B(x, \rho)) \geq c_0 \rho^n, \tag{1.19}$$

where $\widehat{c}_0 = \frac{\omega_n}{V_0}$ and $c_0 = \frac{\rho_{min}}{\rho_{max}}\widehat{c}_0$. Let

$$r_0 = \min\left(\frac{1}{2H}\phi_0, i_0, \frac{\pi}{2\Lambda}\right), \quad r_1 = \frac{1}{2}r_0, \quad \phi_1 = \frac{\lambda_1}{2}\phi_0 \le \frac{1}{2}. \tag{1.20}$$

Then Definition 1.2 implies

$$\Phi(x,y) \ge \lambda_1 \Phi^1(r_0) \ge \phi_1 \quad \text{for } d(x,y) \le r_0. \tag{1.21}$$

A note on the used constants. Below, we will make frequent use of the constants c_j, C_j, etc. Unless otherwise stated, these constants only depend on $n, D, \Lambda, i_0, \rho_{min}, \rho_{max}$. In particular, c_j and C_j do not depend on the parameters $H, \phi_0, \lambda_1, \lambda_2, \mu, \sigma$, or β, as we keep a record of how the coefficients in the inequalities depend on these parameters.

2. Reformulation of the main result with several parameters.

Our aim is to prove the following result yielding Theorem 1 when it is combined with the results in Appendix A on manifold reconstruction with small deterministic errors.

Theorem 2. *Let $n \ge 2$ and $D, \Lambda, i_0, \rho_{min}, \rho_{max}, \lambda_1, \lambda_2, \sigma, \beta > 0$, $\mu \in \mathbb{R}$, $H \ge \Lambda^{-1}$, and $\phi_0 \in (0,1]$ be given. Then there are $C_2 > 1$ and $\widehat{\delta}_1 < 1$ depending on $n, D, \Lambda, i_0, \rho_{min}$, and ρ_{max}, such that the following holds for $\delta_1 \le \widehat{\delta}_1$, $\varepsilon_1 \le \delta_1$, and $\theta \in (0, 1/2)$. Let*

$$N_0 = \left\lfloor C_2 \delta_1^{-n}\left(\log\left(\frac{1}{\theta}\right) + \log\left(\frac{1}{\delta_1}\right)\right)\right\rfloor, \tag{2.1}$$

$$N \ge \left\lfloor C_2\, p_{H,\phi_0,\beta,\lambda_2,\lambda_1}\, \varepsilon_1^{-2n} Q_n(\varepsilon_1)\left(\log\left(\frac{1}{\theta}\right) + \log\left(\frac{1}{\delta_1}\right) + \log\left(\frac{1}{\varepsilon_1}\right)\right)\right\rfloor, \tag{2.2}$$

where $Q_n(\varepsilon_1) = 1$ when $n \ge 3$, and $Q_n(\varepsilon_1) = \log^4(1/\varepsilon_1)$ when $n = 2$. Moreover, $p_{H,\phi_0,\beta,\lambda_1,\lambda_2}$ is given as in (1.7).

Also, let X_j and $\overline{D}_{jk}$, $j,k = 1,2,\dots,N$ be as in Definitions 1.1 and 1.2. Suppose that we are given the samples of the random variables $\overline{D}_{jk}$ for $j,k = 1,2,\dots,N$. Let r_1 be given as in (1.20). Then, with the probability $1-\theta$, the set $\{X_j : j = 1,2,\dots,N_0\}$ is a δ_1-net in M, and one can determine the approximate distances $d^{(a)}(X_j, X_{j'})$ so that the following holds.

For all $j, j' \in \{1,2,\dots,N_0\}$,

$$|d^{(a)}(X_j, X_{j'}) - d_M(X_j, X_{j'})| \le \varepsilon_1 \quad \text{if } d_M(X_j, X_{j'}) < r_1, \tag{2.3}$$

$$d^{(a)}(X_j, X_{j'}) \ge r_1 - \varepsilon_1 \quad \text{if } d_M(X_j, X_{j'}) \ge r_1. \tag{2.4}$$

In the case when the probability $\Phi(x,y)$, when the information on the distance of x and y is not missing, is bounded from below by a positive constant, that is, $\Phi(x,y) \ge \phi_2$ for all $(x,y) \in M \times M$, the inequality (2.3) holds for all $X_j, X_{j'}$, $j, j' \in \{1,2,\dots,N_0\}$.

We note that when $\Phi(x,y)$ is bounded from below by a positive constant, $\Phi(x,y) \ge \lambda_1\phi_0$, we can choose the function Φ^1 to be equal to the constant ϕ_0. Without loss of generality, we can assume that in this case $\phi_2 = \lambda_1\phi_0$.

2.1. A sketch of an algorithmic based on Theorem 2. The proof of Theorem 2 yields an algorithm that, with a large probability, produces distances in the sparse net $\{X_1, \dots, X_{N_0}\} \subset M$ with small deterministic errors. After such distances are constructed, one can proceed to reconstruct a manifold (M^*, g^*), described in Theorem 1 using some of the alternative methods discussed below.

Below, we use the smooth cut-off functions $\psi_1(t) = \psi_1(t)$ and $\widetilde{\psi}_1(t) = 1 - \psi_1(t)$, where $\psi_1 \in C_0^\infty((-2, 2))$ is equal to 1 near zero and satisfies the properties (3.21).

The following algorithm implements the construction in the proof of Theorem 2. The parameters that the algorithm needs are the geometric parameters D, Λ, i_0; the stochastic parameters of the noise σ, β; the bounds for the sampling density ρ_{min}, ρ_{max}; the parameters related to the missing data $H, \phi_0, \lambda_1, \lambda_2$; and the bounds for errors, namely δ_1, ε_1 and the upper bound θ for the probability that the construction is not successful.

The algorithm NRD below uses the additional parameters

$$\kappa = c_1 H^{-(n/2+1)} \phi_0^{n+4}, \quad b = c_2 \phi_0^2, \quad \rho = \frac{1}{4}\kappa\varepsilon_1, \quad u_2 = c_3 \phi_0 \rho^n, \quad L = 4\log^2\left(C_4 \beta^2 \rho^{-1}\right),$$

where c_1 is given in (3.19), c_2 in (3.7), c_3 in (4.1), and C_4 in (4.25). Moreover, C_2 used below in (2.2) is defined in (4.36). The numbers c_1, c_2, c_3, C_2, and C_4 depend on $D, \Lambda, i_0, \rho_{min}, \rho_{max}, \lambda_1, \lambda_2$.

Algorithm NRD (noise reduction in distances). The input for the algorithm is the sample size N and the numbers $\overline{D}_{j,k}$ and $Y_{j,k}$ for $j, k = 1, 2, \dots, N$.

(1) If N satisfies (2.2), continue. Otherwise, stop and report that there are not enough samples.

(2) Set N_0 to be given by (2.1) and set $N_1 = \lfloor N/4 \rfloor$, $N_2 = N - (N_0 + N_1)$.

(3) For $j = 1$ to N_0 do

For $k = N_0 + 1$ to $N_0 + N_1$ do

$$K_{jk}^L = \frac{1}{N_2} \sum_{\ell=N_1+1}^{N_1+N_2} (\min(|\overline{D}_{j,\ell} - \overline{D}_{k,\ell}|^2, L) - 2\sigma^2) Y_{j\ell} Y_{k\ell}, \quad T_{jk} = \sum_{\ell=N_1+1}^{N_1+N_2} Y_{j\ell} Y_{k\ell}.$$

end

end

(4) For $j = 1$ to N_0 do

For $j' = 1$ to N_0 do

$$V_{j,j'} = \frac{1}{N_1} \sum_{\ell=N_0+1}^{N_0+N_1} \widetilde{\psi}_1\left(\frac{T_{jk}}{bN_2}\right) \psi_1\left(\frac{K_{jk}^L}{\rho^2}\right) Y_{j'k}(\overline{D}_{k,j'} - \mu),$$

$$W_{j,j'} = \frac{1}{N_1} \sum_{\ell=N_0+1}^{N_0+N_1} \widetilde{\psi}_1\left(\frac{T_{jk}}{bN_2}\right) \psi_1\left(\frac{K_{jk}^L}{\rho^2}\right) Y_{j'k},$$

$$Q_{j,j'} = \frac{V_{j,j'}}{W_{j,j'}}.$$

end

end

(5) For $j = 1$ to N_0 do
For $j' = 1$ to N_0 do

$$\widetilde{d}_{j,j'} = \begin{cases} Q_{j,j'} & \text{if } W_{j,j'} > u_2, \\ D & \text{otherwise.} \end{cases}$$

end
end
(6) Return the numbers $\widetilde{d}_{j,j'}$, $j, j' \in \{1, 2, \ldots, N_0\}$.

With the probability $1-\theta$, the output satisfies $|\widetilde{d}_{j,j'} - d_M(X_j, X_{j'})| \le \varepsilon_1$ where the points $\mathcal{X} = \{X_1, X_2, \ldots, X_{N_0}\} \subset M$ are a δ_1-dense subset of M. Thus with the probability $1-\theta$, after applying algorithm NRD one has obtained the distance matrix in a δ_1-dense subset $\mathcal{X}$ of M.

After this one can continue using any of the following ways:

(A) One can apply the ISOMAP algorithm to find an imbedding of the set $\mathcal{X}$ with the distance $\widetilde{d}_{ij}$ into a Euclidean space $\mathbb{R}^m$, $m \ge n$.
(B) One can apply the Diffusion Map algorithm [11, 12, 13] to embed $\mathcal{X}$ with the distance $\widetilde{d}_{ij}$ into some sufficiently large-dimensional Euclidean space $\mathbb{R}^m$ and construct local coordinate functions and the metric tensor on the image using the methods presented in [48].
(C) One can use the results presented in [24] (see the procedure ManifoldConstruction) to obtain a constructive procedure that gives a manifold (M^*, g^*) that has the properties stated in Theorem 1.

In case (A), the algorithm is computationally efficient, but the conditions when the result is guaranteed to be close to original manifold are quite restrictive (see [5, 18, 20]).

In case (B), the construction produces with ideal data (that correspond to the continuous case) the original manifold (M, g), the Laplacian Δ_g, and the corresponding heat kernel $e^{-t\Delta_g}$ (see [13]), as well as the metric g (see [37]). The convergence of the reconstructed graph Laplacians to the Laplace operator of the original manifold as $N \to \infty$ is studied in [4, 13, 33, 55].

In case (C), we note that the construction procedure ManifoldConstruction in [24] can be considered as a sketch of an algorithm, and adding the necessary details to make the algorithm numerically implementable requires more work. However, the proof of [24] gives an estimate for the Lipschitz distance of the manifolds (M, g) and (M, g) in terms of the geometric parameters D, Λ, i_0 and the error bound δ_1, ε_1.

To reconstruct a manifold (M^*, g^*) described in Theorem 1 with the error parameter δ, we apply the Algorithm NRD with parameters $\varepsilon_1 = \delta^{3/2}$ and $\delta_1 = \Lambda^{-2/3}\delta^{1/2}/20$ (see the proof of Theorem 1).

Finally, we consider the computational requirements of the algorithm NRD with the above parameters. The algorithm requires $N_0N_1 + 2N_0^2$ steps that each require sums over at most CN_2 terms that contain evaluations of explicit functions of data. We count these evaluations as one operation. As $N_0, N_1, N_2 \le N$, we see that the algorithm requires $O(N_0N_1N_2) \le O(N^3)$ operations. Note that when $\delta \le \theta$ and all parameters other than δ and $N \sim \delta^{-3n}(\log(1/\delta))^5$

are fixed, we have $N_0 = o(N^{1/5})$, and hence the algorithm requires $O(N_0 N_1 N_2) \leq o(N^{2+1/5})$ operations. For the computational requirements of the procedure ManifoldConstruction in the case (C), see [24].

2.2. Probability that the sample points form a dense net. First we estimate the probability that the set $S_0 = \{X_1, \ldots, X_{N_0}\}$ is a δ_1-net, using standard methods based on the collectors problem.

Lemma 2.1. *There is $C_3 \geq 10$ such that if $\theta \in (0, \frac{1}{2})$, $\delta_1 \in (0, \frac{D}{2})$, and*

$$N_0 \geq C_3 \delta_1^{-n} (\log(\delta_1^{-1}) + \log(\theta^{-1})), \tag{2.5}$$

then the probability that the set $\{X_j : \ j = 1, 2, \ldots, N_0\}$ is a δ_1-net in M is larger than $1 - \frac{1}{2}\theta$.

The proof of Lemma 2.1 is based on well-known results. For the reader's convenience , we give in Appendix B its proof using our notations.

3. The modified L^2-norm of the differences of the distance functions. Let $y, z \in M$ be (deterministic) points on M. Denote

$$\Phi_y(x) = \Phi(x, y), \quad \Phi_y^1(x) = \Phi^1(d_M(x, y)).$$

Definition 3.1. *For $x \in M$, let $r_x \in C(M)$. We define the distance function*

$$r_x(y) = d_M(x, y), \quad y \in M.$$

For $y, z \in M$, let

$$k_\Phi(y, z) = \|(r_y - r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|^2_{L^2(M, d\nu)}, \quad A(y, z) = \int_M \Phi_y(x)\Phi_z(x) d\nu(x). \tag{3.1}$$

The map $R : M \to L^\infty(M)$, given by $R(x) = r_x$, defines an isometric embedding $R : M \to R(M) \subset L^\infty(M)$. Below we will consider this map as a function taking values in the Hilbert space $L^\infty(M)$. The function $A(y, z)$ measures the relative density of the points $x \in M$ for which both distances $d_M(x, y)$ and $d_M(x, z)$ are nonmissing in the data that is given to us. In the next lemmas we analyze these functions. Recall that $r_1 = r_0/2$.

Lemma 3.2. *If $d_M(y, z) \leq r_0$, then*

$$A(y, z) = \int_M \Phi_y(x)\Phi_y(x) d\nu(x) \geq a, \quad \textit{where} \quad a = c_0 \phi_1^2 r_1^n. \tag{3.2}$$

Proof. Let $x, y \in M$ be such that $\ell = d_M(y, z) \leq r_1$. Also, let $[yz] = \gamma_{y,\xi}([0, \ell])$, $\xi \in S_y M$ be a length minimizing geodesic from y to z. Let $q = \gamma_{y,\xi}(\ell/2)$. Using properties of Φ given in (1.21) (see also (1.5)–(1.6)), we see that for all $x \in B_M(q, r_1)$ we have $x \in B_M(y, r_0)$ and $x \in B_M(z, r_0)$, and so $\Phi_y(x)\Phi_z(x) \geq \phi_1^2$. Observe that

$$\nu(B_M(q, r_1)) \geq c_0 r_1^n. \tag{3.3}$$

Hence,

$$A(y, z) = \int_M \Phi_y(x)\Phi_y(x) d\nu(x) \geq \int_{B_M(q, r_1)} \Phi_y(x)\Phi_y(x) d\nu(x) \geq \phi_1^2 c_0 r_1^n. \qquad \blacksquare$$

Let X be a random variable having distribution ν. Also, let $Y_{X,y}$ be a random variable, taking values in $\{0,1\}$, that is 1 with the probability $\Phi(X,y)$, and let $Y_{X,z}$ be a random variable, taking values in $\{0,1\}$, that is 1 with the probability $\Phi(X,z)$. Also, let η, η' have the zero mean and variance σ^2. We assume that all X, η, η' are independent random variables. We also assume that under the condition that X is given, random variables $Y_{X,y}$, $Y_{X,z}$, η, and η' are independent.

Lemma 3.3. *Let $y, z \in M$. We have*

$$\mathbb{E}((|(d_M(y,X)+\eta)-(d_M(z,X)+\eta')|^2 - 2\sigma^2)Y_{X,y}Y_{X,z}) = k_\Phi(y,z). \tag{3.4}$$

Proof. We denote $R_y(X) = d_M(y,X)+\eta$ and $R_z(X) = d_M(z,X)+\eta'$. Then

$$\begin{aligned}\mathbb{E}(|R_z(X)-R_y(X)|^2 Y_{X,y}Y_{X,z}) &= \mathbb{E}_{\eta,\eta'}\int_M |(d_M(y,x)+\eta)-(d_M(z,x)+\eta')|^2 \Phi_y(x)\Phi_z(x)d\nu(x)\\ &= \|(r_y-r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|^2_{L^2(M,d\nu)} + 2\sigma^2 A(y,z).\end{aligned}$$

■

3.1. Deterministic estimates for the rough distance function. In this subsection, we consider the rough distance function $k_\Phi(y,z)$.

In the study of metric spaces, Kuratowski observed that the map $R(x) = r_x$ defines an isometric embedding $R : M \to R(M) \subset C(M)$ of the manifold M into the vector space $C(M)$. When there are no missing data, that is, $\Phi = 1$, the following proposition shows that the map $\overline{R} : M \to L^2(M)$, given by $\overline{R}(x) = r_x$, defines a bi-Lipschitz embedding $\overline{R} : M \to \overline{R}(M) \subset L^2(M)$. Note that here $L^2(M)$ is the space $L^2(M, d\nu)$, where ν is a probability measure on M.

Proposition 3.4. *There is a constant $\kappa_M \in (0,1)$, depending on (M,g), such that*

$$\kappa_M d_M(y,z) \le \|r_y - r_z\|_{L^2(M,d\nu)} \le d_M(y,z). \tag{3.5}$$

Due to this, we call the map $\overline{R} : M \to L^2(M)$ the L^2-Kuratowski embedding. The proof of the Proposition 3.4 is the special case of the Proposition 3.5 when $\Phi = 1$, claims (i)–(ii), given below.

Proposition 3.5. (i) *For all $y, z \in M$, we have the inequality*

$$\|(r_y-r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|_{L^2(M,d\nu)} = k_\Phi(y,z)^{1/2} \le A(y,z)d_M(y,z) \le d_M(y,z). \tag{3.6}$$

(ii) *Let*

$$\widehat{a} = \frac{1}{4}\min(\lambda_2 H r_1, a), \quad \kappa = c_1\lambda_1\lambda_2^{-n/2-2}H^{-(n/2+1)}\phi_1^{n+4}, \tag{3.7}$$

where $c_1 = 8^{-n-3}\rho_{min}^{1/2} r_1^{(n+4)n}(\widehat{c}_0)^{1/2}c_6^{(n+4)/2}D^{-1}$ and $c_6 = \min(\frac{1}{8}i_0^{-n}, \frac{1}{4}c_0)$.

If $y, z \in M$ satisfy $A(y,z) \ge \widehat{a}$, then

$$\|(r_y-r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|_{L^2(M,d\nu)} \ge \kappa d_M(y,z). \tag{3.8}$$

(iii) *For all $y, z \in M$ satisfying $d_M(y,z) \le r_1$, where r_1 is defined in (1.20), we have $A(y,z) \ge a \ge \widehat{a}$, and so the inequality (3.8) is valid.*

By the above proposition, if $A(y,z) \geq \widehat{a}$, then $k_\Phi(y,z)^{1/2}$ approximates $d_M(y,z)$. Also, we have $\widehat{a} \geq c_6 r_1^n \phi_1^2$.

Proof. (i) We have by triangular inequality

$$\|(r_y - r_z)\Phi_y^{\frac{1}{2}}\Phi_z^{\frac{1}{2}}\|_{L^2}^2 = \int_M |d_M(y,\,\cdot\,) - d_M(z,\,\cdot\,)|^2 \Phi_y \Phi_z d\nu \leq |d_M(y,z)|^2 A(y,z).$$

As $A(y,z) \leq 1$, this proves the inequality (3.6).

(ii) To prove the inequality in (3.8), we use the following (well-known) corollary of Toponogov's theorem. Similar kinds of formulas are used in section 4.5 of [7]. However, we present the results in the form needed later and give the proof for the reader's convenience.

Lemma 3.6. *Let M be a Riemannian manifold with sectional curvature bounded below by $-\Lambda^2$. Let $x,y,z \in M$ and $\beta = \angle xyz$ be the angle of the length minimizing curves $[xy]$ and $[yz]$ at y. Assume that $d_M(y,z) \leq \frac{1}{2} d_M(x,y)$ and $d_M(x,y) \leq \frac{2}{3}\min(i_0, \pi/(2\Lambda))$. Then*

$$\Big| d_M(x,z) - (d_M(x,y) - d_M(y,z)\cos\beta) \Big| \leq \frac{d_M(y,z)^2}{d_M(x,y)}. \tag{3.9}$$

Proof. Let $\gamma_{y,\xi}([0,\ell])$ be a distance minimizing geodesic from y to z, where $|\xi| = 1$ and $\ell = d_M(y,z)$. Consider functions

$$F(p) = d_M(x,p), \quad f(s) = F(\gamma_{y,\xi}(s)).$$

Let $\ell_0 = \min(i_0, \pi/(2\Lambda))$. Then observe that $d_M(\gamma_{y,\xi}(s), x) \leq \ell_0$ for all $s \in [0,\ell]$.

The gradient of $F(p)$ at $p \in B(0,\ell_0)$ is equal to the normal vector ν of the sphere $\Sigma = \partial B(x,r)$, where $r = d_M(p,x)$, at the point p and the Hessian of F at p and the shape operator $S(p)$ of the sphere Σ have the relation $\mathrm{Hess}(F)(\xi,\eta) = g(S(p)\xi,\eta)$, $\xi,\eta \in T_pM$, and where $g : T_pM \times T_pM \to \mathbb{R}$ is the quadratic form determined by the metric tensor g; see [50]. By the standard comparison estimates [50, Chap. 6, Thm. 2.1], in the space $T_p\Sigma$ we have

$$\frac{\Lambda\cosh(\Lambda r)}{\sinh(\Lambda r)} \leq S(p) \leq \frac{\Lambda\cos(\Lambda r)}{\sin(\Lambda r)}. \tag{3.10}$$

As $\frac{d}{dt}(\tan(t)) = 1/\cos^2(t)$, the mean value theorem implies that for $0 < s < \pi/(2\Lambda)$ we have $\tan(\Lambda s) \geq \Lambda s$, so that $\|S(p)\| \leq 1/F(p)$.

Since $\partial_s f(s) = g(\nabla F(\gamma_{y,\xi}(s)), \dot\gamma_{y,\xi}(s)) = g(\nu(\gamma_{y,\xi}(s)), \dot\gamma_{y,\xi}(s))$, then $\nu(x) = \nabla F(x)$ is the normal of the sphere $\partial B(y,s)$ at the point $x = \gamma_{y,\xi}(s)$. Moreover, $f(0) = d_M(x,y)$ and $\partial_s f(0) = g(-\dot\gamma_{y,\xi}(s), \dot\gamma_{y,\xi}(s)) = -\cos\beta$. Also, since

$$\partial_s^2 f(s) = (\mathrm{Hess}\, F)(\dot\gamma(s),\dot\gamma(s)) + g(\nabla F(\gamma(s)), \nabla_{\gamma(s)}\dot\gamma(s)) = g(S(\gamma(s))\dot\gamma(s),\dot\gamma(s)),$$

where $\gamma = \gamma_{y,\xi}$, we have

$$|\partial_s^2 f(s)| \leq \frac{1}{f(s)} \leq \frac{1}{d_M(x,y) - d_M(y,z)}.$$

Hence, using Taylor's series we see that

$$\left| f(s) - (d_M(x,y) - s\cos\beta) \right| \le \frac{s^2}{2(d_M(x,y) - d_M(y,z))} \le \frac{s^2}{d_M(x,y)}.$$

This proves the claim. ∎

Next we continue the proof of inequality (3.8). We consider the claim in two cases.

Case 1. Assume that $d_M(y,z) \ge r_1/16$. We show that $\kappa' > 0$ such that

$$\|r_y - r_z\|_{L^2(M,d\nu)} \ge \|(r_y - r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|_{L^2(M,d\nu)} \ge \kappa' d_M(y,z). \tag{3.11}$$

Proof of Case 1. Assume that $0 < d_M(y,z) < r_1/16$. Then, if $x \in M$ is such that $r_1/2 < d_M(x,z) < r_1$, we have $d_M(x,y) \ge r_1/4$, and by (3.9),

$$d_M(x,z) \le d_M(x,y) - d_M(y,z)\cos\beta + \frac{1}{4} d_M(y,z),$$

where β is the angle $\angle xyz$. When $\beta < \pi/4$, this yields

$$d_M(x,y) - d_M(x,z) \ge d_M(y,z)\cos\beta - \frac{1}{4} d_M(y,z) \ge \frac{1}{4} d_M(y,z).$$

Thus, let

$$W = \{x \in M;\ r_1/2 < d_M(x,z) < r_1,\ \angle xyz < \pi/4\}.$$

Then

$$\|(r_y - r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|^2_{L^2(M,d\nu)} \ge \frac{1}{16} d(y,z)^2 \phi_1^2 \operatorname{vol}_\nu(W),$$

where by [50, Chap. 6.2, Cor. 2.4] (see also (1.19), (1.20))

$$\begin{aligned}\operatorname{vol}_\nu(W) &\ge \rho_{min}\operatorname{vol}_M(W) \ge \rho_{min} 4^{-n}\omega_n \left(\frac{\sin(\Lambda r_1)}{\Lambda r_1}\right)^{n-1} (r_1^n - (r_1/2)^n) \\ &\ge \rho_{min} 4^{-n}\omega_n \left(2^{-1/2}\right)^{n-1} (r_1^n - (r_1/2)^n) \ge 2^{-4n}\rho_{min}\omega_n r_1^n.\end{aligned}$$

Thus there exists $\kappa' = \phi_1\, 2^{-2n-2}\rho_{min}^{1/2}\omega_n^{1/2} r_1^{n/2}$ such that (3.11) is valid.

Case 2. Assume that $d_M(y,z) \ge r_1/16$. Then we show that κ'' such that

$$\|(r_y - r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|_{L^2(M,d\nu)} = k_\Phi(y,z)^{1/2} \ge \kappa'' d_M(y,z). \tag{3.12}$$

Proof of Case 2. The assumption $d_M(y,z) \ge r_1/16$ and the definition of $\widehat{a}$ imply that $4\,\widehat{a}\,\lambda_2^{-1}H^{-1} \le r_1$. Then $d_M(y,z) \ge \frac{1}{4}\widehat{a}\lambda_2^{-1}H^{-1}$. Denote $d_M(y,z) = \ell$. Then $\ell \ge 2r_a$, where

$$r_a = \frac{1}{8}\widehat{a}\lambda_2^{-1}H^{-1} \le r_1/16.$$

Assume that $A(y,z) \geq \widehat{a}$. Since $d_M(y,x) + d_M(x,z) \geq d_M(y,z) = \ell$, we see that for all $x \in M$ we have either $d_M(x,y) \geq \ell/2$ or $d_M(x,z) \geq \ell/2$. Let us assume that the latter is true. As $s \mapsto \Phi^1(s)$ is nonincreasing and Φ takes values in $[0,1]$, we have $\Phi_y(x)\Phi_z(x) \leq \Phi_z(x) \leq \lambda_2\Phi^1(d_M(x,z)) \leq \lambda_2\Phi^1(\ell/2)$. This yields that

$$\widehat{a} \leq A(y,z) = \int_M \Phi_y(x)\Phi_z(x)d\nu(x) \leq \lambda_2\Phi^1(\ell/2), \tag{3.13}$$

so that $\Phi^1(\ell/2) \geq \widehat{a}/\lambda_2$.

Let $[yz] = \gamma_{y,\xi}([0,\ell])$, $\xi \in S_yM$, be a length minimizing geodesic from y to z. Let $q = \gamma_{y,\xi}(\ell/2)$, $p = \gamma_{y,\xi}(\ell/2 - r_a)$, and $r = r_a/2$. When $d_M(x,p) < r$, we have

$$d_M(y,x) \leq d_M(y,p) + d_M(p,x) \leq \ell - \frac{r_a}{2}, \quad d_M(z,x) \geq d_M(z,p) - d_M(p,x) \geq \ell + \frac{r_a}{2},$$

so that $d_M(z,x) - d_M(y,x) \geq r_a$. Recall that $\Phi(x,y) \geq \lambda_1\Phi^1(d_M(x,y))$, $\Phi^1 : [0,\infty) \to [0,1]$ is nonincreasing and $\|\Phi^1\|_{C^1(\mathbb{R})} \leq H$. As $\Phi^1_y(q) = \Phi^1_z(q) = \Phi^1(\ell/2) \geq \widehat{a}/\lambda_2$ and $B_M(p,r) \subset B_M(q, \frac{3}{2}r_a)$, we have for all $x \in B_M(p,r)$

$$\begin{aligned}\Phi_y(x)\Phi_z(x) &\geq \lambda_1^2(\Phi^1_y(p) - H(r_a + r))\,(\Phi^1_z(p) - H(r_a + r)) \\ &\geq \lambda_1^2\left(\lambda_2^{-1}\widehat{a} - \frac{3}{2}Hr_a\right)^2 \geq \frac{1}{4}\lambda_1^2\lambda_2^{-2}\widehat{a}^{\,2}.\end{aligned}$$

Then, as $\ell/D \leq 1$,

$$\begin{aligned}\|(r_y - r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|^2_{L^2(M,d\nu)} &\geq \int_{B_M(p,r)} |d_M(y,x) - d_M(z,x)|^2\Phi_y(x)\Phi_z(x)\,d\nu(x) \\ &\geq \left(\frac{1}{8}\widehat{a}\lambda_2^{-1}H^{-1}\right)^2 \cdot \frac{1}{4}\lambda_1^2\lambda_2^{-2}\widehat{a}^{\,2} \cdot \rho_{min}\widehat{c}_0\left(\frac{1}{2}\frac{1}{8H}\lambda_2^{-1}\widehat{a}\right)^n \\ &\geq \rho_{min}\frac{1}{4}\lambda_1^2\lambda_2^{-2}\widehat{a}^{\,2} \cdot \widehat{c}_0\left(\frac{1}{16H}\lambda_2^{-1}\widehat{a}\right)^{n+2} \\ &\geq (\kappa'')^2\ell^2 = (\kappa'')^2 d_M(y,z)^2,\end{aligned}$$

where

$$\kappa'' = c_1\lambda_1\lambda_2^{-n/2-2}H^{-(n/2+1)}\phi_1^{n+4} \leq (16^{-n-3}\rho_{min}\lambda_1^2\lambda_2^{-n-4}\widehat{c}_0H^{-(n+2)}\widehat{a}^{\,n+4}D^{-2})^{1/2}$$

and $c_1 = 8^{-n-3}\rho_{min}^{1/2}r_1^{(n+4)n}(\widehat{c}_0)^{1/2}c_6^{(n+4)/2}D^{-1}$. As we have assumed that $D, \Lambda \geq 1$ and we have $\phi_1 \leq 1$, we see that $\kappa'' \leq \kappa'$. This yields (ii) with $\kappa = \kappa''$.

(iii) As $\lambda_1 \leq 1$, Lemma 3.2 implies that if $d_M(y,z) \leq r_1$, then $A(y,z) \geq a \geq \widehat{a}$. ∎

3.2. Probabilistic estimates for the rough distance function. In this subsection, we consider the rough distance function $k_\Phi(X_j, X_k)^{1/2}$.

Next we determine approximately $k_\Phi(X_j,X_k) = \|(r_{X_j} - r_{X_k})\Phi^{1/2}_{X_j}\Phi^{1/2}_{X_k}\|^2_{L^2(M,d\nu)}$ for $(j,k) \in \mathcal{I}^{(0)} \times \mathcal{I}^{(1)}$ by averaging noisy distances over the dense net $S_2 = \{X_j : j \in \mathcal{I}^{(2)}\}$.

For $(j,k) \in I^{(0)} \times I^{(1)}$, we consider the random variables

$$K_{jk} = K^{\sigma}_{jk}, \quad K^{\sigma}_{jk} = \sum_{\ell \in I^{(2)}} \frac{1}{N_2}(|\overline{D}_{j,\ell} - \overline{D}_{k,\ell}|^2 - 2\sigma^2) Y_{j\ell} Y_{k\ell}, \tag{3.14}$$

$$K^L_{jk} = K^{L,\sigma}_{jk}, \quad K^{L,\sigma}_{jk} = \sum_{\ell \in I^{(2)}} \frac{1}{N_2}(\min(|\overline{D}_{j,\ell} - \overline{D}_{k,\ell}|^2, L) - 2\sigma^2) Y_{j\ell} Y_{k\ell}. \tag{3.15}$$

We also consider the random variables

$$A_{jk} = A(X_j, X_k) = \int_M \Phi_{X_j}(x)\Phi_{X_k}(x) d\nu(x), \quad T_{jk} = \sum_{\ell \in I^{(2)}} Y_{j\ell} Y_{k\ell}. \tag{3.16}$$

Roughly speaking, below T_{jk} measures how well K^L_{jk} approximates $k_\Phi(X_j, X_k)$ that further approximates $d_M(X_j, X_k)^2$. We note that if the variance σ^2 is not a priori known, it can be estimated (using the data $\overline{D}_{jk}$) with an error that is small enough for our construction with the probability $1 - \theta$; see Remark 4.10.

3.2.1. Probabilistic notations. To introduce some notation, let us assume for simplicity that the complete probability space $(\Omega, \Sigma, \mathbb{P})$ can be represented as a product $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3$ such that $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2 \times \mathbb{P}_3$ and $\omega = (\omega_1, \omega_2, \omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$. We assume that $\eta = \eta(\omega_1)$, $\eta = \eta'(\omega_2)$ are random variables with variance σ, and $X = X(\omega_3)$. We assume that X is a random variable having distribution ν. Also, assume that X, η, and η' are independent.

For an integrable function $F(\omega_1, \omega_2, \omega_3) = f(\eta(\omega_1), \eta'(\omega_2), X(\omega_3))$ we denote

$$\mathbb{E}_{\eta,\eta'} F = \mathbb{E}_{\eta,\eta'} f(\eta, \eta', X) = \int_{\Omega_1} \int_{\Omega_2} F(\omega_1, \omega_2, \omega_3) d\mathbb{P}_1(\omega_1) d\mathbb{P}_2(\omega_2). \tag{3.17}$$

As X is a random variable, we have that also $\mathbb{E}_{\eta,\eta'} f(\eta, \eta', X)$ is a random variable. The expectation $\mathbb{E}_{\eta,\eta'} f(\eta, \eta', X)$ over variables η, η' is a function of X, and thus it can be considered as the expectation of $f(\eta, \eta', X)$ under the condition that X is known.

Below, we consider conditional expectations using σ-algebras. Let $\mathcal{B}_X \subset \Sigma$ be σ-algebra generated by the random variable $X : \Omega \to \mathbb{R}$, that is, the σ-algebra generated by sets $X^{-1}(S) \subset \Omega$, where $S \subset \mathbb{R}$ is an open set (see [39, Chap. 5]). We recall that $\mathbb{E}(F|\mathcal{B}_X)(\omega)$ is the $\mathcal{B}_X$-measurable random variable that satisfies

$$\int_S \mathbb{E}(F|\mathcal{B}_X)(\omega)\, d\mathbb{P}(\omega) = \int_S F(\omega)\, d\mathbb{P}(\omega)$$

for all sets $S \in \mathcal{B}_X$. In formula (3.17), $\mathbb{E}_{\eta,\eta'} F$ is in fact equal to the conditional expectation $\mathbb{E}(F|\mathcal{B}_X) = \mathbb{E}(F|\mathcal{B}_X)(\omega)$ of the random variable F with respect to the σ-algebra $\mathcal{B}_X \subset \Sigma$; that is, we have $(\mathbb{E}_{\eta,\eta'} F)(\omega_3) = \mathbb{E}(F|\mathcal{B}_X)(\omega)$ with $\omega = (\omega_1, \omega_2, \omega_3)$. As X is a random variable, $\mathbb{E}(Z\,|\mathcal{B}_X)$ is a random variable, too. Recall also the notation $\mathbb{P}(A\,|\mathcal{B}_X) = \mathbb{E}(1_A\,|\mathcal{B}_X)$ for an event $A \in \Sigma$, where $1_A(\omega)$ is the indicator function of the set $A \subset \Omega$. Below we state several times the fact that

$$\mathbb{E}(\mathbb{E}(Z\,|\mathcal{B}_X)) = \mathbb{E}(Z), \quad \mathbb{E}(\mathbb{P}(A\,|\mathcal{B}_X)) = \mathbb{P}(A). \tag{3.18}$$

Below, we will consider the σ-algebra $\mathcal{B}_j \subset \Sigma$ generated by the random variable $X_j : \Omega \to \mathbb{R}$. We also consider the σ-algebra $\mathcal{B}_{jk}$ generated by the random variables X_j and X_k.

By Lemma 3.3, the conditional expectation of K_{jk}, under the condition that X_j and X_k are known, satisfies $\mathbb{E}\big(K_{jk}\,\big|\mathcal{B}_{jk}\big) = k_\Phi(X_j, X_k)$, where $k_\Phi(X_j, X_k) = \|(r_{X_j} - r_{X_k})\Phi_{X_j}^{1/2}\Phi_{X_k}^{1/2}\|^2_{L^2(M,d\nu)}$ is a random variable.

3.2.2. Probabilistic estimates for rough distances K_{jk}^L and reliability values T_{jk}. We use the following form of Hoeffding's inequality.

Lemma 3.7 (Hoeffding's inequality [35]). *Let $Z_1, \dots, Z_N$ be N i.i.d. copies of the random variable Z satisfying $0 \le Z \le L$, where $L > 0$. Then, for $\varepsilon > 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{N}\left(\sum_{i=1}^{N} Z_i\right) - \mathbb{E}[Z]\right| \le \varepsilon\right] \ge 1 - 2\exp(-2N\varepsilon^2 L^{-2}).$$

Below, we will show that K_{jk}^L, defined in (3.15), can be considered as an approximation of $k_\Phi(X_j, X_k)$, which further approximates $d_M(X_j, X_k)^2$, when $A(X_j, X_k)$ is larger than a suitable threshold value. Let

$$\varepsilon_3 \le \frac{1}{4}a, \quad b = \frac{1}{2}a = c_2\phi_1^2, \quad c_2 = c_0 r_1^n. \tag{3.19}$$

For $j \in \mathcal{I}^{(0)}$, we consider the events $\mathcal{E}_j^{(1)} \subset \Omega$ and $\mathcal{E}^{(1)} \subset \Omega$, defined by

$$\mathcal{E}_j^{(1)} = \left\{\omega \in \Omega \,\middle|\, \forall k \in I^{(1)} \left(\left|\frac{T_{jk}}{N_2} - A_{jk}\right| \le \varepsilon_3\right)\right\}, \quad \mathcal{E}^{(1)} = \bigcap_{j \in I^{(0)}} \mathcal{E}_j^{(1)}. \tag{3.20}$$

Below, we use smooth cut-off functions $\psi_\rho(t) = \psi_1(t/\rho^2)$ and $\widetilde{\psi}_1(t) = 1 - \psi_1(t)$, where $\psi_1 \in C_0^\infty(\mathbb{R})$ satisfies

$$\begin{aligned} &\text{supp }(\psi_1) \subset (-2, 2), \quad \|\psi_1\|_{C^1(\mathbb{R})} \le 2, \\ &\psi_1(t) = 1 \quad \text{for } -1 \le t \le 1, \\ &0 \le \psi_1(t) \le 1, \quad \psi_1(-t) = \psi_1(t), \quad \text{and } \psi_1|_{\mathbb{R}_+} \text{ is nonincreasing.} \end{aligned} \tag{3.21}$$

Note that if $\widetilde{\psi}_1(T_{jk}(bN_2)^{-1}) > 0$, then $\frac{T_{jk}}{bN_2} \ge 1$. Also, if $\mathcal{E}_j^{(1)}$ happens and $A_{jk} \ge a$, then

$$\frac{T_{jk}}{N_2} \ge A_{jk} - \varepsilon_3 \ge a - \frac{1}{4}a \ge b. \tag{3.22}$$

Moreover, if $\frac{1}{b}\frac{T_{jk}}{N_2} \ge 1$, then (3.22) implies

$$A_{jk} \ge b - \varepsilon_3 \ge \frac{1}{4}a \ge \widehat{a}. \tag{3.23}$$

For $y, z \in M$, let $Y(z, y)$ be a random variable that is 1 with the probability $\Phi(y, z)$, and 0 with the probability $1 - \Phi(y, z)$. Suppose the random variables $Y(y, z)$ are independent for $y, z \in M$. Let

$$T(y, z) = \sum_{\ell \in I^{(2)}} Y(y, X_\ell) Y(z, X_\ell). \tag{3.24}$$

As $\mathbb{E}\frac{1}{N_2}T(y,z) = A(y,z)$, Hoeffding's inequality implies

$$\mathbb{P}\left[\left|\frac{T(y,z)}{N_2} - A(y,z)\right| \le \varepsilon_3\right] \ge 1 - 2\exp(-2N_2\varepsilon_3^2). \tag{3.25}$$

As T_{jk} and $T(X_j, X_k)$ have the same distributions, and $A_{jk} = A(X_j, X_k)$ for $j \in \mathcal{I}^{(0)}$ and $k \in \mathcal{I}^{(1)}$, inequality (3.25) implies for the conditional probability, under the condition that X_j and X_k are known, that

$$\mathbb{P}\left[\left|\frac{T_{jk}}{N_2} - A_{jk}\right| \le \varepsilon_3 \,\middle|\, \mathcal{B}_{jk}\right] \ge 1 - 2\exp(-2N_2\varepsilon_3^2). \tag{3.26}$$

Thus we have by (3.18) $\mathbb{P}(\mathcal{E}_j^{(1)}) \ge 1 - 2N_1\exp(-2N_2\varepsilon_3^2)$. Hence,

$$\mathbb{P}(\mathcal{E}^{(1)}) \ge 1 - p^{(1)}, \quad p^{(1)} = 2N_0N_1\exp(-2N_2\varepsilon_3^2). \tag{3.27}$$

We recall that by Lemma 3.2 and Proposition 3.5,

$$\left(d_M(y,z) \le r_1 \implies A(y,z) \ge a \ge \widehat{a}\right) \quad \text{and} \tag{3.28}$$

$$\left(A(y,z) \ge \widehat{a} \implies d_M(y,z) \ge \|(r_y - r_z)\Phi_y^{1/2}\Phi_z^{1/2}\|_{L^2(M,d\nu)} = k_\Phi(y,z)^{\frac{1}{2}} \ge \kappa d_M(y,z)\right).$$

Lemma 3.8. *Let*

$$L > 2\max(D^2, \sigma), \quad \varepsilon_2 > 0, \quad \varepsilon(L) := \beta e^{-(L^{1/2}-D)/2}(D^2 + 6\beta^2), \tag{3.29}$$

and consider the events $\mathcal{E}_j^{(2)} \subset \Omega$, $j \in I^{(0)}$, *and* $\mathcal{E}^{(2)} \subset \Omega$,

$$\mathcal{E}_j^{(2)} = \{\forall k \in I^{(1)} : \ |K_{jk}^L - k_\Phi(X_j, X_k)| \le \varepsilon_2 + \varepsilon(L)\}, \quad \mathcal{E}^{(2)} = \bigcap_{j \in I^{(0)}} \mathcal{E}_j^{(2)}. \tag{3.30}$$

Then

$$\mathbb{P}(\mathcal{E}^{(2)}) \ge 1 - p^{(2)}, \quad p^{(2)} = 2N_0N_1\exp(-2N_2\varepsilon_2^2L^{-2}). \tag{3.31}$$

Proof. Denote $\eta_{jk\ell} = \eta_{j\ell} - \eta_{k\ell}$ and $D_{jk\ell} = d_M(X_j, X_\ell) - d_M(X_k, X_\ell)$. Then

$$\mathbb{E}\eta_{jk\ell} = 0, \quad \mathbb{E}\eta_{jk\ell}^2 = 2\sigma^2, \quad \mathbb{E}e^{|\eta_{jk\ell}|} \le (\mathbb{E}e^{|\eta_{j\ell}-\mu|})(\mathbb{E}e^{|\eta_{k\ell}-\mu|}) \le \beta^2. \tag{3.32}$$

Let $r = L^{1/2}$. Observe that $|D_{jk\ell}| \le D$, so that if $|D_{jk\ell} + \eta_{jk\ell}| \ge r$, then $|\eta_{jk\ell}| > r - D$. By (3.32), $Y = e^{|\eta_{jk\ell}|}$ satisfies

$$\mathbb{P}(|\eta_{jk\ell}| > r - D) \le \frac{\mathbb{E}(e^{|\eta_{jk\ell}|})}{e^{r-D}} = \beta^2 e^{-(r-D)}.$$

Thus, using the fact that $\eta_{jk\ell}$ and $D_{jk\ell}$ are independent, we see, using the Schwartz inequality, that

$$\begin{aligned}&\mathbb{E}\left(|(\min((D_{jk\ell}+\eta_{jk\ell})^2,L)-(D_{jk\ell}+\eta_{jk\ell})^2)Y_{j\ell}Y_{k\ell}|\ \middle|\ \mathcal{B}_{jk}\right)\\&\le \mathbb{E}(\chi_{|\eta_{jk\ell}|>r-D}(D_{jk\ell}+\eta_{jk\ell})^2|\ \mathcal{B}_{jk}) \le (\mathbb{P}(|\eta_{jk\ell}|>r-D))^{\frac12}\,(\mathbb{E}((D_{jk\ell}+\eta_{jk\ell})^4|\ \mathcal{B}_{jk}))^{\frac12}\\&\le \beta e^{-(r-D)/2}(D^2+6\beta^2)\le \varepsilon(L).\end{aligned}$$

By Lemma 3.3, the above shows that

$$k_\Phi^L(X_j,X_k):=\mathbb{E}((\min((D_{jk\ell}+\eta_{jk\ell})^2,L)-2\sigma^2)Y_{j\ell}Y_{k\ell}|\ \mathcal{B}_{jk})$$

satisfies

$$\left|k_\Phi^L(X_j,X_k)-k_\Phi(X_j,X_k)\right|\le\varepsilon(L). \tag{3.33}$$

By arguing as in (3.25)–(3.26), we see that Hoeffding's inequality implies

$$\mathbb{P}\left[|K_{jk}^L-k_\Phi^L(X_j,X_k)|\le\varepsilon_2\middle|\mathcal{B}_{jk}\right]\ge 1-2\exp(-2N_2\varepsilon_2^2L^{-2}). \tag{3.34}$$

Using this, (3.18), and (3.33), and summing over $j\in I^{(0)}$, we obtain (3.31). ∎

4. Determination of the approximate distances in the sparse net. Next we assume that

$$\rho\le r_1=r_0/2\le\phi_0/(4H),\quad u_0=4c_3\phi_0\rho^n,\quad u_1=u_0/2,\quad u_2=u_0/4, \tag{4.1}$$

where $c_3=2^{-3-2n}c_0$.

Next we define $Q_{j,j'}$, which will turn to the approximate distances $d_M(X_j,X_{j'})$ for points X_j and $X_{j'}$, where $j,j'\in I^{(0)}$, which are sufficiently close to each other.

Definition 4.1. *Let $\rho\in(0,1)$ satisfy (4.1). For $j,j'\in I^{(0)}$, let*

$$Q_{j,j'}=\frac{V_{j,j'}}{W_{j,j'}},\quad \textit{where} \tag{4.2}$$

$$V_{j,j'}=\frac{1}{N_1}\sum_{k\in I^{(1)}}\widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_\rho(K_{jk}^L)Y_{j'k}(\overline{D}_{k,j'}-\mu), \tag{4.3}$$

$$W_{j,j'}=\frac{1}{N_1}\sum_{k\in I^{(1)}}\widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_\rho(K_{jk}^L)Y_{j'k}. \tag{4.4}$$

In the case when $W_{j,j'}$ is zero, we define $Q_{j,j'}$ to be D. To emphasize that μ and σ appear in the definitions of $Q_{j,j'}$ and $V_{j,j'}$, we sometimes denote $Q_{j,j'}=Q_{j,j'}^{(\mu,\sigma)}$ and $V_{j,j'}=V_{j,j'}^{(\mu,\sigma)}$. We also use notation $W_{j,j'}=W_{j,j'}^{(\sigma)}$.

We define for $j,j'\in I^{(0)}$

$$d^{app}(X_j,X_{j'})=\begin{cases}Q_{j,j'} & \text{if } W_{j,j'}>u_2,\\ D & \text{otherwise.}\end{cases} \tag{4.5}$$

Roughly speaking, above the function $\widetilde{\psi}_1(T_{jk}(bN_2)^{-1})$ measures the reliability of the terms $\psi_\rho(K^L_{jk})$ in formulas (4.3) and (4.4), and $W_{j,j'}$ measures the reliability of $Q_{j,j'}$ in formula (4.5). The numbers $d^{app}(X_j, X_{j'})$ will be the final approximation for the distances $d_M(X_j, X_{j'})$ for all pairs $(X_j, X_{j'})$ of points that are close to each other. Observe that $Q_{j,j'}$ and $W_{j,j'}$ can be computed from the given data.

For technical purposes, we define deterministic (indexed with (d)) and random (indexed with (r)) functions

$$W^{(d),-}(y,z) = \int_M \widetilde{\psi}_1(A(y,x)b^{-1})\Phi(z,x)\,\psi_{\rho/2}(k_\Phi(y,x))d\nu(x), \tag{4.6}$$

$$W^{(r),-}(y,z) = \frac{1}{N_1}\sum_{k\in I^{(1)}} \widetilde{\psi}_1(A(y,X_k)b^{-1})\Phi(z,X_k)\,\psi_{\rho/2}(k_\Phi(y,X_k)). \tag{4.7}$$

The motivation behind defining functions $W^{(d),-}$ and $W^{(r),-}$ is that we can use Hoeffding's inequality to estimate how close $W^{(d),-}(X_j, X_{j'})$ and $W^{(r),-}(X_j, X_{j'})$ are when X_j and $X_{j'}$ are known. Also, we show that we have $W_{j,j'} \geq W^{(r),-}(X_j, X_{j'})$ with a large probability.

Lemma 4.2. *If $d_M(x,z) < r_1$, then*

$$W^{(d),-}(y,z) \geq u_0. \tag{4.8}$$

Moreover, when we have $\Phi(x,y) \geq \lambda_1\phi_0$ for all $x, y \in M$, the inequality (4.8) holds for all $x, y \in M$.

Proof. Let

$$w^-(x,y,z) = \widetilde{\psi}_1(A(y,x)b^{-1})\Phi(z,x)\,\psi_{\rho/2}(k_\Phi(y,x))$$

and recall that by (4.1), $\rho \leq r_1$.

When $d_M(x,y) < \rho/4$, by Lemma 3.5 (iii) we have that $A(x,y) \geq a$. Also, in the case when $\Phi(x,y) \geq \lambda_1\phi_0 \geq \phi_1$ for all $(x,y) \in M \times M$, we have $A(x,y) \geq a$. Thus in both cases, $\widetilde{\psi}_1(A(y,x)b^{-1}) = 1$. Moreover, by 3.5 (i), we have $k_\Phi(y,x) \leq (d_M(y,x))^2 < (\rho/2)^2$, so that $\psi_{\rho/2}(k_\Phi(y,x)) = 1$.

Also, by (1.21), when $d_M(x,z) < r_1 = r_0/2$, we have $\Phi(z,x) \geq \phi_1$.

Thus, when $d_M(x,z) < r_1$ or $\Phi(x,y) \geq \lambda_1\phi_0 \geq \phi_1$, and when we have $d_M(x,y) < \rho/4$, it holds that $w_-(x,y,z) \geq \phi_1$. Hence we see that

$$W^{(d),-}(y,z) = \int_M w_-(x,y,z)d\nu(x) \geq \phi_1 \cdot \nu\Big(B_M\Big(y, \frac{\rho}{4}\Big)\Big) \geq \phi_1 \cdot c_0(\rho/4)^n = u_0. \qquad \blacksquare$$

Let us write for $j, j' \in I^{(0)}$ the following:

$$Q_{j,j'} = Q^1_{j,j'} + Q^2_{j,j'}, \text{ where } Q^1_{j,j'} = \frac{V^1_{j,j'}}{W_{j,j'}}, \quad Q^2_{j,j'} = \frac{V^2_{j,j'}}{W_{j,j'}}, \tag{4.9}$$

$$V^1_{j,j'} = \frac{1}{N_1}\sum_{k\in I^{(1)}} \widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_\rho(K^L_{jk})Y_{j'k}d_M(X_k, X_{j'}),$$

$$V^2_{j,j'} = \frac{1}{N_1}\sum_{k\in I^{(1)}} \widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_\rho(K^L_{jk})Y_{j'k}(\eta_{k,j'} - \mu),$$

and consider the terms $Q^1_{j,j'}$ and $Q^2_{j,j'}$ separately.

First we will show that $Q^2_{j,j'}$ is small with a large probability when $W_{j,j'} \geq u_2$. To that end, let $h_0 > 0$ and $\mathcal{E}^{(3)}_{j,j'} \subset \Omega$, $(j,j') \in I^{(0)} \times I^{(0)}$, and $\mathcal{E}^{(3)} \subset \Omega$ be the events

$$\mathcal{E}^{(3)}_{j,j'} = \{(W_{j,j'} \geq u_2) \implies (|Q^2_{j,j'}| \leq h_0)\}, \quad \mathcal{E}^{(3)} = \bigcap_{(j,j') \in I^{(0)} \times I^{(0)}} \mathcal{E}^{(3)}_{j,j'}. \tag{4.10}$$

Lemma 4.3. *For any $h_0 \in (0,1)$ we have*

$$\mathbb{P}(\mathcal{E}^{(3)}) \leq 1 - p^{(3)}, \quad p^{(3)} = 2N_0^2 \exp(-e^{-2\beta} N_1 u_2 h_0^2/4). \tag{4.11}$$

Proof. Let us next recall some basic facts: Let $a = (a_k)_{k=1}^{N_1}$ satisfy $0 \leq a_k \leq 1$ and

$$S_{N_1} = \Big(\sum_{k=1}^{N_1} a_k\Big)^2 \Big/ \Big(\sum_{k=1}^{N_1} a_k^2\Big), \quad Z_{N_1} = \sum_{k=1}^{N_1} a_k, \quad V_{N_1} = \frac{1}{Z_{N_1}} \sum_{k=1}^{N_1} a_k(\eta_k - \mu), \tag{4.12}$$

where η_k are i.i.d. variables $\mathbb{E}(\eta_k - \mu) = 0$ and $\mathbb{E}e^{|\eta_k - \mu|} \leq \beta$. Since $0 \leq a_k \leq 1$, we have $a_k^2 \leq a_k$, so that $\sum_{k=1}^{N_1} a_k^2 \leq \sum_{k=1}^{N_1} a_k$, and

$$S_{N_1} = \Big(\sum_{k=1}^{N_1} a_k\Big)^2 \Big/ \Big(\sum_{k=1}^{N_1} a_k^2\Big) \geq \Big(\sum_{k=1}^{N_1} a_k\Big)^2 \Big/ \Big(\sum_{k=1}^{N_1} a_k\Big) \geq \sum_{k=1}^{N_1} a_k = Z_{N_1}. \tag{4.13}$$

Then using Jensen's inequality for the random variable $R = e^{\eta_k - \mu}$ with concave function $h : [0,\infty) \to \mathbb{R}$, $h(s) = s^t$ with $t \in [0,1]$ (for which the standard Jensen's inequality reverses), we obtain $\mathbb{E}(R^t) \leq (\mathbb{E}R)^t$. This yields

$$\mathbb{E}(\exp(t(\eta_k - \mu))) \leq \beta^t \leq e^{\beta t}$$

for $t \in [0,1]$. Hence, as $\mathbb{E}(\eta_k - \mu) = 0$, the moment generating function of $(\eta_k - \mu)$, $M(t) = \mathbb{E}(\exp(t(\eta_k - \mu)))$ satisfies, by considerations involving Taylor series and the mean value theorem, $|M_k(t) - 1| \leq c' t^2$, $t \in [0,1]$, where $c' \leq e^{2\beta}$. Then, by using the independency of random variables η_k, we see that the moment generating function of V_{N_1} satisfies for $s \in [0, Z_{N_1}]$

$$\mathbb{E}\exp(sV_{N_1}) = \prod_{k=1}^{N_1} M\Big(s\frac{a_k}{Z_{N_1}}\Big) \leq \exp\Big(c' \sum_{k=1}^{N_1} \Big(s\frac{a_k}{Z_{N_1}}\Big)^2\Big) \leq \exp\Big(e^{2\beta}\frac{s^2}{S_{N_1}}\Big) \leq \exp\Big(e^{2\beta}\frac{s^2}{Z_{N_1}}\Big),$$

where we use the fact that $\sum_{k=1}^{N_1}(a_k/Z_{N_1})^2 = 1/S_{N_1}$. Similarly, by considering $-(\eta_k - \mu)$ instead of $\eta_k - \mu$, we see that $\mathbb{E}\exp(-sV_{N_1}) \leq \exp(e^{2\beta}\frac{s^2}{S_{N_1}}) \leq \exp(e^{2\beta}\frac{s^2}{Z_{N_1}})$.

Then we see that for $0 < \lambda \leq 1$ and $s \in [0, Z_{N_1}]$,

$$\mathbb{P}(|V_{N_1}| > \lambda) \leq \mathbb{P}(sV_{N_1} > s\lambda) + \mathbb{P}(sV_{N_1} < -s\lambda) \leq 2\exp(e^{2\beta}s^2 Z_{N_1}^{-1} - s\lambda).$$

When $s = \lambda Z_{N_1} e^{-2\beta}/2$, the above implies

$$\begin{aligned} \mathbb{P}(|V_{N_1}| > \lambda) &\le 2\exp(e^{2\beta}\cdot 2^{-2}\lambda^2 Z_{N_1}^2 e^{-4\beta}\cdot Z_{N_1}^{-1} - 2^{-1}\lambda^2 Z_{N_1} e^{-2\beta}) \\ &\le 2\exp(-\lambda^2 Z_{N_1} e^{-2\beta}/4). \end{aligned} \tag{4.14}$$

Next we consider a fixed $j, j' \in I^{(0)}$, and let

$$a_k = \widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_\rho(K_{jk}^L)Y_{j'k},$$

for $k \in I^{(1)}$, $\lambda = h_0$, and let $Z_{N_1} = N_1 W_{j,j'}$ be defined analogously to (4.12). Then we see that $Q_{j,j'}^2 = V_{N_1}$, where V_{N_1} is defined analogously to (4.12) with $\eta_k = \eta_{k,j'}$. Note that $\eta_{k,j'}$ are independent of variables T_{jk}, K_{jk}^L, and $Y_{j',k}$. Also, V_{N_1} is a function of these variables. Let $\mathcal{B}^*$ be the σ-algebra generated by all random variables T_{jk}, K_{jk}^L, and $Y_{j',k}$, $j, j' \in I^{(0)}$, $k \in I^{(2)}$. As Z_{N_1} is measurable with respect to the σ-algebra $\mathcal{B}^*$, by applying (3.18) and (4.14), we see

$$\mathbb{P}\Big((W_{j,j'} \ge u_2) \implies (|Q_{j,j'}^2| \le h_0)\Big) \ge 1 - 2\exp(-e^{-2\beta} N_1 u_2 h_0^2/4).$$

By doing this analysis for all $j, j' \in I^{(0)}$ and summing up the results, we obtain the claim. ∎

Next we analyze $Q_{j,j'}^1$. We assume that L is so large and $\varepsilon_2, \varepsilon_3$ are so small that

$$\varepsilon_2 + \varepsilon(L) \le \frac{1}{200}\rho^2, \quad \varepsilon_3 \le \frac{b}{4}u_1. \tag{4.15}$$

We denote (see (4.6))

$$W_{j,j'}^{(d),-} = W^{(d),-}(X_j, X_{j'}), \quad W_{j,j'}^{(r),-} = W^{(r),-}(X_j, X_{j'}), \quad W_{j,j'} = W(X_j, X_{j'}).$$

Let us consider the events $\mathcal{E}_{j,j'}^{(4)} \subset \Omega$ and $\mathcal{E}^{(4)} \subset \Omega$; then

$$\mathcal{E}_{j,j'}^{(4)} = \Big\{\omega \in \Omega :\ (W_{j,j'}^{(d),-} \ge u_0) \implies (W_{j,j'}^{(r),-} \ge \frac{1}{2}u_0)\Big\}, \quad \mathcal{E}^{(4)} = \bigcap_{(j,j')\in I^{(0)}\times I^{(0)}} \mathcal{E}_{j,j'}^{(4)}.$$

Observe that

$$\Big\{\omega \in \Omega :\ |W^{(d),-}(X_j, X_{j'}) - W^{(r),-}(X_j, X_{j'})| \le \frac{1}{2}u_0\Big\} \subset \mathcal{E}_{j,j'}^{(4)}. \tag{4.16}$$

We see using Hoeffding's inequality for $y, z \in M$ that

$$\mathbb{P}\Big(|W^{(d),-}(y,z) - W^{(r),-}(y,z)| \le \frac{1}{2}u_0\Big) \ge 1 - 2\exp(-2N_1(u_1/2)^2).$$

Thus we see using (4.16) for all $j, j' \in I^{(0)}$ and (3.18), and summing the results,

$$\mathbb{P}(\mathcal{E}^{(4)}) \ge 1 - p^{(4)}, \quad p^{(4)} = 2N_0^2\exp(-2N_1(u_1/2)^2).$$

Below, to shorten notation, let us denote $\mathcal{E}^{(5)} = \bigcap_{k\in\{1,2,3,4\}} \mathcal{E}^{(k)}$.

Lemma 4.4. *Assume the event $\mathcal{E}^{(5)}$ happens. Then, if (4.15) is valid, we have for all $(j,j') \in I^{(0)} \times I^{(0)}$ the implication $(W^{(d),-}_{j,j'} \geq u_0) \implies (W_{j,j'} \geq u_2)$.*

Proof. As $\mathcal{E}^{(5)}$ happens, also the event $\mathcal{E}^{(2)}$ happens. If $0 \leq s \leq \frac{1}{20}\rho^2$, then $\psi_\rho(s - \frac{1}{10^2}\rho^2) = 1$ and $\psi_{\rho/2}(s) = 1$. On the other hand, if $s > \frac{1}{20}\rho^2$, we have $4(s - \frac{1}{10^2}\rho^2) > s$. Thus, as the function $\psi_1|_{\mathbb{R}_+}$ is nonincreasing,

$$\psi_\rho\left(s - \frac{1}{10^2}\rho^2\right) \geq \psi_{\rho/2}\left(4\left(s - \frac{1}{10^2}\rho^2\right)\right) \geq \psi_{\rho/2}(s)$$

for all $s \geq 0$. Hence, we have

$$\psi_\rho(s - 2(\varepsilon_2 + \varepsilon(L))) \geq \psi_\rho\left(s - \frac{1}{10^2}\rho^2\right) \geq \psi_{\rho/2}(s).$$

Thus, as $\mathcal{E}^{(2)}$ happens,

$$\psi_\rho(K^L_{jk}) \geq \psi_\rho(k_\Phi(X_j, X_k) - 2(\varepsilon_2 + \varepsilon(L))) \geq \psi_{\rho/2}(k_\Phi(X_j, X_k)). \tag{4.17}$$

When $\mathcal{E}^{(5)}$ happens, also $\mathcal{E}^{(4)}$ happens, and we have the implication $(W^{(d),-}_{j,j'} \geq u_0) \implies (W^{(r),-}_{j,j'} \geq u_1)$.

Recall that $\|\widetilde{\psi}_1\|_{C^1(\mathbb{R})} \leq 2$. Thus, when $\mathcal{E}^{(1)}$ happens,

$$|\widetilde{\psi}_1(T_{jk}(bN_2)^{-1}) - \widetilde{\psi}_1(A(X_j, X_k)b^{-1})| \leq 2b^{-1}\varepsilon_3. \tag{4.18}$$

Then

$$\begin{aligned}&|\widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_{\rho/2}(k_\Phi(X_j, X_k)) - \widetilde{\psi}_1(A(X_j, X_k)b^{-1})\psi_{\rho/2}(k_\Phi(X_j, X_k))| \\ &\leq 2b^{-1}\varepsilon_3 \cdot \psi_{\rho/2}(k_\Phi(X_j, X_k)) \leq 2b^{-1}\varepsilon_3.\end{aligned}$$

This and (4.17) imply

$$\begin{aligned}\widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_\rho(K^L_{jk}) &\geq \widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_{\rho/2}(k_\Phi(X_j, X_k)) \\ &\geq \widetilde{\psi}_1(A(X_j, X_k)b^{-1})\psi_{\rho/2}(k_\Phi(X_j, X_k)) - 2b^{-1}\varepsilon_3.\end{aligned} \tag{4.19}$$

Recall that by the assumption of the claim, $\varepsilon_3 < \frac{b}{4}u_1$. Computing the average of the inequalities (4.19) times $Y_{j',k}$ over $k \in I^{(1)}$, we obtain

$$W_{j,j'} \geq W^{(r),-}_{j,j'} - 2b^{-1}\varepsilon_3 \geq W^{(r),-}_{j,j'} - \frac{1}{2}u_1.$$

Thus when $\mathcal{E}^{(5)}$ and thus $\mathcal{E}^{(4)}$ happen, we have

$$(W^{(d),-}_{j,j'} \geq u_0) \implies (W^{(r),-}_{j,j'} \geq u_1) \implies \left(W_{j,j'} \geq u_1 - \frac{1}{2}u_1 = u_2\right). \qquad \blacksquare$$

Lemma 4.5. *When the event $\mathcal{E}^{(5)}$ happens and (4.15) is valid, it holds for all $j, j' \in I^{(0)}$ that if $W(X_j, X_{j'}) \geq u_2$, then*

$$|Q_{j,j'} - d_M(X_j, X_{j'})| \leq \frac{2}{\kappa}\rho + h_0. \tag{4.20}$$

Proof. In this proof assume that the event $\mathcal{E}^{(5)}$ happens. We will first show using Lemma 3.8 that if $W(X_j, X_{j'}) \geq u_2$, then

$$|Q^1_{j,j'} - d_M(X_j, X_{j'})| \leq \frac{2}{\kappa}\rho. \tag{4.21}$$

Consider next the indexes $(j,k) \in I^{(0)} \times I^{(1)}$ for which $\psi_\rho(K^L_{jk}) > 0$ and $\widetilde{\psi}_1(T_{jk}(bN_2)^{-1})) > 0$. Then $\psi_\rho(K^L_{jk}) > 0$ implies $K^L_{jk} < 2\rho^2$, and hence, as the event $\mathcal{E}^{(2)}$ happens,

$$k_\Phi(X_j, X_k) < 2\rho^2 + \varepsilon_2 + \varepsilon(L). \tag{4.22}$$

Moreover, if $\widetilde{\psi}_1(T_{jk}(bN_2)^{-1}) > 0$, then $T_{jk}N_2^{-1} > b$. As the event $\mathcal{E}^{(5)}$ happens, we have

$$A(X_j, X_k) = A_{jk} \geq T_{jk}N_2^{-1} - \varepsilon_3 \geq b - \varepsilon_3 \geq \frac{1}{4}a \geq \widehat{a};$$

see Proposition 3.5 and (3.19). Then by Proposition 3.5,

$$k_\Phi(X_j, X_k)^{1/2} \geq \kappa d_M(X_j, X_k), \tag{4.23}$$

which implies with (4.22) that

$$d_M(X_j, X_k) \leq \frac{1}{\kappa}k_\Phi(X_j, X_k)^{1/2} \leq \frac{1}{\kappa}\sqrt{2\rho^2 + \varepsilon_2 + \varepsilon(L)}. \tag{4.24}$$

We denote $v_{j,j',k} = \widetilde{\psi}_1(T_{jk}(bN_2)^{-1})\psi_\rho(K^L_{jk})Y_{j'k}$ and observe that if $v_{j,j',k} > 0$, then the distance $d_M(X_j, X_k)$ satisfies (4.24).

Assume next that $W_{j,j'} \geq u_2$. Then, we have $\sum_{k\in I^{(1)}} v_{j,j',k} = W(X_j, X_{j'}) \geq u_2 > 0$ and

$$Q^1(X_j, X_{j'}) = V^1(X_j, X_{j'})/W(X_j, X_{j'}) = \frac{\sum_{k\in I^{(1)}} d_M(X_k, X_{j'})\, v_{j,j',k}}{\sum_{k\in I^{(1)}} v_{j,j',k}}.$$

This means that we can consider $Q^1(X_j, X_{j'})$ as a weighted average of distances $d_M(X_k, X_{j'})$. Hence, when the event $\mathcal{E}^{(5)}$ happens, we see that for all $(j,j') \in I^{(0)} \times I^{(0)}$ such that $W_{j,j'} \geq u_2$, we have

$$\begin{aligned} |Q^1_{j,j'} - d_M(X_j, X_{j'})| &= \left| \frac{\sum_{k\in I^{(1)}} (d_M(X_k, X_{j'}) - d_M(X_j, X_{j'}))\, v_{j,j',k}}{\sum_{k\in I^{(1)}} v_{j,j',k}} \right| \\ &\leq \frac{\sum_{k\in I^{(1)}} d_M(X_k, X_j)\, v_{j,j',k}}{\sum_{k\in I^{(1)}} v_{j,j',k}} \leq \frac{1}{\kappa}\sqrt{2\rho^2 + \varepsilon_2 + \varepsilon(L)}, \end{aligned}$$

where we have used (4.24) in the last inequality. By (4.15), here $\varepsilon_2 + \varepsilon(L) < \frac{1}{100}\rho^2$, and the inequality (4.21) follows.

As we have assumed that the event $\mathcal{E}^{(5)}$ happens, also the event $\mathcal{E}^{(3)}$ happens; see (4.10). Thus if $W_{j,j'} \geq u_2$, then $Q^2_{j,j'} < h_0$. Combining the above, we see that

$$|Q_{j,j'} - d_M(X_j, X_{j'})| \leq |Q^1_{j,j'} - d_M(X_j, X_{j'})| + |Q^2_{j,j'}| \leq \frac{2}{\kappa}\rho + h_0. \qquad \blacksquare$$

The above proposition means that if $W_{j,j'} \geq u_2$, then the number $Q_{j,j'}$ approximates $d_M(X_j, X_{j'})$ with a large probability when all parameters are suitably chosen.

Next we consider the proof of Theorem 2. We use below

(4.25)
$$\rho = \frac{1}{4}\kappa\varepsilon_1, \quad h_0 = \frac{\varepsilon_1}{2}, \quad \varepsilon_2 = \frac{\rho^2}{800} = \frac{\kappa^2\varepsilon_1^2}{2^7 \cdot 10^2}, \quad \varepsilon_3 = \frac{b}{4}u_1 = \frac{a}{16}\phi_1 c_0 (\kappa\varepsilon_1/4)^n, \quad L = 4\log^2\left(\frac{C_4\beta^2}{\rho}\right),$$

where $C_4 = 1200e^{2D}$. Note that $\rho \leq 1$. Then $\varepsilon(L) = \rho^2/400$, implying that $\varepsilon_2 + \varepsilon(L) = \rho^2/200$. Below we assume that $\varepsilon_1 < \widehat{\varepsilon}_1$, where $\widehat{\varepsilon}_1 = \min(1, 8\kappa(\phi_1 c_0)^{-1/n})$. Then we have $\varepsilon_3 < \frac{1}{4}a$; see (3.19).

Below, let N_0 be

$$N_0 = \lfloor 2C_3\delta_1^{-n}(\log(\delta_1^{-1}) + \log(\theta^{-1}))\rfloor. \tag{4.26}$$

Next we consider the probability of $\mathcal{E}^{(5)}$. We see that $\mathbb{P}(\mathcal{E}^{(5)}) \geq 1 - p^{(5)}$, where $p^{(5)} = p^{(1)} + p^{(2)} + p^{(3)} + p^{(4)}$. Next we consider the probabilities one by one.

In the inequalities below, we use that $t \log t \leq t^2$, so that for $t > e$ we have $\log(t \log t) \leq 2 \log t$. Also, we use that for $t, s \geq 2$ we have $\log(t + s) \leq (\log t) \cdot (\log s)$ and $\log t \leq t$.

Lemma 4.6. *There is $C_7 > 1$ such that we have $p^{(3)} < \theta/8$ when*

$$N_1 \geq C_7\phi_1^{-1}\kappa^{-n}\varepsilon_1^{-n-2}\left(\log\left(\frac{1}{\delta_1}\right) + \log\left(\frac{1}{\theta}\right)\right). \tag{4.27}$$

Proof. Below, we use that $t \leq -\log(1-t)$ for $0 < t < 1$. We see that $p^{(3)} < \theta/8$ if

$$N_1 \geq R_1 = \frac{2^{2n+4}e^{2\beta}\log(\frac{16N_0^2}{\theta})}{\phi_1 c_0 (\kappa\varepsilon_1)^n h_0^2} = \frac{2^{2n+8}e^{2\beta}\log(\frac{16N_0^2}{\theta})}{\phi_1 c_0 \kappa^n \varepsilon_1^{n+2}}. \tag{4.28}$$

We state that N_0 is given in (4.26). We see that

$$R_1 = \frac{2^{2n+8}e^{2\beta}\log(\frac{16N_0^2}{\theta})}{\phi_1 c_0 \kappa^n \varepsilon_1^{n+2}} \leq C_9\phi_1^{-1}\kappa^{-n}\varepsilon_1^{-n-2}\left(\log\left(\frac{1}{\delta_1}\right) + \log\left(\frac{1}{\theta}\right)\right) = P_1,$$

where C_9 is suitable. Thus (4.28) is valid when $N_1 \geq P_1$. This yields that claim. $\blacksquare$

Lemma 4.7. *There is $C_{10} > 2C_7$ such that we have $p^{(4)} < \theta/8$ when*

$$N_1 = \left\lfloor C_{10}\phi_1^{-2}\kappa^{-2n}\varepsilon_1^{-2n}\left(\log\left(\frac{1}{\delta_1}\right) + \log\left(\frac{1}{\theta}\right)\right)\right\rfloor. \tag{4.29}$$

Note that when N_1 is chosen as in (4.29), the inequality (4.27) is also valid.

Proof. Next we estimate $p^{(4)} = 2N_0^2 \exp(-2N_1(u_1/2)^2)$. We see that $p^{(4)} < \theta/8$ if

$$N_1 \geq R_2 = \frac{\log(\frac{16N_0^2}{\theta})}{2(u_1/2)^2} = \frac{2\log(\frac{16N_0^2}{\theta})}{\phi_1^2 c_0^2((\kappa\varepsilon_1)/4)^{2n}}. \tag{4.30}$$

Also, we see that

$$R_2 \leq \left\lfloor C_{10}\phi_1^{-2}\kappa^{-2n}\varepsilon_1^{-2n}\left(\log\left(\frac{1}{\delta_1}\right) + \log\left(\frac{1}{\theta}\right)\right)\right\rfloor = P_2,$$

where C_{10} is suitably chosen. Thus (4.30) is valid when $N_1 \geq P_2$. ∎

Lemma 4.8. *There is $C_{12} > 0$ such that $p^{(2)} < \theta/8$ when*

$$N_2 \geq C_{12}\phi_1^{-1}\log^4(e+\beta^2)\,(\kappa\varepsilon_1)^{-4}\left(\log\left(\frac{1}{\kappa\varepsilon_1}\right)\right)^4\left(\log\left(\frac{1}{\theta}\right)+\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\kappa\varepsilon_1}\right)\right). \tag{4.31}$$

Proof. Let N_0 and N_1 be given as in (4.26) and (4.29). We see that $p^{(2)} \leq \theta/8$ if

$$N_2 \geq R_3 = \frac{1}{2}\varepsilon_2^{-2}L^2\log\left(\frac{16N_0N_1}{\theta}\right). \tag{4.32}$$

We have

$$\begin{aligned}
R_3 &\leq C_{13}(\kappa^2\varepsilon_1^2)^{-2}L^2\log\left(\frac{16}{\theta}\cdot 2C_3\delta_1^{-n}\left(\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\theta}\right)\right)\right.\\
&\quad\left.\cdot C_{10}\phi_1^{-2}\kappa^{-2n}\varepsilon_1^{-2n}\left(\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\theta}\right)\right)\right)\\
&\leq C_{12}(\kappa^2\varepsilon_1^2)^{-2}\log^4\left(\frac{C_4\beta^2}{\kappa\varepsilon_1}\right)\left(\log\left(\frac{1}{\theta}\right)+\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\kappa\varepsilon_1}\right)+\log\frac{1}{\phi_1}\right)\\
&\leq C_{12}\phi_1^{-1}\log^4(e+\beta^2)\,(\kappa^2\varepsilon_1^2)^{-2}\log^4\left(\frac{1}{\kappa\varepsilon_1}\right)\left(\log\left(\frac{1}{\theta}\right)+\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\kappa\varepsilon_1}\right)\right) = P_3,
\end{aligned}$$

where C_{12} and C_{13} are suitable. Thus (4.32) is valid when $N_2 \geq P_3$. This yields the claim. ∎

Lemma 4.9. *There is $C_{14} > 0$ such that we have $p^{(1)} < \theta/8$ when*

$$N_2 \geq C_{14}\phi_1^{-3}a^{-2}(\kappa\varepsilon_1)^{-2n}\left(\log\left(\frac{1}{\theta}\right)+\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\kappa\varepsilon_1}\right)\right). \tag{4.33}$$

Proof. Using (4.25), we see that the inequality

$$p^{(1)} = 2N_0N_1\exp(-2N_2\varepsilon_3^2) = 2N_0N_1\exp\left(-2N_2\frac{\phi_1^2}{2^{4n+8}}c_0^2a^2(\kappa\varepsilon_1)^{2n}\right) < \frac{1}{8}\theta \tag{4.34}$$

is valid when $N_2 \geq R_4 = 2^{4n+7}\phi_1^{-2}c_0^{-2}a^{-2}(\kappa\varepsilon_1)^{-2n}\log(\frac{16N_0N_1}{\theta})$. We see that

$$\begin{aligned}
R_4 &\leq \frac{2^{4n+7}}{\phi_1^2c_0^2a^2(\kappa\varepsilon_1)^{2n}}\log\left(\frac{16}{\theta}\cdot 2C_3\delta_1^{-n}\left(\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\theta}\right)\right)\right.\\
&\quad\left.\cdot C_{10}\phi_1^{-2}\kappa^{-2n}\varepsilon_1^{-2n}\left(\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\theta}\right)\right)\right)\\
&\leq C_{14}\phi_1^{-2}(\log(\phi_1^{-1}))a^{-2}(\kappa\varepsilon_1)^{-2n}\left(\log\left(\frac{1}{\theta}\right)+\log\left(\frac{1}{\delta_1}\right)+\log\left(\frac{1}{\kappa\varepsilon_1}\right)\right) = P_4,
\end{aligned}$$

where C_{14} is suitable. Thus (4.34) is valid when $N_2 \geq P_4$. This yields the claim. ∎

Next we prove Theorems 2 and 1.

Proof of Theorem 2. We observe that when $\mathcal{E}^{(5)}$ happens, by Lemmas 4.2 and 4.4, for all X_j and $X_{j'}$ such that $d_M(X_j, X_{j'}) < r_1$, we have $W_{j,j'} \geq u_2$.

We recall that $\kappa = c_1 H^{-(n/2+1)} \lambda_1 \lambda_2^{-n/2-2} r_1^{(n+4)n} \phi_1^{n+4} \leq 1/2$, $a = c_0 \phi_1^2 r_1^n \leq 1/2$, $r_1 = \frac{1}{2} r_0 \geq \frac{\pi}{4\Lambda}$, $\phi_1 = \frac{\lambda_1}{2}\phi_0$. Let N_0 and N_1 be given as in (4.26) and (4.29). The conditions (4.33) and (4.31) are valid when

(4.35)

$$\begin{aligned} N_2 &\geq \left\lfloor C_{15} \log^4(e+\beta)\, \phi_1^{-2} (\log(\phi_1^{-1})) a^{-2} \kappa^{-2n} (\log(\kappa^{-1}))^5 \varepsilon_1^{-2n} Q_n(\varepsilon_1) \left(\log\left(\frac{1}{\theta}\right) \right.\right. \\ &\quad \left.\left. + \log\left(\frac{1}{\delta_1}\right) + \log\left(\frac{1}{\varepsilon_1}\right)\right)\right\rfloor \\ &\geq \left\lfloor C_{16} \log^4(e+\beta)\, \frac{H^{n^2+2n+1}}{\phi_0^{2n^2+8n+7}} \frac{\lambda_2^{n^2+4n+1}}{\lambda_1^{2n^2+10n+7}} \varepsilon_1^{-2n} Q_n(\varepsilon_1) \left(\log\left(\frac{1}{\theta}\right) + \log\left(\frac{1}{\delta_1}\right) + \log\left(\frac{1}{\varepsilon_1}\right)\right)\right\rfloor, \end{aligned}$$

where $Q_n(\varepsilon_1) = 1$ when $n \geq 3$, and $Q_n(\varepsilon_1) = \log^4(1/\varepsilon_1)$ when $n = 2$, and $C_{15}, C_{16} > 1$ are suitable numbers depending on n, D, i_0, and r_0. Then $p^{(5)} = p^{(1)} + p^{(2)} + p^{(3)} + p^{(4)} \leq \frac{1}{2}\theta$. When

$$C_2 = 2C_3 + C_{10} + C_{16}, \tag{4.36}$$

this and Lemma 2.1 prove that with probability $1 - \theta$ we have for all $(j, j') \in I^{(0)} \times I^{(0)}$ that inequality (2.3) holds when $d_M(X_j, X_{j'}) < r_1$, and the inequality (2.4) holds when $d_M(X_j, X_{j'}) \geq r_1$.

In the case when $\Phi(x, y) \geq \lambda_1 \phi_0$, for $(x, y) \in M \times M$, we see that when the event $\mathcal{E}^{(5)}$ happens, Lemmas 4.2 and 4.4 yield that we will have with probability $1 - \theta$ that $W_{j,j'} > u_2$ for all (j, j'). This implies that the inequality (2.3) holds for all pairs $(X_j, X_{j'})$ with $j, j' \in \{1, 2, \ldots, N_0\}$. ∎

Proof of Theorem 1. Let $\varepsilon_1 = \delta^{3/2}$ and $\delta_1 = \Lambda^{-2/3}\delta^{1/2}/20$. For N_0, N_1 given in (4.26) and (4.29),

(4.37)

$$N_0 \leq \left\lfloor C_{17} \delta^{-n/2} \left(\log\left(\frac{1}{\theta}\right) + \log\left(\frac{1}{\delta}\right)\right)\right\rfloor, \quad N_1 \leq \left\lfloor C_{18} p_{H,\phi_0,\beta} \delta^{-3n} \left(\log\left(\frac{1}{\theta}\right) + \log\left(\frac{1}{\delta}\right)\right)\right\rfloor,$$

with suitable C_{17} and C_{18}. Moreover, N_2 satisfies (4.35) when we choose a suitable $C_{19} \geq C_{18}$ and

$$N_2 = \left\lfloor C_{19} p_{H,\phi_0,\beta} \delta^{-3n} Q_n(\delta) \left(\log\left(\frac{1}{\theta}\right) + \log\left(\frac{1}{\delta}\right)\right)\right\rfloor. \tag{4.38}$$

Let $\widehat{\delta} = \varepsilon_1 = \delta^{3/2}$ and $\widehat{r} = (\widehat{\delta}/\Lambda^2)^{1/3} = \Lambda^{-2/3}\delta^{1/2}$. Then by Theorem 2, with the probability $1 - \theta$ the set $\mathcal{X} = \{X_j : j = 1, 2, \ldots, N_0\}$ is a δ_1-dense subset of M, and the

approximate distances $\widetilde{d}(X_j, X_{j'}) = d^{app}(X_j, X_{j'})$, $j, j' \in \{1, 2, \ldots, N_0\}$ (see (4.5)) satisfy the conditions given in Proposition A.1 in Appendix A. Thus with probability $1-\theta$ we can apply Proposition A.1 (see also [24, Corollary 1.10]) with $\widehat{\delta} = \delta^{3/2}$ and $\widehat{r} = (\widehat{\delta}/\Lambda^2)^{1/3}$ to construct a Riemannian manifold (M^*, g^*) that approximates the original manifold (M, g) so that the claims (1)–(3) in Theorem 1 are satisfied. ∎

Remark 4.10. Recall that above, in particular in the claims of Theorems 1 and 2, we have assumed that the dimension n, the variance σ^2, and the expectation μ of the noise are a priori known. However, if these parameters are not known, we can use the noisy distance $\overline{D}_{jk}$ to estimate these parameters. Next, we explain how this can done.

First we consider the dimension n. Observe that the bounds for N_0, N_1, and N_2 are increasing functions of n. If we have an upper bound n_0 for the dimension n and use N_0, N_1, and N_2 as required for n_0, the algorithm NRD returns a discrete set X and an approximate distance function $d_X : X \times X \to \mathbb{R}$ that can be considered as a δ_1-net of the manifold M and the distance function d_M with a small deterministic error. For the pair (X, d_X) we can apply the algorithms described in [24] to determine the dimension n of M.

Let us next consider the estimation of σ. Using Lemma 2.1, Proposition 3.5, (3.13), (3.15), (3.20), and (3.30), we see that when $\delta_1 \le \kappa\rho/20$ and the event $\mathcal{E}^{(5)}$ happens, σ^2 can be estimated with accuracy $\rho^2/400$ by choosing σ to have the value

$$\sigma_{est} = \sup\Big\{\sigma \ge 0 : \ K_{jk}^{L,\sigma} \ge 0 \text{ for all } (j,k) \text{ satisfying } T_{jk}/N_2 \ge b\Big\}. \tag{4.39}$$

In other words, we choose σ so that the reconstructed values $K_{jk}^{L,\sigma}$, that should approximate $k_\Phi(X_j, X_k) = \|(r_{X_j} - r_{X_k})\Phi_{X_j}^{1/2}\Phi_{X_k}^{1/2}\|^2_{L^2(M,dv)} \ge 0$, are nonnegative when $T_{jk}/N_2 \ge b$.

Next we consider the estimation of μ. To do this, observe that by the above considerations

$$|K_{jk}^{L,\sigma_{est}} - K_{jk}^{L,\sigma}| \le \rho^2/400$$

for all (j,k) satisfying $T_{jk}/N_2 \ge b$. Moreover, by Lemma 4.2 and the proof of Lemma 4.4 (see, in particular, formula (4.17) and the definition of the event $\mathcal{E}^{(2)}$), we see that for all X_j and $X_{j'}$, $j, j' \in I^{(0)}$, such that $d_M(X_j, X_{j'}) < r_1$, we have $W_{j,j'}^{(\sigma_{est})} \ge u_2$. These yield that when $\delta_1 \le \kappa\rho/20$ and the event $\mathcal{E}^{(5)}$ happens, the parameter μ can be estimated with accuracy $\varepsilon_1 + \delta_1$ by choosing μ to be

$$\mu_{est} = \Big\{\mu \in \mathbb{R} : \ Q_{j,j'}^{(\mu,\sigma_{est})} \ge 0 \text{ for all } j, j' \in I^{(0)} \text{ satisfying } W_{j,j'}^{\sigma_{est}} > u_2\Big\}. \tag{4.40}$$

Note that here we use the fact that $\frac{2}{\kappa}\rho + h_0 = \varepsilon_1$ and that the points $\{X_j : \ j \in I^{(0)}\}$ form a δ_1-net in M when the event $\mathcal{E}^{(5)}$ happens.

5. Discussion and future directions.

The questions studied in this paper are related to imaging applications. In particular, many inverse problems, where one aims to determine the coefficient functions of a partial differential equation from indirect measurements, can be formulated as geometric problems for Riemannian manifolds. For example, when waves propagate in an inhomogeneous medium, the travel times between the points of the medium define

a non-Euclidean metric called the *travel time metric*. Thus the problem of determining the wave speed function inside a body from the external or boundary measurements leads to the problem of determining a Riemannian metric on a manifold. An example where the physical structures are represented by abstract Riemannian manifolds is seen in medical ultrasound imaging, where the acoustic properties inside a body are imaged. The typical ultrasound image corresponds to an image of the body represented in the non-Euclidean travel time coordinates, more precisely, in the Riemannian normal coordinates of the travel time metric [51]. In these non-Euclidean coordinates, the image rays (i.e., the geodesics) emanating from the location of the source device are straight lines. The global structure of the manifold is represented by a collection of such images where the location of the source device is moved. Similar inverse problems of determining the wave speed inside a body from external measurements are encountered in seismic imaging of the Earth [57] and in magnetic resonance imaging (MRI) based elastography used in medical imaging, where the elastic properties of the body of a patient are determined by observing the propagation of elastic waves sent into the body [36]. The relation of the boundary measurements for a wave equation and the distances between the points in a δ-net in the interior of the manifold are considered, e.g., in [2, 17, 40, 41]. It would be interesting to study how the random errors in the boundary measurements propagate in the imaging algorithms to the errors in the reconstructed travel time distances of the points in the interior of the domain and, moreover, whether the statistics of these errors satisfy the assumptions of Theorem 1 in this paper.

Also, in this paper we have assumed that the variance σ of the noise is constant. An interesting open problem is the reconstruction of the manifold when the variance depends on the geodesic distance of the points or, more generally, only satisfy upper and lower bounds that depend on the distance of the points. Indeed, it would be realistic to assume that noise accumulates along the geodesic path and hence that shorter distances would be less noisy.

Appendix A. Reconstruction of a manifold with a small deterministic errors. Here, we give the results on the reconstruction of a Riemannian manifold when one is given distances with small deterministic errors. The following result is an improvement of Corollary 1.10 in [24].

Proposition A.1. *There are $C_n' > 0$ depending on n and $c_1'(n,K) > 0$ depending on n, K such that the following holds: Let $0 < \widehat{\delta} < c_1'(n,K)$, $\widehat{r} = (\widehat{\delta}/K)^{1/3}$, and M be a compact n-dimensional manifold with $|\mathrm{Sec}(M)| \le K$ and $\mathrm{inj}(M) > 2\widehat{r}$. Let $\mathcal{X} = \{x_j\}_{j=1}^N$ be an $\widehat{r}/20$-dense subset of M. Moreover, let $\widetilde{d}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \cup \{0\}$ be an approximate local distance function that satisfies*

$$|\widetilde{d}(x,y) - d_M(x,y)| \le \widehat{\delta} \quad \text{if } d_M(x,y) < \widehat{r}, \tag{A.1}$$

$$\widetilde{d}(x,y) > \widehat{r} - \widehat{\delta} \quad \text{if } d_M(x,y) \ge \widehat{r}. \tag{A.2}$$

Then, given the values $\widetilde{d}(x_j,x_k)$, $j,k = 1,2,\dots,N$, one can construct a compact n-dimensional Riemannian manifold (M^*,g^*) such that the following hold.

(1) *There is a diffeomorphism $F : M^* \to M$ satisfying*

$$\frac{1}{L} \le \frac{d_M(F(x),F(y))}{d_{M^*}(x,y)} \le L \quad \text{for } x,y \in M^*,\ L = 1 + C_n' K^{1/3} \widehat{\delta}^{\,2/3}.$$

(2) $|\operatorname{Sec}(M^*)| \le C_n' K$.
(3) *The injectivity radius* $\operatorname{inj}(M^*)$ *of* M^* *satisfies*

$$\operatorname{inj}(M^*) \ge \min\{(C_n' K)^{-1/2}, (1 - C_n' K^{1/3} \widehat{\delta}^{\,2/3}) \operatorname{inj}(M)\}.$$

Proof. A result similar to the claim is proven in [24, Corollary 1.10] under the assumption that the set $\mathcal{X}$ is a $\widehat{\delta}$-dense subset of M, instead of $\widehat{r}/20$-dense as it is assumed in the claim. Moreover, by [24, Corollary 1.10], it is sufficient to construct numbers $\widetilde{D}_{j,k}$, $j,k = 1,2,\dots,\widetilde{N}$, such that the following is true: There is a $\widehat{\delta}$-net $\mathcal{Y} = \{y_j : j = 1,2,\dots,\widetilde{N}\} \subset M$ such that the conditions (A.1) and (A.2) are valid for the function $\widetilde{d}' \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{0\}$ defined by $\widetilde{d}'(y_j, y_k) = \widetilde{D}_{j,k}$.

Next we construct the required $\widehat{\delta}$-net $\mathcal{Y} \subset M$ and an approximate distance function $\widetilde{d}'$ on $\mathcal{Y} \times \mathcal{Y}$. We assume that $c_1'(n,K)$ is chosen so small that $\widehat{\delta}/\widehat{r} = K\widehat{r}^2 < \frac{1}{150}$.

For $p \in M$ we denote by E_p the restriction of the Riemannian exponential map $\exp_p$ to the $\widehat{r}$-ball in T_pM centered at the origin. This restriction is a diffeomorphism onto the $\widehat{r}$-ball centered at p in M. It distorts distances by at most $\frac{1}{2}\widehat{\delta}$, namely for all $u, v \in T_pM$ such that $|u|, |v| < \widehat{r}$, we have

$$\left| d_M(E_p(u), E_p(v)) - |u - v| \right| < \tfrac{1}{2} K \widehat{r}^3 = \tfrac{1}{2}\widehat{\delta}. \tag{A.3}$$

This inequality holds as long as $\operatorname{inj}(M) > 2\widehat{r}$ and $K\widehat{r}^2 < \pi/2$; see [24, section 4] for a proof.

For every $p \in \mathcal{X}$, define $X_p = \{x \in \mathcal{X} : \widetilde{d}(p,x) < \widehat{r}/6 - \widehat{\delta}\}$. By (A.1) and (A.2), X_p is contained in the $\widehat{r}/6$-neighborhood of p. Define $\widetilde{X}_p = E_p^{-1}(X_p)$, let $\widetilde{V}_p$ be the convex hull of $\widetilde{X}_p$ in T_pM, and let $V_p = E_p(\widetilde{V}_p)$. Since $\mathcal{X}$ is $\widehat{r}/20$-dense in M, (A.1) and (A.3) imply that for every $u \in T_pM$ such that $|u| < \widehat{r}/6 - \widehat{r}/20 - 2\widehat{\delta}$ there exists $v \in \widetilde{X}_p$ such that $|u - v| < \widehat{r}/20 + \widehat{\delta}/2$. This implies that $\widetilde{V}_p$ contains the ball of radius $\widehat{r}/6 - 2\widehat{r}/20 - 2\widehat{\delta} - \widehat{\delta}/2 > \widehat{r}/20$ centered at the origin. (Here we use the assumption that $\widehat{\delta}/\widehat{r} < \frac{1}{150}$.) Hence V_p contains the $\widehat{r}/20$-ball centered at p and, therefore, $\bigcup_{p \in \mathcal{X}} V_p = M$.

We represent points of $\widetilde{V}_p$ as linear combinations of points of $\widetilde{X}_p$ as follows. Let X_p^n be the set of all n-tuples of points of X_p, and let Δ^n be the standard coordinate simplex in $\mathbb{R}^n$:

$$\Delta^n = \left\{ (t_1, \dots, t_n) \in \mathbb{R}^n : t_1, \dots, t_n \ge 0, \sum t_i \le 1 \right\}.$$

For $\alpha = (a_1, \dots, a_n) \in X_p^n$ and $\tau = (t_1, \dots, t_n) \in \Delta^n$, let

$$S_p(\alpha, \tau) = \sum t_i E_p^{-1}(a_i).$$

This defines a map $S_p \colon X_p^n \times \Delta^n \to T_pM$. For a fixed $\alpha = (a_1, \dots, a_n) \in X_p^n$, the range of $S_p(\alpha, \cdot)$ is a (possibly degenerate) affine simplex in T_pM with vertices $0, E_p^{-1}(a_1), \dots, E_p^{-1}(a_n)$. Since $0 \in \widetilde{X}_p$, the union of all such affine simplices is precisely the convex hull of $\widetilde{X}_p$. Thus $S_p(X_p^n \times \Delta^n) = \widetilde{V}_p$.

Fix an ε'-dense finite set $\Sigma \subset \Delta^n$, where $\varepsilon' = \widehat{\delta}/(3\widehat{r}\sqrt{n})$, and define $Y_p = E_p(S_p(X_p^n \times \Sigma)) \subset M$. Since $\widetilde{V}_p$ is contained in the $\widehat{r}/6$-ball, $S_p(\alpha, \cdot)$ is Lipschitz with Lipschitz constant $\widehat{r}\sqrt{n}/6$. Therefore $S_p(X_p^n \times \Sigma)$ is $\widehat{\delta}/2$-dense in $\widetilde{V}_p$. Hence, by (A.3), Y_p is $\widehat{\delta}$-dense in V_p.

Now define $\mathcal{Y} \subset M$ by $\mathcal{Y} = \bigcup_{p\in\mathcal{X}} Y_p$. Since the sets V_p cover M and Y_p is $\widehat{\delta}$-dense in V_p for each p, $\mathcal{Y}$ is a $\widehat{\delta}$-net in M. The points of $\mathcal{Y}$ are indexed by triples (p, α, τ) where $p \in \mathcal{X}$, $\alpha \in X_p^n$, $\tau \in \Sigma$. This index set can be enumerated algorithmically using the known data.

Our first goal is to compute approximate *squared* distances $Q(x, y)$ between sufficiently close pairs of points $x, y \in \mathcal{Y}$. Fix $p, q \in \mathcal{X}$ such that $\widetilde{d}(p, q) < 2\widehat{r}/3 - 2\widehat{\delta}$ (the case $p = q$ is not excluded). By (A.1) and (A.2) we have $d_M(p, q) < 2\widehat{r}/3 - \widehat{\delta}$. Hence, by the triangle inequality, $d_M(x, y) < \widehat{r} - \widehat{\delta}$ for all $x \in V_p$ and $y \in V_q$. In particular V_q is contained in the range of E_p. By (A.3),

$$\left|d_M(x,y)^2 - |E_p^{-1}(x) - E_p^{-1}(y)|^2\right| < \widehat{\delta}\widehat{r} \qquad \text{for all } x \in V_p \text{ and } y \in V_q. \tag{A.4}$$

We compute the values $Q(x, y)$ for all $x \in X_p \cup Y_p$ and $y \in X_q \cup Y_q$ in several steps. First consider $x \in X_p$ and $y \in X_q$. In this case we simply define $Q(x, y) = \widetilde{d}(x, y)^2$. Then, by (A.1),

$$|Q(x,y) - d_M(x,y)^2| < 2\widehat{\delta}\widehat{r}. \tag{A.5}$$

Hence, by (A.4),

$$\left|Q(x,y) - |E_p^{-1}(x) - E_p^{-1}(y)|^2\right| < 3\widehat{\delta}\widehat{r}, \qquad x \in X_p,\ y \in X_q. \tag{A.6}$$

Now consider $x \in Y_p$ and $y \in X_q$. By the definition of Y_p we have $x = E_p(S_p(\alpha, \tau))$ for some $\alpha = (a_1, \dots, a_n) \in X_p^n$ and $\tau = (t_1, \dots, t_n) \in \Sigma$. We define $Q(x, y)$ using the values of Q that we have from the previous step. We introduce the following notation: $a_0 = p$, $t_0 = 1 - \sum_{i=1}^n t_i$, $v_i = E_p^{-1}(a_i)$ for $i = 0, \dots, n$ (in particular, $v_0 = 0$), $v = E_p^{-1}(x) = \sum t_i v_i$, and $w = E_p^{-1}(y)$. In this notation, $v - w = \sum_{i=0}^n t_i(v_i - w)$, hence

$$|v-w|^2 = \sum_{0\le i,j\le n} t_i t_j \langle v_i - w, v_j - w\rangle = \frac{1}{2}\sum_{0\le i,j\le n} t_i t_j (|v_i - w|^2 + |v_j - w|^2 - |v_i - v_j|^2),$$

where $\langle \cdot, \cdot \rangle$ is the scalar product in T_pM. With this identity in mind, we define

$$Q(x,y) = \frac{1}{2}\sum_{0\le i,j\le n} t_i t_j (Q(a_i, y) + Q(a_j, y) - Q(a_i, a_j)). \tag{A.7}$$

Since $y \in X_q$ and $a_i, a_j \in X_p$, the values $Q(a_i, y)$, $Q(a_k, y)$, and $Q(a_i, a_j)$ are defined as in the previous step (in the case of $Q(a_i, a_j)$, this is the previous step with $q = p$). Since $\sum_{i=0}^n t_i = 1$, applying (A.6) to the terms in the right-hand side of (A.7) yields that

$$|Q(x,y) - |E_p^{-1}(x) - E_p^{-1}(y)|^2| = |Q(x,y) - |v-w|^2| < \tfrac{9}{2}\widehat{\delta}\widehat{r} < 5\widehat{\delta}\widehat{r}. \tag{A.8}$$

Therefore, by (A.4),

$$|Q(x,y) - d_M(x,y)^2| < 6\widehat{\delta}\widehat{r}. \tag{A.9}$$

We now have the values $Q(x, y)$ satisfying (A.9) for all $x \in Y_p$ and $y \in X_q$. Exchanging the roles of p and q we similarly find $Q(x, y)$ for all $x \in X_p$ and $y \in Y_q$.

Finally, consider $x \in Y_p$ and $y \in Y_q$. Again, let $x = E_p(S_p(\alpha, \tau))$, where $\alpha = (a_1, \dots, a_n) \in X_p^n$ and $\tau = (t_1, \dots, t_n) \in \Sigma$, and define $a_0 = p$ and $t_0 = 1 - \sum_{i=1}^n t_i$. From the previous steps we already have values $Q(a_i, y)$ and $Q(a_i, a_j)$. Therefore, we can define $Q(x, y)$ by the same formula (A.7). Then, starting from (A.9) instead of (A.5), we obtain the same estimates as above but with different constants: (A.6) with $7\widehat{\delta}\widehat{r}$ in the right-hand side, (A.8) with $11\widehat{\delta}\widehat{r}$ in the right-hand side, and finally (A.9) with $12\widehat{\delta}\widehat{r}$ in the right-hand side:

$$|Q(x, y) - d_M(x, y)^2| < 12\widehat{\delta}\widehat{r}, \qquad x \in Y_p,\ y \in Y_q. \tag{A.10}$$

Now one might take the square root of $Q(x, y)$ as an approximate distance between x and y; however, this approximation is not good enough. For a better one, we use an algorithm described in [24, section 4] to construct a map $F\colon Y_p \cup Y_q \to \mathbb{R}^n$ that preserves distances up to an error $O(\widehat{\delta})$. Let us outline how the algorithm works in the present set-up.

First define an approximate scalar product $P(x, y)$ for all pairs $x, y \in Y_p \cup Y_q$ by

$$P(x, y) = \tfrac{1}{2}(Q(p, x) + Q(p, y) - Q(x, y)).$$

By (A.4), (A.10), and the Euclidean identity $\langle u, v\rangle = \frac{1}{2}(|u|^2 + |v|^2 - |u - v|^2)$ for $u, v \in T_pM$, this approximates the scalar product of $E_p^{-1}(x)$ and $E_p^{-1}(y)$ in T_pM:

$$|P(x, y) - \langle E_p^{-1}(x), E_p^{-1}(y)\rangle| < 20\widehat{\delta}\widehat{r}. \tag{A.11}$$

Then, since $E_p^{-1}(X_p)$ is a $\widehat{\delta}/2$-net in $E_p^{-1}(V_p)$, and the latter contains the $\widehat{r}/6$-ball centered at the origin, we can find points $a_1, \dots, a_n \in Y_p$ such that the vectors $v_i := E_p^{-1}(a_i)$, $i = 1, \dots, n$, approximate an orthonormal basis of T_pM rescaled by the factor $\widehat{r}/6$:

$$|(\widehat{r}/6)^{-2}\langle v_i, v_j\rangle - \delta_{ij}| < C_1\widehat{\delta}/\widehat{r}, \qquad 1 \le i, j \le n, \tag{A.12}$$

where δ_{ij} is the Kronecker delta, and $C_1 = C_1(n) > 0$ is a suitable constant. (A straightforward modification of the algorithm from [24, section 2.4] can be used to find such points efficiently.)

The inequalities (A.12) imply that the linear map $L\colon T_pM \to \mathbb{R}^n$ defined by

$$L(v) = (\widehat{r}/6)^{-1}(\langle v, v_1\rangle, \dots, \langle v, v_n\rangle)$$

is $(C_2\widehat{\delta}/\widehat{r})$-close in the operator norm to a linear isometry from T_pM to $\mathbb{R}^n$ for some constant $C_2 = C_2(n) > 1$; see [24, Lemma 2.6]. Hence L distorts distances within the $\widehat{r}$-ball by at most $2C_2\widehat{\delta}$. We approximate $L \circ E_p^{-1}$ by a map $F\colon Y_p \cup Y_q \to \mathbb{R}^n$ defined by

$$F(x) = (\widehat{r}/6)^{-1}(P(x, a_1), \dots, P(x, a_n)), \qquad x \in Y_p \cup Y_q,$$

and compute $\widetilde{d}'(x, y) = |F(x) - F(y)|$ for all $x, y \in Y_p \cap Y_q$. By (A.11) we have $|F(x) - L(E_p^{-1}(x))| < 120\sqrt{n}\,\widehat{\delta}$ for all $x, y \in Y_p \cap Y_q$. Hence, by (A.3) and the above mentioned property of L,

$$|\widetilde{d}'(x, y) - d_M(x, y)| < C_5\widehat{\delta}, \tag{A.13}$$

where $C_5 = 2C_2 + 120\sqrt{n} + 1$.

The domain of the function $\widetilde{d}'$ defined by the above procedure includes all pairs $x, y \in \mathcal{Y}$ such that $d_M(x,y) < \widehat{r}/4$. Indeed, if $d_M(x,y) < \widehat{r}/4$ and $p, q \in X$ are such that $x \in Y_p$, $y \in Y_q$, then by the triangle inequality we have $d_M(p,q) < \widehat{r}/4 + 2\widehat{r}/6 < 2\widehat{r}/3 - 3\widehat{\delta}$, and hence, by (A.1), $\widetilde{d}(p,q) < 2\widehat{r}/3 - 2\widehat{\delta}$. Thus for any such pair x, y the value $\widetilde{d}'(x,y)$ is defined and satisfies (A.13).

To finish the construction, set $\widetilde{d}'(x,y) = \widehat{r}$ for all remaining pairs $x, y \in \mathcal{Y}$. Now the function $\widetilde{d}'$ is defined on $\mathcal{Y} \times \mathcal{Y}$, and it satisfies the assumptions [24, Corollary 1.10] for $\widehat{r}' = \widehat{r}/4$ in place of $\widehat{r}$, $\widehat{\delta}' = C_5\widehat{\delta}$ in place of $\widehat{\delta}$, and $K' = \widehat{\delta}'/(\widehat{r}')^3 = 2^6 C_5 K$ in place of K. Applying [24, Corollary 1.10] with these modified parameters finishes the proof of Proposition A.1. ∎

The proofs in Proposition A.1 and [24, Corollary 1.10] are constructive. For the reader's convenience, we sketch the main ideas of the construction of (M^*, g^*) in [24].

Assume that we are given a set $\mathcal{X} = \{x_j\}_{j=1}^N$ and the function $\widetilde{d}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \cup \{0\}$ such that there are points $\mathcal{Z} = \{z_j\}_{j=1}^N \subset M$ that are an $\widehat{r}/20$-dense subset of M, and

$$|\widetilde{d}(x_i, x_j) - d_M(z_i, z_j)| \le \widehat{\delta} \quad \text{if } d_M(z_i, z_j) < \widehat{r}, \tag{A.14}$$

$$\widetilde{d}(x_i, x_j) > \widehat{r} - \widehat{\delta} \quad \text{if } d_M(z_i, z_j) \ge \widehat{r}. \tag{A.15}$$

Then we can construct a manifold (M^*, g^*) by taking the following steps:

1. Choose a maximal subset $\mathcal{X}_0 = \{x_{j_1}, x_{j_2}, \dots, x_{j_L}\} \subset \mathcal{X}$ that satisfies $\widetilde{d}(x_{j_\ell}, x_{j_{\ell'}}) \ge r/100$ for $\ell \ne \ell'$. By renumbering the points, we can assume that $\mathcal{X}_0 = \{x_1, x_2, \dots, x_L\}$.

2. Find the $\widetilde{d}$-balls $B(x_\ell, \widehat{r}) = \{y \in \mathcal{X} : \ \widetilde{d}(y, x_\ell) < \widehat{r}\}$, $\ell = 1, \dots, L$ in $\mathcal{X}$ and choose disjoint Euclidean balls $D_\ell \subset \mathbb{R}^n$ of radius $\widehat{r}$, $\ell = 1, 2, \dots, L$. Then, construct embeddings $f_\ell : B(x_\ell, \widehat{r}) \to D_\ell$ such that $\big||f_\ell(y) - f_\ell(y')| - \widetilde{d}(y, y')\big| \le \widehat{\delta}$ for all $y, y' \in B(x_\ell, \widehat{r})$.

Then the balls D_ℓ can be considered as local coordinate charts. For $x \in B(x_{\ell_1}, \widehat{r}) \cap B(x_{\ell_2}, \widehat{r})$, the function $f_{\ell_2} \circ f_{\ell_1}^{-1}$ maps the points $f_{\ell_1}(B(x_{\ell_2}, \widehat{r}) \cap B(x_{\ell_2}, \widehat{r})) \subset D_{\ell_1}$ to D_{ℓ_1}. These functions are the discrete approximations of the transition functions between the coordinate charts.

3. Using the approximative transition functions $f_{\ell_2} \circ f_{\ell_1}^{-1}$, construct an embedding Φ from the union of the balls D_ℓ, $\ell = 1, \dots, L$ into a Euclidean space $\mathbb{R}^m$, $m = L(n+1)$, so that if $B(x_{\ell_1}, \widehat{r}) \cap B(x_{\ell_2}, \widehat{r}) \ne \emptyset$, then the images $\Phi(D_{\ell_1})$ and $\Phi(D_{\ell_2})$ are close. This embedding is similar to a classical Whitney embedding of a manifold into a Euclidean space.

4. It turns out that the images $\Phi(D_\ell)$, $\ell = 1, \dots, L$, are close to a nn-dimensional submanifold in $\mathbb{R}^m$. Using an interpolation algorithm for the point cloud $\bigcup_{\ell=1}^L \Phi(f_\ell(B(x_\ell, \widehat{r})))$, find an n-dimensional manifold $M^* \subset \mathbb{R}^m$ that is close to the set $\bigcup_{\ell=1}^L \Phi(D_\ell) \subset \mathbb{R}^m$. Let P_{M^*} be the projector in the Fermi coordinates around M^* that maps a point in a neighborhood of M^* to the closest point of M^*.

5. Push forward the Euclidean metric tensors g_ℓ on the balls D_ℓ to the metric tensors $(P_{M^*} \circ \Phi)_*(g_\ell)$ on M^* in the map $P_{M^*} \circ \Phi$. Then, construct g^* on M^* by computing a weighted average of the obtained tensors $(P_{M^*} \circ \Phi)_*(g_\ell)$ using a smooth partition of unity.

The details and a more algorithmic version of this construction are given in [24].

Appendix B. Proof of Lemma 2.1.

Proof. Recall that by (1.18) we have $\mathrm{vol}_g(M) \le V_0$. Also, by (1.19), the ν-volume of a metric ball $B_M(x, \delta_1/6) \subset M$ is bounded from below by $c_0(\delta_1/6)^n$.

Let $\{z_1, \dots, z_m\}$ be a maximal $(\delta_1/3)$-separated subset in M. Then $m \le m_0 = 6^n c_0^{-1} \delta_1^{-n} = C_5 \delta_1^{-n}$.

Let $V_k = \{y \in M : \ \mathrm{dist}(y, z_k) < \mathrm{dist}(y, z_j) \text{ for all } j \ne k\}$ be the open Voronoi sets corresponding to points z_k, and let W_k, $k = 1, 2, \dots, m$, be such disjoint sets that $V_k \subset W_k \subset \overline{V}_k$, and that the union of the sets W_k is M. Note that then the balls $B_M(z_j, \delta_1/6)$ are disjoint, and thus there is $C_6 = C_6(n, \Lambda, D, i_0, \rho_{max}, \rho_{min})$ so that $\nu(W_k) \ge c_0(\delta_1/6)^{-n} \ge 1/(C_6 m)$ for all $k = 1, 2, \dots, m$. Observe that if for all $k = 1, 2, \dots, m$ there is X_j, $j \le N_0$ such that $X_j \in W_k$, then the set $\{X_j : \ j = 1, 2, \dots, N_0\}$ is a δ_1-net in M.

We can use the classical collectors problem to estimate the probability of the event A_{m,N_0} to estimate that all sets W_k contain at least one point X_j, $j = 1, 2, \dots, N_0$. The tail estimates are used to give a solution for this problem, and for the reader's convenience we give the details of this below (see also [15, 16, 47] for related results).

Let us choose an infinite sequence of i.i.d. random variables $X_1, X_2, \dots$ having distribution ν on M, and let T be the smallest number such that all sets W_k, $k = 1, 2, \dots, m$ contain at least one point X_j, $j = 1, 2, \dots, T$. Let E_i^r denote the event that the ith set W_i does not contain any of the first r points $X_1, \dots, X_r$. Let $m_1 = C_6 m_0 \ge C_6 m$ and $b = 1 + \frac{\log(2(C_6\theta)^{-1})}{\log m_1}$. Then

$$\mathbb{P}\left[E_i^r\right] \le \left(1 - \frac{1}{C_6 m}\right)^r \le \left(1 - \frac{1}{m_1}\right)^r \le e^{-r/m_1}.$$

For $r = \lfloor b m_1 \log m_1 \rfloor + 1$, we have $\mathbb{P}\left[E_i^r\right] \le e^{(-bm_1 \log m_1)/m_1} = m_1^{-b}$. Then,

$$\mathbb{P}\left[T > b m_1 \log m_1\right] = \mathbb{P}\left[\bigcup_{i=1}^{m} E_i^r\right] \le \frac{m_1}{C_6} \cdot \mathbb{P}[E_1^r] \le \frac{1}{C_6} m_1^{-b+1} \le \frac{\theta}{2}.$$

Observe that $m_1 = C_6 C_5 \delta_1^{-n}$ and

$$\begin{aligned} b m_1 \log m_1 &= (1 + \frac{\log(2(C_6\theta)^{-1})}{\log(C_6 C_5 \delta_1^{-n})}) C_6 C_5 \delta_1^{-n} \log(C_6 C_5 \delta_1^{-n}) \\ &= (\log(C_6 C_5 \delta_1^{-n}) + \log(2(C_6\theta)^{-1})) C_6 C_5 \delta_1^{-n}. \end{aligned}$$

Therefore, when (2.5) is valid with a suitable C_3, we have $N_0 > b m_1 \log m_1$. Hence, $\mathbb{P}(T \le N_0) \ge 1 - \frac{\theta}{2}$. This proves Lemma 2.1. ∎

Acknowledgment. The authors would like to thank the reviewers for their useful suggestions.

REFERENCES

[1] E. Aamari and C. Levrard, *Nonasymptotic rates for manifold, tangent space and curvature estimation*, Ann. Statist., 47 (2019), pp. 177–204.

[2] M. Anderson, A. Katsuda, Y. Kurylev, M. Lassas, and M. Taylor, *Boundary regularity for the Ricci equation, geometric convergence, and Gel'fand's inverse boundary problem*, Invent. Math., 158 (2004), pp. 261–321.
[3] M. Belkin and P. Niyogi, *Laplacian eigenmaps and spectral techniques for embedding and clustering*, in Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, 2002, pp. 585–691.
[4] M. Belkin and P. Niyogi, *Towards a theoretical foundation for Laplacian-based manifold methods*, J. Comput. System Sci., 74 (2008), pp. 1289–1308.
[5] M. Bernstien, V. de Silva, J. Langford, and J. Tenenbaum, *Graph Approximations to Geodesics on Embedded Manifolds*, Technical Report, Stanford University, Stanford, CA, 2000.
[6] M. Brand, *Charting a manifold*, in Advances in Neural Information Processing Systems 15, MIT Press, Cambridge, 2002, pp. 985–992.
[7] D. Burago, Y. Burago, and S. Ivanov, *A Course in Metric Geometry*, Grad. Stud. Math. 33, AMS, Providence, RI, 2001.
[8] M. Carreira-Perpinan and Z. Lu, *Manifold Learning and Missing Data Recovery through Unsupervised Regression*, in Proceedings of the 11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, 2011.
[9] S. Cheng, T. Dey, and E. Ramos, *Manifold reconstruction from point samples*, in Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 2005, pp. 1018–1027.
[10] D. Chigirev and W. Bialek, *Optimal manifold representation of data: An information theoretic approach*, in Advances in Neural Information Processing Systems 16, S. Thrun et al., eds., MIT Press, Cambridge, 2004, pp. 164–168.
[11] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. of Nat. Acad. Sci., 102 (2005), pp. 7426–7431.
[12] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data Part* II: *Multiscale methods*, Proc. of Nat. Acad. Sci., 102 (2005), pp. 7432–7438.
[13] R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.
[14] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman & Hall, London, 1994.
[15] P. Erdos and A. Renyi, *On a classical problem of probability theory*, Magyar Tud. Akad. Mat., Magyar Tud. Akad. Mat. Kutató Int. Közl., 6 (1961), pp. 215–220.
[16] L. Flatto and D. Newman, *Random coverings*, Acta Math., 138 (1977), pp. 241–264.
[17] M. de Hoop, P. Kepley, and L. Oksanen, *On the construction of virtual interior point source travel time distances from the hyperbolic Neumann-to-Dirichlet map*, SIAM J. Appl. Math., 76 (2016), pp. 805–825, https://doi.org/10.1137/15M1033010.
[18] D. Donoho and C. Grimes, *When does geodesic distance recover the true hidden parametrization of families of articulated images?*, in Proceedings of ESANN 2002, Bruges, Belgium, 2002.
[19] D. Donoho and D. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data*, in Proceedings of the National Academy of Sciences 100, pp. 5591–5596.
[20] D. Donoho and C. Grimes, *Image manifolds which are isometric to Euclidean space*, J. Math. Imaging Vision, 23 (2005), pp. 5–24.
[21] Ch. Fefferman, *A sharp form of Whitney's extension theorem*, Ann. of Math., 161 (2005), pp. 509–577.
[22] Ch. Fefferman, *C^m-extension by linear operators*, Ann. of Math., 166 (2007), pp. 779–835.
[23] Ch. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan, *Fitting a putative manifold to noisy data*, in Proceedings of the 31st Annual Conference On Learning Theory, Vol. 75, 2018, pp. 688–720.
[24] C. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan, *Reconstruction and interpolation of manifolds* I: *The geometric Whitney problem*, Found. Comput. Math., (2019), https://doi.org/10.1007/s10208-019-09439-7.
[25] C. Fefferman, S. Ivanov, M. Lassas, and H. Narayanan, *Fitting a Manifold of Large Reach to Noisy Data*, preprint, https://arxiv.org/abs/1910.05084, 2019.
[26] Ch. Fefferman and B. Klartag, *Fitting C^m-smooth function to data* I, Ann. of Math., 169 (2009), pp. 315–346.

[27] Ch. Fefferman, S. Mitter, and H. Narayanan, *Testing the manifold hypothesis*, J. Amer. Math. Soc., 29 (2016), pp. 983–1049.
[28] C. Genovese, M. Christopher, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman, *Manifold estimation and singular deconvolution under Hausdorff loss*, Ann. Statist., 40 (2012), pp. 941–963.
[29] C. Genovese, M. Christopher, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman, *Minimax manifold estimation*, J. Mach. Learn. Res., 13 (2012), pp. 1263–1291.
[30] C. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman, *Nonparametric ridge estimation*, Ann. Statist., 42 (2014), pp. 1511–1545.
[31] A. Gilbert and R. Sonthalia, *Unrolling Swiss Cheese: Metric repair on manifolds with holes*, in Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing, 2018.
[32] A. Haddad, D. Kushnir, and R. Coifman, *Texture separation via a reference set*, Appl. Comput. Harmon. Anal., 36 (2014), pp. 335–347.
[33] M. Hein, J.-Y. Audibert, and U. von Luxbur, *From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians*, in Learning Theory, Lecture Notes in Comput. Sci. 3559, Springer, Berlin, 2005, pp. 470–485.
[34] M. Hein and M. Maier, *Manifold denoising*, in Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, 2007, pp. 561–568.
[35] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. , 58 (1963), pp. 13–30.
[36] P. Hoskins, *Principles of ultrasound elastography*, Ultrasound, 20 (2012), pp. 8–15.
[37] D. Joncas, M. Meila, and J. McQueen, *Improved graph Laplacian via geometric self-consistency*, in Proceedings of the 31st Annual Conference on Neural Information Processing Systems, 2017.
[38] P. Jones, M. Maggioni, and R. Schul, *Universal local parametrizations via heat kernels and eigenfunctions of the Laplacian*, Ann. Acad. Sci. Fenn. Math., 35 (2010), pp. 131–174.
[39] O. Kallenberg, *Foundations of Modern Probability Theory*, Springer-Verlag, New York, 2002.
[40] A. Katchalov, Y. Kurylev, and M. Lassas, *Inverse Boundary Spectral Problems*, Chapman & Hall/CRC Monogr. Surv. Pure Appl. Math. 123, Chapman & Hall/CRC, Boca Raton, FL, 2001.
[41] Y. Kurylev, L. Oksanen, and G. Paternain, *Inverse problems for the connection Laplacian*, J. Differential Geom., 110 (2018), pp. 457–494.
[42] D. Kushnir, A. Haddad, and R. Coifman, *Anisotropic diffusion on sub-manifolds with application to earth structure classification*, Appl. Comput. Harmon. Anal., 32 (2012), pp. 280–294.
[43] T. Lin and H. Zha, *Riemannian manifold learning*, IEEE Trans Pattern Anal Mach Intell., 5 (2008), pp. 796–809.
[44] T. Lin, H. Zha, and S. Lee, *Riemannian manifold learning for nonlinear dimensionality reduction*, in Computer Vision - ECCV 2006, ECCV 2006, Lecture Notes in Comput. Sci. 3951, A. Leonardis, H. Bischof, and A. Pinz, eds., Springer, Berlin, Heidelberg, pp. 44–55.
[45] J. Nash, *C^1-isometric imbeddings*, Ann. of Math., 60 (1954), pp. 383–396.
[46] J. Nash, *The imbedding problem for Riemannian manifolds*, Ann. of Math., 63 (1956), pp. 20–63.
[47] D. Newman and L. Shepp, *The double dixie cup problem*, Amer. Math. Monthly, 67 (1960), pp. 58–61.
[48] D. Perraul-Joncas and M. Meila, *Non-linear Dimensionality Reduction: Riemannian Metric Estimation and the Problem of Geometric Discovery*, preprint, https://arxiv.org/abs/1305.7255, 2013.
[49] J. McQueen, M. Meila, and D. Perrault-Joncas, *Nearly isometric embedding by relaxation*, in Advances in Neural Information Processing Systems 29, MIT Press, Cambridge, 2016, pp. 2639–2647.
[50] P. Petersen, *Riemannian Geometry*, Grad. Texts in Math. 171, Springer-Verlag, New York, 1998.
[51] X. Qu, T. Azuma, H. Nakamura, H. Imoto, S. Tamano, S. Takagi, S.-I. Umemura, I. Sakuma, and Y. Matsumoto, *Bent ray ultrasound tomography reconstruction using virtual receivers for reducing time cost*, in Proceedings of SPIE, Vol. 9419, Medical Imaging 2015: Ultrasonic Imaging and Tomography, 94190F.
[52] S. Roweis and L. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290 (2000), 2323.
[53] S. Roweis, L. Saul, and G. Hinton, *Global coordination of local linear models*, in Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, 2001, pp. 889–896.

[54] J. Shawe-Taylor and N. Christianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK, 2004.
[55] A. Singer, *From graph to manifold Laplacian: the convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134.
[56] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math., 125 (1987), pp. 153–169.
[57] P. Stefanov, G. Uhlmann, and A. Vasy, *Boundary rigidity with partial data*, J. Amer. Math. Soc., 29 (2016), pp. 299–332.
[58] J. Tenenbaum, V. de Silva, and J. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 290 (2000), pp. 2319–2323.
[59] N. Verma, *Distance preserving embeddings for general n-dimensional manifolds*, J. Mach. Learn. Res., 14 (2013), pp. 2415–2448.
[60] H. Zha and Z. Zhang, *Continuum isomap for manifold learnings*, Comput. Statist. Data Anal., 52 (2007), pp. 184–200.
[61] Z. Zhang and H. Zha, *Principal manifolds and nonlinear dimension reduction via local tangent space alignment*, SIAM J. Sci. Comput., 26 (2005), pp. 313–338, https://doi.org/10.1137/S1064827502419154.