

Reconstruction of Manifold Distances from Noisy Observations

Charles Fefferman* Jonathan Marty† Kevin Ren*

Dedicated to the memory of Yaroslav Kurylev

Abstract

We consider the problem of reconstructing the intrinsic geometry of a manifold from noisy pairwise distance observations. Specifically, let $\mathcal{M}$ denote a diameter 1 d -dimensional manifold and μ a probability measure on $\mathcal{M}$ that is mutually absolutely continuous with the volume measure. Suppose $X_1, \dots, X_N$ are i.i.d. samples of μ and we observe noisy-distance random variables $d'(X_j, X_k)$ that are related to the true geodesic distances $d(X_j, X_k)$. With mild assumptions on the distributions and independence of the noisy distances, we develop a new framework for recovering all distances between points in a sufficiently dense subsample of M . Our framework improves on previous work which assumed i.i.d. additive noise with known moments. Our method is based on a new way to estimate L_2 -norms of certain expectation-functions $f_x(y) = \mathbb{E}d'(x, y)$ and use them to build robust clusters centered at points of our sample.

Using a new geometric argument, we establish that, under mild geometric assumptions—bounded curvature and positive injectivity radius—these clusters allow one to recover the true distances between points in the sample up to an additive error of $O(\varepsilon \log \varepsilon^{-1})$. We develop two distinct algorithms for producing these clusters. The first achieves a sample complexity $N \asymp \varepsilon^{-2d-2} \log(1/\varepsilon)$ and runtime $o(N^3)$. The second introduces novel geometric ideas that warrant further investigation. In the presence of missing observations, we show that a quantitative lower bound on sampling probabilities suffices to modify the cluster construction in the first algorithm and extend all recovery guarantees. Our main technical result also elucidates which properties of a manifold are necessary for the distance recovery, which suggests further extension of our techniques to a broader class of metric probability spaces.

*Department of Mathematics, Princeton University, Princeton, NJ 08544 USA. {cf, kr5621}@math.princeton.edu

†Program in Applied and Computational Mathematics (PACM), Princeton University, Princeton, NJ 08544, USA. jm2866@math.princeton.edu

Contents

1	Introduction	3
1.1	Main Result	4
1.2	Intuition	5
1.3	Potential Applications	6
2	Preliminaries	7
2.1	Probability	7
2.2	Manifold	7
3	Algorithm 1: Norm estimation via inner products	9
3.1	Identifying Distances in Metric Spaces	9
3.2	Proof of Theorem 3.1, non-missing case	10
3.2.1	Estimating norms from inner products	10
3.2.2	Cluster Construction	12
3.3	In the Case of Missing Noise	14
3.4	Finding distances from clusters	16
3.5	Proof of Theorem 1.1	21
4	Algorithm 2: Regularized Optimization	22
4.1	Accessing Objective Functions	24
4.2	Obtaining a Cluster Center	25
4.3	Cluster Structure Constants	27
4.4	Separability and Closeness Constants	28
4.5	Relationship Between Constants	30
4.6	Cluster Efficacy	31
4.7	Cluster-Cluster Distances	33
4.8	Proof of Theorem 4.1	34
5	Conclusion and Discussion	36
5.1	Discussion of Theorem 1.1 and comparison with [13]	37
5.2	Discussion of Theorem 4.1	39
5.3	Recovering the Manifold Scale	39
5.4	More Open Questions	40
6	Omitted Proofs	40

1 Introduction

Let $(\mathcal{M}, \mu, d)$ be geodesic probability space. Suppose that we are given noisy versions of the distances $d'(x_j, x_k)$ for a set of points $\{x_i\}_{i \in [N]}$, which are i.i.d. samples of μ . Assuming modest constraints on the distribution of the $d'(x_j, x_k)$, this paper asks whether it is possible to approximate the pairwise geodesic distances $d(x_j, x_k)$ for a smaller subset of the x_i which are sufficiently dense in $\mathcal{M}$. If $(\mathcal{M}, d)$ denotes a Riemannian manifold with geodesic distance, one can use these approximate distances to construct an approximation of $\mathcal{M}$ in the Gromov-Hausdorff sense, see [11].

There have been several previous approaches to learning abstract Riemannian manifolds from geodesic distance data. The Diffusion Maps [7, 8, 6] and EigenMap [2] algorithms build weighted graphs based on the distance information. These weights define a diffusion or differential operator on the manifold. By approximating the associated eigenfunctions, each point can be non-isometrically embedded into $\mathbb{R}^m$. Further, the embeddings given by both approaches are used to approximate the metric in [19]. In [21] it is assumed that the manifold is covered by a single coordinate chart and we can access the coordinates for all sampled points as well as "similarity measures", which depend on the distances. The relationship between the underlying metric and the observed similarity measures is then modeled. The approximation of a Riemannian manifold in a Lipschitz sense based on access to pairwise distances between points in a dense net with small deterministic errors was treated in [11]. [12] addresses doing so when the distances are between a dense net of an open subset of the manifold with certain properties and a dense net of the whole manifold. In [13], the approach of [11] was extended to noisy distances of the form $d(x_j, x_k) + \eta_{jk}$, where the η_{jk} are i.i.d. and we have access to a few of their moments. [13] also accounts for the case where a subset of the noisy distances are missing.

In this paper we develop methods for recovering the geodesic distances between a sufficiently dense set of points on $\mathcal{M}$ given access to only noisy distances. Our sample size (number of draws from μ) and runtime are comparable to [13] while significantly loosening assumptions on the noisy distances, foregoing much of the required independence and knowledge of moments (for a more detailed discussion of how our approach relates to that of [13], see Section 5). Further, we are able to extend past the case of Riemannian manifolds to finite diameter geodesic probability spaces which satisfy key properties. These properties are characteristic of manifolds with curvature bounds, but can hold for more general spaces. Our methods take advantage of how certain statistics of the $d'(x, y)$ change as we vary the choice of (x, y) . It is possible to access these statistics by taking inner products of key functions against μ and these inner products can in turn be approximated by taking averages of the noisy distances. Our main result is stated in Section 1.1. Intuition for our approach is given in Section 1.2. Our methods have potential applications to a diverse set of inverse problems, which are discussed in Section 1.3.

1.1 Main Result

We propose two slightly different algorithms for manifold recovery, which essentially generalize [13, Theorem 1]. See Section 5 for a detailed explanation of the extend to which Theorem 1.1 is a generalization. The result statement for both algorithms are very similar, and so Algorithm 2's statement can be found at Theorem 4.1.

Theorem 1.1. (Algorithm 1) Fix constants $d \geq 2$, $\rho, C_1, C_2, C_3 \geq 1$, $0 < i_0, r_0 \leq 1$, $\Lambda \geq 0$, $\varepsilon > 0$.

Let $(\mathcal{M}, \mu, d)$ be a d -dimensional Riemannian manifold with diameter 1, $|\text{Sec}_{\mathcal{M}}| \leq \Lambda^2$, and injectivity radius $\geq i_0$. Let μ be a probability measure on $\mathcal{M}$ which is ρ -mutually absolutely continuous with the uniform volume measure, with Radon-Nikodym derivative bounded in $[\rho, 1/\rho]$. We draw N points X μ -uniformly at random from $\mathcal{M}$. For $x, y \in X$, let $p(x, y) \in [0, 1]$ denote the probability that observation $d'(x, y)$ is present, and denote $m(x, y)$ to equal 1 if observation $d'(x, y)$ is present, and 0 otherwise. Suppose we have a noisy distance function $d'(x, y)$ defined for all $x \neq y \in X$ satisfying the following properties:

- (Symmetry) $d'(x, y) = d'(y, x)$;
- (Independence) $\{d'(x, y)\}_{\{x, y\} \subset X}$ is a set of mutually independent random variables.
- (Subgaussian) $d'(x, y)$ is sub-Gaussian with Orlicz norm $\|d'(x, y)\|_{\psi_2} \leq C_1$;

and furthermore whose expectation $f(x, y) = \mathbb{E}d'(x, y)$ satisfies the following properties:

- (Bounded expectation) $|f(x, y)| \leq C_2$ for all $x, y \in X$;
- (Bi-lipschitz) If $x, y, z \in X$ with $d(x, y) \geq d(x, z)$, then

$$\frac{1}{C_3}(d(x, y) - d(x, z)) \leq f(x, y) - f(x, z) \leq C_3 d(y, z).$$

No missing data: If $p(x, y) = 1$, then with probability $\geq 1 - \theta$, we can recover all distances $d(x, y)$ up to additive error $O(\varepsilon \log \varepsilon^{-1})$ and learn a manifold representation with error $O((\varepsilon \log \varepsilon^{-1})^{2/3})$, and furthermore we can take $N = \varepsilon^{-2d-2} \log \varepsilon^{-1}$, with runtime $o(N^3)$.

Missing data: If p satisfies the following quantitative “robustly nonzero” properties:

1. $p(x, y) \geq \phi$ for all x, y with $d(x, y) \leq r_0$;
2. for any distinct $x, y, z, w \in X$, we have $p(z, w) > \lambda_1 p(x, y)$ for all $d(z, w) < d(x, y) + \lambda_2 p(x, y)$;

and if $\varepsilon < c(i_0, \Lambda, d)\lambda_2\phi r^d$ then with probability $\geq 1 - \theta$, we can recover all distances $d(x, y)$ up to additive error $O(\frac{\varepsilon}{r_0^2} \log \varepsilon^{-1})$ and learn a manifold representation with error $O((\frac{\varepsilon}{r_0^2} \log \varepsilon^{-1})^{2/3})$, and furthermore we can take $N = \varepsilon^{-2d-2} \log \varepsilon^{-1}$, with runtime $o(N^{4.5})$.

Remark 1.2. The theorem statement is not the most general possible. See Section 5 for some examples of further extensions of Theorem 1.1.

1.2 Intuition

Our methods rely on constructing clusters of points which are close together. From there, we can approximate the average noisy distances between given points and all points in a given cluster and use these estimates to model the true distances between a sufficiently dense set of points on the manifold. However, a key conceptual impediment is that we make no strong assumptions on the distribution of $d'(x_j, x_k)$ for any pair of points. By contrast, in [13] it is assumed that $d'(x_j, x_k) = d(x_j, x_k) + \eta_{jk}$ where the η_{jk} are i.i.d. and we know their mean and variance, in which case it's clear how our point-cluster distances would relate to the true distances between, say, a point in the cluster and our given point. In our case the meaning of this average is unclear and, moreover, it's even less clear how one would go about constructing a cluster.

Consider $f(x, y) = \mathbb{E}d'(x, y)$ and let $f_x(y) = f(x, y)$. If we assume that the $d'(x, y) \sim K_{d(x, y)}$ and the $r \mapsto \mathbb{E}K_r$ is a bilipschitz function, then the $f_x : \mathcal{M} \rightarrow \mathbb{R}$ are bilipschitz functions w.r.t. the intrinsic metric. Observe that, on account of $\mathcal{M}$ having finite diameter, the f_x are bounded functions. This boundedness and the finite volume of $\mathcal{M}$ imply that $f_x \in L^2(\mathcal{M}, \mu)$ where μ denotes the uniform probability distribution. This L^2 structure is the key ingredient in our algorithm.

To get a sense of how we can use this L^2 structure to overcome the challenges mentioned in the first paragraph of this section, observe that we can approximate the inner product using our data. Take $x, y \in \mathcal{M}$ and sample a set $Z = \{z_i\}$ with the $z_i \sim \mu$ i.i.d. For a pair $x, y \in \mathcal{M}$ with $x \neq y$, consider the random variable

$$I(x, y) = \frac{1}{|Z|} \sum_{z \in Z} d'(x, z) d'(z, y)$$

and note that $\mathbb{E}I(x, y) = \langle f_x, f_y \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the L^2 inner product. Hence by sampling enough points we can approximate $\langle f_x, f_y \rangle$ for as many pairs (x, y) as we want to whatever degree of accuracy desired. What can we do with this information?

Consider the inner product $\langle f_x, f_y \rangle = \frac{1}{2}(\|f_x\|_2^2 + \|f_y\|_2^2 - \|f_x - f_y\|_2^2)$ for $x, y \in \mathcal{M}$. Note that the $\|f_x - f_y\|_2^2$ acts as a proxy for $d(x^*, a)$. To see this, observe that taking $d(x, a) > 3\delta$ means that on a δ -ball around x , $|f_x - f_a|$ is above some constant. Lower bounding the volume of this ball, which can be done via a sectional curvature upper bound, gives a lower bound on $\|f_x - f_a\|_2^2$. This follows from the lower Lipschitz constant of $r \mapsto \mathbb{E}K_r$. Alternatively, for any given $c > 0$, setting sufficiently $d(x, a)$ small enough means that $|f_x - f_a| < c$ everywhere on the manifold (follows from the upper Lipschitz constant of $r \mapsto \mathbb{E}K_r$). Thus, $\|f_x - f_y\|_2^2$ acts as a proximity sensor.

In the case that $\mathcal{M}$ is a symmetric space we can take direct advantage of this property, since all the f_x have the same L^2 norm, which we term S . As such, for any pair $x, y \in \mathcal{M}$, $\langle f_x, f_y \rangle = \frac{1}{2}(\|f_x\|_2^2 + \|f_y\|_2^2 - \|f_x - f_y\|_2^2) = S - \frac{1}{2}\|f_x - f_y\|_2^2$. Thus, the function $y \rightarrow \langle x, y \rangle$ acts as a proximity sensor and we can use it to build clusters around any point $x \in \mathcal{M}$.

However, in the general case we need to deal with the changing values of the $\|f_x\|_2^2$. We can overcome this difficulty through optimization. Let $X^* = \arg \max_{x \in \mathcal{M}} \|f_x\|_2^2$. We can find a member

$x^* \in X^*$ by taking

$$(x^*, y^*) = \arg \max_{x, y \in \mathcal{M}, x \neq y} \langle f_x, f_y \rangle$$

Arguing based on the optimality of x^* and Lipschitz-ness of $x \rightarrow \|f_x\|_2^2$, it can be shown that the $\langle f_{x^*}, f_y \rangle$ can be used to distinguish points that are close to x^* from those that are far. Hence, we can use $y \mapsto \langle f_{x^*}, f_y \rangle$ and functions like it to build clusters, in this case around x^* .

This argmax idea is directly implemented in Algorithm 2 (Theorem 4.1). In fact, a slight variation of the argmax trick allows us to implement a simpler Algorithm 1 (Theorem 1.1).

As for relating the noisy point-cluster distances to true ones, having built these clusters the bilipschitz-ness of $r \mapsto \mathbb{E}K_r$ can be utilized to approximate the original distances. This is formalized in Section 3.4.

1.3 Potential Applications

The ability to reconstruct the intrinsic metric based on noisy pairwise distances is relevant to a broad class of inverse problems which attempt to recover material properties of some body based on observed or experimental data. For example, suppose I have two points in some inhomogenous medium. The time that it takes to traverse between these can be modelled by an underlying *travel-time metric* of a manifold (the medium). Hence, recovering the wave-speed function within the medium allows for the reconstruction of the metric. Recovering the travel-time metric accurately has applications in both medical and seismic imaging contexts [22, 23, 14]. Accurate reconstruction of the travel-time metric for seismic waves has the potential to unveil nuances in the Earth's structure. For example, measurements of these travel times has shown that seismic waves within the inner core travel faster along the Earth's spin-axis than in other directions [9]. Another application is magnetic resonance imaging based elastography, which measures mechanical properties of tissue based on its response to acoustic waves [14]. For a more detailed discussion of these topics, see Section 5 of [13].

Existing methods to solve these inverse problems rely on measurements taken at the boundary to infer the intrinsic distances or travel times between points in the manifold [23, 22, 15]. As an example of such boundary measurements, one might probe tissue using an ultrasound machine, moving the transducer to get readings at different angles [22]. In addition to algorithmic progress, there is a line of analysis which studies the connection between boundary measurements of a wave equation and the distances between points in the interior [1, 10, 16, 17]. However, random errors in the boundary measurements can propagate to the estimates of pairwise distances, reducing the efficacy of these approaches. In the context of ultrasound imaging, diffuse reverberation, in which the emitted sound wave reflects multiple times before returning to the transducer, is known to reduce the efficacy of sound speed estimation and Doppler imaging, among other techniques [4, 15, 20]. Thus, there is a need to develop methods which can either denoise the input data for these methods or reconstruct the Riemannian metric from their corrupt outputted distances. The techniques in this paper are relevant to the latter problem.

Acknowledgments. C.F. and J.M. are partially supported by NSF Grant DMS-1608782 and AFOSR Grant FA9550-12-1-0425. K.R. is supported by a NSF GRFP fellowship, a Simons Foundation Dissertation Fellowship in Mathematics, and a Cubist/Point72 PhD Fellowship. We thank Matti Lassas, Hariharan Narayanan, Hau Tieng Wu, and Aalok Gangopadhyay for helpful discussions.

2 Preliminaries

2.1 Probability

Definition 2.1 (Orlicz norm). If X is a sub-Gaussian random variable, we define its Orlicz norm $\|X\|_{\psi_2}$ by

$$\|X\|_{\psi_2} := \inf\{c \geq 0 : \mathbb{E}\left(e^{(X-\mathbb{E}X)^2/c^2}\right) \leq 2\}.$$

Theorem 2.2 (Sub-Gaussian Hoeffding Inequality). [24, Theorem 2.6.2] Let $X_1, \dots, X_n$ be independent sub-Gaussian random variables with means $\mathbb{E}[X_i]$ and Orlicz norms $\|X_i\|_{\psi_2}$. Then for all $t \geq 0$,

$$\Pr\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2 \exp\left(-\frac{ct^2}{2 \sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right),$$

where c is a universal constant.

Theorem 2.3 (Hoeffding Inequality). [24, Theorem 2.2.6] Let $X_1, \dots, X_n$ be independent random variables such that $|X_i| \leq C$ almost surely. Then for all $t \geq 0$,

$$\Pr\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{8nC^2}\right).$$

Theorem 2.4 (Bernstein). [3] Let $X_1, \dots, X_n$ be independent random variables such that $|X_i - \mathbb{E}X_i| \leq C$ almost surely. Then for all $t \geq 0$,

$$\Pr\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{i=1}^n \text{Var}(X_i) + \frac{2}{3}Ct}\right).$$

2.2 Manifold

For completeness, we include proof sketches in Section 6.

Proposition 2.5. Let $\mathcal{M}$ be a measure metric space and μ be a probability measure on $\mathcal{M}$. For $r > 0$, let $\text{VR}(r)$ be the infimal possible measure of an r -ball on $\mathcal{M}$. Suppose we sample N many points X i.i.d from μ , and let Y be the first N_0 many points of X . With probability $\geq 1 - (Ne^{-\frac{1}{8}\text{VR}(\delta)N} + \frac{1}{\text{VR}(\varepsilon/4)}e^{-\frac{1}{8}\text{VR}(\frac{\varepsilon}{2})|Y|})$, we can guarantee:

- every ball $B(x, \delta)$ with $x \in X$ contains at least $\frac{1}{2}\text{VR}(\delta)N$ many elements of X ;

- the set Y is ε -dense in $\mathcal{M}$.

Furthermore, for any measurable set Λ , we have

$$\Pr[|\Lambda \cap Y| \leq \frac{1}{2}\mu(\Lambda)|Y|] \leq e^{-\frac{1}{8}\mu(\Lambda)|Y|}.$$

Proposition 2.6. *For $d \geq 2$, let $\mathcal{M}$ be a d -dimensional Riemannian manifold with diameter 1, $|\text{Sec}_{\mathcal{M}}| \leq \Lambda^2$, and injectivity radius $\geq i_0$. Let μ be a probability measure on M which is ρ -mutually absolutely continuous with the manifold volume measure (i.e. the Radon-Nikodym derivative is in $[\rho, 1/\rho]$). Then:*

- (a) $\mu(B(x, r)) \geq c(\Lambda, i_0, d)\rho r^d$ for all $r > 0$;
- (b) for all $x, y \in \mathcal{M}$, $\mu(\{d(x, z) \geq d(y, z) + \frac{d(x, y)}{4}\}) \geq c(\Lambda, i_0, d)\rho i_0^d$;
- (c) for all $u \leq \frac{d(x, y)}{8}$, there exists $\kappa > 0$ depending only on Λ, i_0, d, ρ, u such that for all $x, y \in \mathcal{M}$, $\mu(\{d(x, z) - \frac{d(x, y)}{2} \in [\frac{u}{2}, \frac{3u}{2}], d(y, z) - \frac{d(x, y)}{2} \in [-\frac{3u}{2}, -\frac{u}{2}]\}) \geq c(\Lambda, i_0, d)\rho u^d$.

Definition 2.7 (Manifold Recovery). We say that we can learn a manifold representation of a manifold (M, g) with error (C, δ) if we can construct a manifold (M^*, g^*) such that

1. There is a diffeomorphism $F : M^* \rightarrow M$ satisfying

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

where $L = 1 + C\delta$, that is, the Lipschitz distance of the metric spaces (M^*, g^*) and (M, g) satisfies

$$d_{\text{Lip}}((M^*, g^*), (M, g)) \leq \log L.$$

2. The sectional curvature Sec_{M^*} of M^* satisfies

$$|\text{Sec}_{M^*}| \leq C.$$

3. The injectivity radius $\text{inj}(M^*)$ of M^* satisfies

$$\text{inj}(M^*) \geq \min\{C^{-1}, (1 - C\delta) \text{inj}(M)\}.$$

Proposition 2.8. *[13, Proposition A.1] For $K > 0$ and $d \geq 2$, the following holds true for sufficiently small $\delta > 0$. Let X be a $\frac{1}{20}(\frac{\delta}{K})^{1/3}$ -dense set on a d -dimensional manifold $\mathcal{M}$ with $|\text{Sec}_{\mathcal{M}}| \leq K$, and injectivity radius $\geq 2(\frac{\delta}{K})^{1/3}$. If we can recover all distances $d(x, y)$ up to additive error δ , then we can learn a manifold representation of $\mathcal{M}$ with error $(C_d K, (\frac{\delta}{K})^{2/3})$, where C_d is a constant depending only on d .*

3 Algorithm 1: Norm estimation via inner products

We give a brief sketch for Algorithm 1.

1. Find a cluster $\mathcal{C}(x)$ around each $x \in Y$ such that

$$B(x, \varepsilon') \subseteq \mathcal{C}(x) \subseteq B(x, \varepsilon),$$

for some $\varepsilon' < c\varepsilon$ to be chosen later.

2. Compute “distance proxy” numbers $A(x, y)$ such that for all $x, y, z \in Y$ and $d(x, y) \geq d(x, z) + 17\varepsilon$, then $A(x, y) > A(x, z)$.
3. Recover all $d(x, y)$ with $x, y \in Y$ up to $O(\varepsilon \log \varepsilon^{-1})$.

Step 1 will be motivated in Section 3.2.1 and proved in Section 3.2.2. Step 2 will also be proved in Section 3.2.2, and Step 3 will be completed in Section 3.4.

3.1 Identifying Distances in Metric Spaces

We prove the following theorem, which is about an abstract metric space satisfying certain properties that random samples of manifolds have (thanks to Theorem 2.6 and Theorem 2.5).

Theorem 3.1. *Let (X, d) be a metric space with $|X| = N$ and Y denote the first N_0 elements of X . For $x, y \in X$, let $p(x, y) \in [0, 1]$ denote the probability that observation $d'(x, y)$ is present, and denote $m(x, y)$ to equal 1 if observation $d'(x, y)$ is present, and 0 otherwise. Suppose that X satisfies the following a priori properties.*

- every ball $B(x, \delta)$ with $x \in Y$ contains at least αN many elements of X ;
- if $x, y \in X$ and $d(x, y) \leq r_0$, then $p(x, y) \geq \phi$;
- if $x, y \in X$ are distinct and satisfy $d(x, y) \leq \delta$, then $\frac{1}{|Y|} \sum_{z \in Y} p(x, z)p(y, z) > 2c_1$;
- if $d(x, y) > 4\varepsilon$, $d(x, v) \leq \delta$, $d(y, w) \leq \delta$, and $\frac{1}{|Y|} \sum_{z \in Y} p(x, z)p(y, z) > c_1$, the set $\Lambda_{x, y, v, w} := \{z : d(x, z) \geq d(y, z) + \varepsilon, p(x, z)p(y, z)p(v, z)p(w, z) > c_2\}$ satisfies $|\Lambda_{x, y, v, w} \cap Y| \geq \kappa|Y|$ for each distinct $x, y, v, w \in X$;

Suppose we have a noisy distance function $d'(x, y)$ satisfying the following properties:

- (Identity) $d'(x, x) = 0$;
- (Symmetry) $d'(x, y) = d'(y, x)$;
- (Subgaussian) $d'(x, y)$ is sub-Gaussian with Orlicz norm $\|d'(x, y)\|_{\psi_2} \leq C_1$;

- (“Independence”) $\{d'(x, y)\}_{\{x, y\} \subset X}$ is a set of random variables such that for every $x \in X$, the variables $\{d'(x, y)\}_{y \in X}$ are independent conditioned on any subset of the other variables $\{d'(y, z)\}_{y, z \neq x}$.

and furthermore whose expectation $f(x, y) = \mathbb{E}d'(x, y)$ satisfies the following properties:

- (Bounded expectation) $|f(x, y)| \leq C_2$ for all $x, y \in X$;
- (Lower bound) $f(x, y) - f(x, z) \geq \sigma$ for all distinct $x, y, z \in X$ with $d(x, y) \geq d(x, z) + \varepsilon$.
- (Uniform continuity) If $y, z \in X$ satisfy $d(y, z) \leq \delta$, then $|f(x, y) - f(x, z)| \leq \frac{\kappa\sigma^2}{8C_2}$ for all $x \neq y, z$.

Suppose that $N_0 \geq C \frac{C_1^4 + C_2^4}{\kappa^2 \sigma^4 C_2^2} \log \frac{N}{\theta} + \frac{C}{C_1^2} \log \frac{N}{\theta}$ and $\alpha N \geq \frac{CC_1^2}{\phi\sigma^2} \log \frac{N_0}{\theta}$, where $C > 0$ is an absolute constant. Then with probability $\geq 1 - \theta$, we can compute numbers $A(x, y) \in \mathbb{R} \cup \{+\infty\}$ in time $N^3 N_0^2$ such that:

1. for every $x, y \in Y$, if $d(x, y) \leq r_0 - 8\varepsilon$, then $A(x, y) \neq \infty$;
2. for every $x, y, z \in Y$ satisfying $d(x, y) \geq d(x, z) + 17\varepsilon$ and $A(x, y), A(x, z) \neq +\infty$, we have $A(x, y) > A(x, z)$.

If all $p(x, y) = 1$ (the non-missing case), we can in fact compute numbers $A(x, y) \in \mathbb{R}$ such that for all $x, y, z \in Y$ and $d(x, y) \geq d(x, z) + 17\varepsilon$, then $A(x, y) > A(x, z)$ in time $N^2 N_0$. We also obtain slightly relaxed guarantees $N_0 \geq C \frac{C_1^4 + C_1^2 C_2^2}{\kappa^2 \sigma^4} \log \frac{N}{\theta}$ and $\alpha N \geq \frac{CC_1^2}{\sigma^2} \log \frac{N_0}{\theta}$.

Remark 3.2. Think of $\sigma \sim \varepsilon$. It turns out that combining with Theorem 2.5, we need to take $N_0 \sim \varepsilon^{-\max\{d, 4\}} (d \log \frac{1}{\varepsilon} + \log \frac{1}{\theta})$, $\alpha \sim \varepsilon^{-2d}$, and $N \sim \varepsilon^{-2d-2} (\log \varepsilon^{-1} + \log \theta^{-1})$.

3.2 Proof of Theorem 3.1, non-missing case

3.2.1 Estimating norms from inner products

Definition 3.3. Fix a subset $Y \subseteq X$. For functions $f, g : Y \rightarrow \mathbb{R}$, we define

$$\langle f, g \rangle = \frac{1}{|Y|} \sum_{y \in Y} f(y)g(y).$$

The motivation for considering inner products is that they can be estimated effectively via the following concentration statement for inner products. The statement and proof technique may be of independent interest. The statement is not the most general possible in order to give a clean exposition; the reader is encouraged to relax the constraints of the proposition. The proof is deferred to Section 6.

Proposition 3.4 (Sub-Gaussian Hoeffding Inequality for Inner Products). *Let $X_1, \dots, X_n, Y_1, \dots, Y_n$ be sub-Gaussian random variables with means $\mathbb{E}[X_i], \mathbb{E}[Y_i]$ and Orlicz norms $\|X_i\|_{\psi_2}, \|Y_i\|_{\psi_2} \leq K$. Suppose that:*

- $Y_1, \dots, Y_n$ are mutually independent;
- $X_1, \dots, X_n$ are mutually independent conditioned on $Y_1, \dots, Y_n$;
- $\mathbb{E}[X_i | Y_1, \dots, Y_n] = \mathbb{E}[X_i]$, $|\mathbb{E}[X_i]| \leq L$, and $|\mathbb{E}[Y_i]| \leq L$ for all $1 \leq i \leq n$.

(These assumptions hold if $X_1, \dots, X_n, Y_1, \dots, Y_n$ are all mutually independent.) Then for all $t \geq 0$,

$$\Pr\left(\left|\sum_{i=1}^n (X_i Y_i - \mathbb{E}[X_i] \mathbb{E}[Y_i])\right| \geq t\right) \leq 5 \exp\left(-\frac{ct^2}{16K^2((K^2 + L^2)n + t)}\right),$$

where c is a universal constant.

Remark 3.5. If $X_1, \dots, X_n, Y_1, \dots, Y_n$, then each $X_i Y_i$ is exponential, hence we can use the sub-exponential Bernstein inequality. However, our approach works for our weaker mutual dependencies, at the cost of worse constants.

In our context, we have a noisy collection of vectors $\{f_x\}_{x \in X}$ defined by $f_x(y) = d'(x, y)$ and by Theorem 3.4 we can efficiently estimate inner products $\langle f_x, f_y \rangle$ with $x \neq y$; unfortunately, this is not enough to accurately determine norms $\|f_x - f_y\|^2$, which is the target quantity for determining which pairs (f_x, f_y) are close. Instead, we resort to estimating the norm via the following lemma.

Lemma 3.6 (Separation Lemma). *Fix $L, \delta > 0$, and suppose that for every $x \in X$, we have a function $f_x : Y \rightarrow \mathbb{R}$. Suppose that:*

- (Bounded) Every $\|f_x\|_2 \leq L$;
- (Dense) For any $x \in X$, there exists $y \neq x \in X$ with $\|f_x - f_y\|_2 \leq \delta$.

Then for any $x \neq y$,

$$\frac{1}{2}\|f_x - f_y\|_2(\|f_x - f_y\|_2 - 2\delta) \leq \sup_{z \in X, z \neq x, y} |\langle f_x - f_y, f_z \rangle| \leq L\|f_x - f_y\|_2.$$

Proof. We first look at the left inequality. If $\|f_x - f_y\|_2 \leq 2\delta$ then it is trivial, so assume $\|f_x - f_y\|_2 > 2\delta$. Note that one of $|\langle f_x, f_x - f_y \rangle| \geq \frac{1}{2}\|f_x - f_y\|^2$ or $|\langle f_y, f_x - f_y \rangle| \geq \frac{1}{2}\|f_x - f_y\|^2$; otherwise,

$$\|f_x - f_y\|^2 = \langle f_x, f_x - f_y \rangle - \langle f_y, f_x - f_y \rangle < \frac{1}{2}\|f_x - f_y\|^2 + \frac{1}{2}\|f_x - f_y\|^2 = \|f_x - f_y\|^2.$$

Without loss of generality, assume $|\langle f_x, f_x - f_y \rangle| \geq \frac{1}{2}\|f_x - f_y\|^2$. Using the dense assumption, pick $z \neq x, y$ such that $\|f_x - f_z\|_2 \leq \delta$. Then by Cauchy-Schwarz,

$$|\langle f_x - f_y, f_z \rangle| \geq |\langle f_x - f_y, f_x \rangle| - |\langle f_x - f_y, f_x - f_z \rangle| \geq \frac{1}{2}\|f_x - f_y\|_2^2 - \delta\|f_x - f_y\|_2.$$

This proves the first claim.

For the right inequality, we use Cauchy-Schwarz and our boundedness assumption to get, for all $z, w \in X$,

$$|\langle f_x - f_y, f_z \rangle| \leq \|f_x - f_y\|_2 \|f_z\|_2 \leq L \|f_x - f_y\|_2. \quad \square$$

Remark 3.7. If our bounded assumption is $\|f_x\|_\infty \leq L$, then our result can be improved to $\leq L \|f_x - f_y\|_1$.

3.2.2 Cluster Construction

Now, we begin the proof of Theorem 3.1 in the non-missing data case. For convenience, define $d'(x, x) = 0$ and $f(x, x) = 0$ for all $x \in X$.

Definition 3.8. Fix a subset $Y \subset X$. We define

$$\mathcal{C}(x) = \{y \in X : \sup_{z \in X, z \neq x, y} \left| \frac{1}{|Y|} \sum_{v \in Y} (d'(x, v) - d'(y, v)) d'(z, v) \right| \leq \frac{1}{5} \kappa \sigma^2\}.$$

Note that $x \in \mathcal{C}(x)$.

Proposition 3.9. *With probability at least $1 - 2|X|^2 \exp(-|Y| \delta_2^4 / (25600 C_1^2 (C_1^2 + C_2^2 + 1)))$, we have that for all $x \in X$,*

$$B(x, \delta) \subset \mathcal{C}(x) \subset B(x, 4\epsilon). \quad (2)$$

After a precomputation cost of $|X|^2 |Y|$, each cluster $\mathcal{C}(x)$ can be computed in $|X|^2$ time.

Proof. If we precompute every $\sum_{v \in Y} d'(x, v) d'(y, v)$ for each pair $x, y \in X$, each computation of $\frac{1}{|Y|} \sum_{v \in Y} (d'(x, v) - d'(y, v)) (d'(z, v) - d'(w, v))$ can be done in $O(1)$ time. For each $y \in X$, it takes $O(|X|)$ time to compute the supremum. Now we turn to proving (2).

Note that by independence,

$$\mathbb{E} \left[\frac{1}{|Y|} \sum_{v \in Y} d'(x, v) d'(z, v) \right] = \frac{1}{|Y|} \sum_{v \in Y} f(x, v) f(z, v) = \langle f_x, f_z \rangle,$$

where $f_x : Y \rightarrow \mathbb{R}$ is defined by $f_x(y) = f(x, y)$. Thus, by Theorem 3.4, with probability at least $1 - 5|X|^2 \exp(-c|Y| \kappa^2 \sigma^4 / (25600 C_1^2 (C_1^2 + C_2^2 + 1)))$, we have that for all distinct $x \neq z \in X$,

$$\left| \frac{1}{|Y|} \sum_{v \in Y} d'(x, v) d'(z, v) - \langle f_x, f_z \rangle \right| \leq \frac{\kappa \sigma^2}{40}. \quad (3)$$

By using triangle inequality on a few applications of (3), we have for every distinct $x, y, z \in X$,

$$\left| \frac{1}{|Y|} \sum_{v \in Y} (d'(x, v) - d'(y, v)) d'(z, v) - \langle f_x - f_y, f_z \rangle \right| \leq \frac{\kappa \sigma^2}{20}. \quad (4)$$

Observe that $\|f_x\|_2 = \sum_{y \in Y} |f(x, y)|^2 \leq C_2$ and for every $x \in X$, there exists $y \neq x \in X$ such that $d(x, y) \leq \delta$, hence by theorem assumption,

$$\|f_x - f_y\|_2 = \sqrt{\left(\frac{1}{|Y|} \sum_{z \in Y \setminus \{x, y\}} |f(x, z) - f(y, z)|^2 \right) + \frac{1}{|Y|} \cdot 2f(x, y)^2} \leq \sqrt{\left(\frac{\kappa\sigma^2}{8C_2} \right)^2 + \frac{2C_2^2}{|Y|}} \leq \frac{\kappa\sigma^2}{8C_2} + \frac{2C_2}{\sqrt{|Y|}}. \quad (5)$$

If $y \in B(x, \delta)$, then $\|f_x - f_y\| \leq \frac{\kappa\sigma^2}{8C_2}$ using the same calculation as in (5), so by Lemma 3.6 and (4) (and using that $N_0 = |Y|$ is sufficiently large),

$$\sup_{z \in X, z \neq x, y} \left| \frac{1}{|Y|} \sum_{v \in Y} (d'(x, v) - d'(y, v))d'(z, v) \right| \leq C_2 \cdot \left(\frac{\kappa\sigma^2}{8C_2} + \frac{2C_2}{\sqrt{|Y|}} \right) + \frac{1}{20}\kappa\sigma^2 \leq \frac{1}{5}\kappa\sigma^2.$$

If $d(x, y) \geq 4\varepsilon$, then by assumption we have $2C_2 \geq \|f_x - f_y\|$ and $\|f_x - f_y\|^2 \geq \kappa\sigma^2$, hence by Lemma 3.6, we have

$$\sup_{z \in X, z \neq x, y} \left| \frac{1}{|Y|} \sum_{v \in Y} (d'(x, v) - d'(y, v))d'(z, v) \right| \geq \frac{1}{2}\kappa\sigma^2 - 2C_2 \cdot \frac{\kappa\sigma^2}{8C_2} - \frac{1}{20}\kappa\sigma^2 = \frac{1}{5}\kappa\sigma^2.$$

□

Recall that we have conditioned on $d'(x, y)$ with $\{x, y\} \cap Y \neq \emptyset$; in particular, we have not conditioned on $d'(x, y)$ with $\{x, y\} \cap Y = \emptyset$. We can ensure $\mathcal{C}(x) \cap Y = \emptyset$. Then, we compute the noisy averages (for $x, y \in Y$)

$$A(x, y) := \frac{1}{|\mathcal{C}(y)|} \sum_{y' \in \mathcal{C}(y)} d'(x', y'),$$

where $x' = x'(x)$ is an arbitrary point in $\mathcal{C}(x)$ that we pick once and for all. Note that

$$\mathbb{E}A(x, y) = \frac{1}{|\mathcal{C}(y)|} \sum_{y' \in \mathcal{C}(y)} \mathbb{E}d'(x', y').$$

Lemma 3.10. $\mathbb{E}A(x, y) - \mathbb{E}A(x, z) \geq \sigma$ for all $x, y, z \in Y$ with $d(x, y) \geq d(x, z) + 17\varepsilon$.

Proof. If $d(x, y) \geq d(x, z) + 17\varepsilon$, then for every $x' \in \mathcal{C}(x)$, $y' \in \mathcal{C}(y)$, and $z' \in \mathcal{C}(z)$, we have $d(x', y') \geq d(x', z') + \varepsilon$ by triangle inequality and the fact that $\mathcal{C}(x) \subset B(x, 4\varepsilon)$, etc. Thus by the fact that an average of a sequence is between its minimum and maximum element and using our theorem assumption, we have

$$\mathbb{E}A(x, y) - \mathbb{E}A(x, z) \geq \min_{y' \in \mathcal{C}(y)} f(x', y') - \max_{z' \in \mathcal{C}(z)} f(x', z') \geq \sigma. \quad \square$$

Now by Theorem 2.2, with probability $1 - 2|Y|^2 \exp(-c\alpha N \sigma^2 / C_1^2)$, we have for all $x, y \in Y$,

$$|A(x, y) - \mathbb{E}A(x, y)| \leq \frac{\sigma}{3}.$$

Thus, if $d(x, y) \geq d(x, z) + 17\varepsilon$, we get $A(x, y) > A(x, z)$, as desired.

Finally, to make the probability of success at least $1 - \theta$, it suffices to ensure that:

$$\begin{aligned} 5|X|^2 \exp(-c|Y|\kappa^2 \sigma^4 / (C_1^4 + C_1^2 C_2^2)) &\leq \frac{\theta}{2}, \\ 2|Y|^2 \exp(-c\alpha N \cdot \sigma^2 / C_1^2) &\leq \frac{\theta}{2}. \end{aligned}$$

One can check that the choices of N_0 and N in the theorem statement work.

3.3 In the Case of Missing Noise

Using Theorem 3.4 applied to the random variables $d'(x, z)m(x, z)m(y, z)m(v, z)m(w, z)$ (which has Orlicz norm $\leq O(C_1 + C_2)$) and $d'(v, z)$, with probability $\geq 1 - 5|X|^4 \exp(-c|Y|\kappa^2 \sigma^4 c_2^2 / (C_1^4 + C_2^4))$, we can guarantee for all $x, y, v, w \in X$,

$$\begin{aligned} \left| \frac{1}{|Y|} \sum_{z \in Y} (d'(x, z)d'(v, z)m(x, z)m(y, z)m(v, z)m(w, z) \right. \\ \left. - f(x, z)f(v, z)p(x, z)p(y, z)p(v, z)p(w, z)) \right| < \frac{1}{40}\kappa\sigma^2 c_2, \quad (6) \end{aligned}$$

which implies

$$\begin{aligned} \left| \frac{1}{|Y|} \sum_{z \in Y} ((d'(x, z) - d'(y, z))(d'(v, z) - d'(w, z))m(x, z)m(y, z)m(v, z)m(w, z) \right. \\ \left. - (f(x, z) - f(y, z))(f(v, z) - f(w, z))p(x, z)p(y, z)p(v, z)p(w, z)) \right| < \frac{1}{10}\kappa\sigma^2 c_2, \quad (7) \end{aligned}$$

Using Theorem 2.3, with probability $\geq 1 - 2|X|^2 \exp(-|Y|c_1^2/32)$, we have for all $x, y \in X$,

$$\left| \frac{1}{|Y|} \sum_{z \in Y} (m(x, z)m(y, z) - p(x, z)p(y, z)) \right| < 0.5c_1. \quad (8)$$

Let

$$\mathcal{C}(x) = \{y \in X : \frac{1}{|Y|} \sum_{z \in Y} m(x, z)m(y, z) \geq 1.5c_1 \text{ and}$$

$$\sup_{\substack{v, w \in X \\ |\{v, w, x, y\}|=4}} \left| \frac{1}{|Y|} \sum_{z \in Y} (d'(x, z) - d'(y, z))(d'(v, z) - d'(w, z))m(x, z)m(y, z)m(v, z)m(w, z) \right| \leq \frac{3}{8}\kappa\sigma^2 c_2\}.$$

Computing each cluster takes time $N^3 N_0$, so it takes total time $N^3 N_0^2$ to compute $\mathcal{C}(x)$ for each $x \in Y$. It turns out that this runtime dominates the algorithm. Now, we show that $B(x, \delta) \subset \mathcal{C}(x) \subset B(x, 4\varepsilon)$.

Suppose $d(x, y) < \delta$. Then by Theorem 3.1 assumption, we have $\frac{1}{|Y|} \sum_{z \in Y} m(x, z)m(y, z) \geq 2c_1$, so by (8), we have $\frac{1}{|Y|} \sum_{z \in Y} p(x, z)p(y, z) \geq 1.5c_1$. Furthermore, for all $v, w \in X$, we have

$$\left| \frac{1}{|Y|} \sum_{z \in Y} (f(x, z) - f(y, z))(f(v, z) - f(w, z))p(x, z)p(y, z)p(v, z)p(w, z) \right| \leq 2C_2 \cdot \frac{\kappa\sigma^2}{8C_2} + \frac{2C_2^2}{|Y|} \leq 0.26\kappa\sigma^2 c_2.$$

Combining this with (7) shows that $y \in \mathcal{C}(x)$. Now suppose $d(x, y) > 4\varepsilon$. Choose $v \in B(x, \delta)$ and $w \in B(y, \delta)$ such that x, y, v, w are distinct (this is doable by the first assumption in Theorem 3.1). Then either $\frac{1}{|Y|} \sum_{z \in Y} m(x, z)m(y, z) < 1.5c_1$, in which case $y \notin \mathcal{C}(x)$, or $\frac{1}{|Y|} \sum_{z \in Y} m(x, z)m(y, z) \geq 1.5c_1$, in which case $\frac{1}{|Y|} \sum_{z \in Y} p(x, z)p(y, z) \geq c_1$ by (8) and so by Theorem 3.1 assumption, we get

$$\begin{aligned} & \left| \frac{1}{|Y|} \sum_{z \in Y} (f(x, z) - f(y, z))(f(v, z) - f(w, z))p(x, z)p(y, z)p(v, z)p(w, z) \right| \\ & \geq \left| \frac{1}{|Y|} \sum_{z \in Y} |f(x, z) - f(y, z)|^2 p(x, z)p(y, z)p(v, z)p(w, z) \right| \\ & - \left| \frac{1}{|Y|} \sum_{z \in Y} (f(x, z) - f(y, z))(f(x, z) - f(v, z))p(x, z)p(y, z)p(v, z)p(w, z) \right| \\ & - \left| \frac{1}{|Y|} \sum_{z \in Y} (f(x, z) - f(y, z))(f(y, z) - f(w, z))p(x, z)p(y, z)p(v, z)p(w, z) \right| \\ & \geq \kappa c_2 \sigma^2 - 4C_2 \cdot \frac{\kappa\sigma^2 c_2}{8C_2} - \frac{4C_2^2}{|Y|} \geq 0.49\kappa c_2 \sigma^2. \end{aligned}$$

Combining this with (7) and (8) shows that $y \notin \mathcal{C}(x)$.

Recall that we have conditioned on $d'(x, y)$ with $\{x, y\} \cap Y \neq \emptyset$; in particular, we have not conditioned on $d'(x, y)$ with $\{x, y\} \cap Y = \emptyset$. Truncate the clusters to ensure that $\mathcal{C}(x) \cap Y = \emptyset$. Then, we compute the noisy averages (for $x, y \in Y$)

$$A(x, y) := \begin{cases} \frac{\sum_{y' \in \mathcal{C}(y)} d'(x', y') m(x', y')}{\sum_{y' \in \mathcal{C}(y)} m(x', y')}, & \text{if } \sum_{y' \in \mathcal{C}(y)} m(x', y') > \frac{1}{2} \alpha N \phi; \\ +\infty & \text{otherwise.} \end{cases}$$

where $x' = x'(x)$ is some arbitrary point we pick once and for all in $\mathcal{C}(x)$. (Note that A is not symmetric.) Intuitively, $A(x, y)$ is close to $f(x, y)$, but only if the denominator is sufficiently large.

By Theorem 2.2, we have with probability $\geq 1 - 2|Y|^2 \exp(-\frac{\varepsilon}{2} \alpha N \phi \cdot \sigma^2 / C_1^2)$,

$$A(x, y) \neq \infty \implies \left| A(x, y) - \frac{\sum_{y' \in \mathcal{C}(y)} f(x', y') m(x', y')}{\sum_{y' \in \mathcal{C}(y)} m(x', y')} \right| \leq \frac{\sigma}{3}. \quad (9)$$

Now we verify the conclusion of Theorem 3.1. First, if $d(x, y) \leq r_0 - 8\varepsilon$, then $d(x', y') \leq r_0$ for any $x' \in \mathcal{C}(x)$ and $y' \in \mathcal{C}(y)$, so $p(x', y') \geq \phi$ by Theorem 3.1 assumption. Hence, $\sum_{y' \in \mathcal{C}(y)} p(x', y') \geq \alpha N \phi$, so by applying Theorem 2.4 to the Bernoulli independent random variables $m(x', y')$, whose variance is $p(x', y')(1 - p(x', y')) \leq p(x', y')$, we get

$$\Pr\left(\sum_{y' \in \mathcal{C}(y)} m(x', y') < \frac{1}{2} \sum_{y' \in \mathcal{C}(y)} p(x', y')\right) \leq 2 \exp\left(-\frac{3}{32} \sum_{y' \in \mathcal{C}(y)} p(x', y')\right).$$

Thus, we get $\sum_{y' \in \mathcal{C}(y)} m(x', y') \geq \frac{1}{2} \alpha N \phi$, which implies $A(x, y) \neq \emptyset$, with probability $\geq 1 - 2 \exp(-\frac{3}{32} \alpha N \phi)$.

Next, we claim if $x, y, z \in Y$ satisfy $d(x, y) \geq d(x, z) + 17\varepsilon$ and $A(x, y), A(x, z) \neq +\infty$, then $A(x, y) > A(x, z)$. Indeed, by (9),

$$\frac{\sum_{y' \in \mathcal{C}(y)} f(x', y') m(x', y')}{\sum_{y' \in \mathcal{C}(y)} m(x', y')} - \frac{\sum_{y' \in \mathcal{C}(y)} f(x', y') m(x', y')}{\sum_{y' \in \mathcal{C}(y)} m(x', y')} \geq \min_{y' \in \mathcal{C}(y)} f(x', y') - \max_{z' \in \mathcal{C}(z)} f(x', z') \geq \sigma.$$

Finally, to make the probability of success at least $1 - \theta$, it suffices to ensure that:

$$\begin{aligned} 5|X|^4 \exp(-c|Y|\kappa^2 \sigma^4 c_2^2 / (C_1^4 + C_2^4)) &\leq \frac{\theta}{4}, \\ 2|X|^2 \exp(-|Y|c_1^2/32) &\leq \frac{\theta}{4}, \\ 2|Y|^2 \exp(-\frac{c}{2} \alpha N \phi \cdot \sigma^2 / C_1^2) &\leq \frac{\theta}{4}, \\ 2|Y|^2 \exp(-\frac{3}{32} \alpha N \phi) &\leq \frac{\theta}{4}. \end{aligned}$$

One can check that the choices of N_0 and N in the theorem statement work.

3.4 Finding distances from clusters

Let Y be a set. Suppose we have the following oracle:

Oracle 3.11. Fix $\varepsilon > 0$. We have an oracle $\mathcal{O}_1(x, y) \in \mathbb{R}$ such that if $x, y, z \in Y$ and $d(x, y) \geq d(x, z) + \varepsilon$, then $\mathcal{O}_1(x, y) > \mathcal{O}_1(x, z)$.

Using Theorem 3.11, we can solve our original problem.

Theorem 3.12. Let Y be a diameter 1 metric space satisfying: if $x, y \in Y$ and $a, b \geq 0$ satisfy $a + b = d(x, y)$, then there exists $z \in Y$ satisfying $|d(x, z) - a| \leq \varepsilon$ and $|d(y, z) - b| \leq \varepsilon$. Assume Theorem 3.11 for Y and a given $0 < \varepsilon \leq 2^{-20}$. Then for any $x, y \in Y$, we can recover $d(x, y)$ up to $O(\varepsilon \log \varepsilon^{-1})$.

In fact, we will need an intermediate oracle.

Oracle 3.13. Fix $\varepsilon > 0$. For any $x, y \in Y$, we can find $z \in Y$ such that $\frac{d(x, y)}{2} - \varepsilon \leq d(x, z), d(y, z) \leq \frac{d(x, y)}{2} + \varepsilon$ in time $|Y|$.

Proposition 3.14. *Let Y be a metric space satisfying: if $x, y \in Y$ and $a, b \geq 0$ satisfy $a + b = d(x, y)$, then there exists $z \in Y$ satisfying $|d(x, z) - a| \leq \varepsilon$ and $|d(y, z) - b| \leq \varepsilon$. Then Theorem 3.11 for ε implies Theorem 3.13 for $\frac{9\varepsilon}{2}$.*

Proof. For any $x, y \in Y$, we choose $z \in Y$ to minimize $\mathcal{O}_1(x, z)$ subject to $\mathcal{O}_1(z, y) < \mathcal{O}_1(z, x)$. This z satisfies $d(z, y) \leq d(z, x) + \varepsilon$ (by contrapositive of oracle assumption), so $d(x, z) \geq \frac{d(x, z) + d(y, z) - \varepsilon}{2} \geq \frac{d(x, y) - \varepsilon}{2}$. On the other hand, there exists $z^* \in Y$ such that $|d(x, z^*) - \frac{d(x, y)}{2} + \frac{3\varepsilon}{2}| \leq \varepsilon$ and $|d(y, z^*) - \frac{d(x, y)}{2} - \frac{3\varepsilon}{2}| \leq \varepsilon$. Then $\mathcal{O}_1(z^*, y) < \mathcal{O}_1(z^*, x)$, so $\mathcal{O}_1(x, z) \leq \mathcal{O}_1(x, z^*)$, which means $d(x, z) \leq d(x, z^*) + \varepsilon \leq \frac{d(x, y)}{2} + \frac{7\varepsilon}{2}$. Hence, $\frac{d(x, y) - \varepsilon}{2} \leq d(x, z) \leq \frac{d(x, y) + 7\varepsilon}{2}$. Combined with $d(x, y) - d(x, z) \leq d(y, z) \leq d(z, x) + \varepsilon$, this implies $\frac{d(x, y) - 7\varepsilon}{2} \leq d(y, z) \leq \frac{d(x, y) + 9\varepsilon}{2}$. $\square$

Oracle 3.15. *For any $x, y, z \in Y$, we can find a real number r such that $|\min\{d(x, z), d(x, y)\} - rd(x, y)| \leq \varepsilon$ in $|Y| \log \varepsilon^{-1}$ time.*

Proposition 3.16. *Theorem 3.11 for $\varepsilon \leq 2^{-20}$ implies Theorem 3.15 for $O(\varepsilon \log \varepsilon^{-1})$.*

Proof. By Proposition 3.14, we may additionally assume Oracle 3.13 for 4.5ε . Call the respective oracles $\mathcal{O}_1$ and $\mathcal{O}_2$. Setting $n = \lceil \log \varepsilon^{-1} \rceil$, define a function $f : [0, 1] \cap 2^{-n}\mathbb{Z} \rightarrow Y$ by $f(0) = x$, $f(1) = y$, and recursively define $f(a + 2^{-k-1}) = \mathcal{O}_2(f(a), f(a + 2^{-k}))$ whenever $a \in 2^{-k}\mathbb{Z}$. Then using Theorem 3.13 we inductively see that

$$|d(f(a), f(a + 2^{-k})) - 2^{-k}d(x, y)| \leq 9\varepsilon \quad (10)$$

whenever $a \in 2^{-k}\mathbb{Z}$. Indeed, the base case $k = 0$ is obvious, and if true for k , then for $k + 1$,

$$\begin{aligned} d(f(a), f(a + 2^{-k-1})) &= d(f(a), \mathcal{O}_2(f(a), f(a + 2^{-k}))) \\ &\in [\frac{1}{2}d(f(a), f(a + 2^{-k})) - 4.5\varepsilon, \frac{1}{2}d(f(a), f(a + 2^{-k})) + 4.5\varepsilon] \\ &\in [2^{-(k+1)}d(x, y) - 9\varepsilon, 2^{-(k+1)}d(x, y) + 9\varepsilon]. \end{aligned}$$

Thus, by using (10) and the binary representation of $a = \sum_{i=1}^{k_a} 2^{-a_i} \in 2^{-n}\mathbb{Z}$, we have

$$\begin{aligned} d(x, f(a)) &\leq \sum_{i=1}^{k_a} d(f(\sum_{j=1}^{i-1} 2^{-a_j}), f(\sum_{j=1}^i 2^{-a_j})) \leq ad(x, y) + 9k_a\varepsilon \leq ad(x, y) + 9n\varepsilon, \\ d(y, f(a)) &\leq (1 - a)d(x, y) + 9k_{1-a}\varepsilon \leq (1 - a)d(x, y) + 9n\varepsilon, \end{aligned}$$

where k_a is the number of ones in the binary representation of a . By pitting this against the bound $d(x, y) \leq d(x, f(a)) + d(y, f(a))$, we obtain

$$|d(x, f(a)) - ad(x, y)| \leq 9n\varepsilon. \quad (11)$$

Now choose the largest $a \in 2^{-n}\mathbb{Z}$ such that $\mathcal{O}_1(x, z) > \mathcal{O}_1(x, f(a))$; we claim that $|\min\{d(x, y), d(x, z)\} - ad(x, y)| \leq 9n\varepsilon + 16\varepsilon$. If $a = 1$, then $\mathcal{O}_1(x, z) > \mathcal{O}_1(x, y)$, so $d(x, z) \geq d(x, y) - \varepsilon$, which means

$|\min\{d(x, y), d(x, z)\} - ad(x, y)| \leq \varepsilon$. If $a < 1$, then $\mathcal{O}_1(x, f(a + 2^{-n})) \geq \mathcal{O}_1(x, z) > \mathcal{O}_1(x, f(a))$, which forces

$$d(x, f(a + 2^{-n})) + \varepsilon \geq d(x, z) \geq d(x, f(a)) - \varepsilon.$$

Using (11) shows that $|d(x, z) - ad(x, y)| \leq (9n + 1)\varepsilon$. This also shows that $|\min\{d(x, y), d(x, z)\} - ad(x, y)| \leq \varepsilon$, as desired. Finally, the $|Y| \log \varepsilon^{-1}$ runtime is from binary search. $\square$

Proof of Theorem 3.12. Use Propositions 3.14 and 3.16 to get oracles $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3$. Choose any $x_1 \in Y$, and let $x_2 \in Y$ maximize $\mathcal{O}_1(x_1, x_2)$. Note that $d(x_1, x_2) \geq \frac{1}{2} - \varepsilon$. Let $x' = \mathcal{O}_2(x_1, x_2)$. For any $y \in Y$, we test if $\mathcal{O}_1(x_1, y) \leq \mathcal{O}_1(x_1, x')$; if this is true, then set $x = x_2$; otherwise, set $x = x_1$.

Claim. $d(x, y) \geq \frac{1}{4} - 7\varepsilon$.

Proof. If $\mathcal{O}_1(x_1, y) \leq \mathcal{O}_1(x_1, x')$, then $d(x_1, y) \leq d(x_1, x') + \varepsilon \leq \frac{d(x_1, x_2)}{2} + 5.5\varepsilon$, hence $d(x_2, y) \geq \frac{d(x_1, x_2)}{2} - 5.5\varepsilon \geq \frac{1}{4} - 7\varepsilon$.

If $\mathcal{O}_1(x_1, y) > \mathcal{O}_1(x_1, x')$, then $d(x_1, y) \geq d(x_1, x') - \varepsilon \geq \frac{d(x_1, x_2)}{2} - 5.5\varepsilon \geq \frac{1}{4} - 7\varepsilon$. This proves the Claim.

Now, for $z \in Y$, we use Theorem 3.15 to get $\frac{d(y, z)}{d(y, x)}$ and $\frac{d(y, x)}{d(x_1, x_2)}$ up to additive error $O(\varepsilon \log \varepsilon^{-1})$ (since both $d(y, x)$ and $d(x_1, x_2)$ are large), hence we get $\frac{d(y, z)}{d(x_1, x_2)}$ up to additive error $O(\varepsilon \log \varepsilon^{-1})$. Thus, we can figure out which pair $y, z \in Y$ maximizes $d(y, z)$ and normalize to find all $d(y, z)$ up to additive error $O(\varepsilon \log \varepsilon^{-1})$. $\square$

For the missing data case, we need a slightly stronger oracle.

Theorem 3.17. Suppose $r \in (0, \frac{1}{100}]$ and let (Y, d) be a diameter $D \in [r, 1]$ metric space satisfying the following “geodesic” property: for every $x, y \in Y$ and numbers $a_1, \dots, a_n \geq 0$ with $\sum_{i=1}^n a_i = d(x, y)$, there exist $x = x_0, x_1, x_2, \dots, x_n = y \in Y$ such that $|d(x_{i-1}, x_i) - a_i| \leq \varepsilon$ for each $1 \leq i \leq n$. Suppose also that there exist numbers $A(x, y) \in \mathbb{R} \cup \{+\infty\}$ such that:

1. for every $x, y \in Y$, if $d(x, y) \leq r$, then $A(x, y) \neq \infty$;
2. for every $x, y, z \in Y$ satisfying $d(x, y) \geq d(x, z) + \varepsilon$ and $A(x, y), A(x, z) \neq +\infty$, we have $A(x, y) > A(x, z)$.

If $\varepsilon < \frac{r^2 D}{100}$, then we can determine the distances $d(x, y)$ up to additive error $O(\frac{\varepsilon}{r^2} \log \varepsilon^{-1})$.

Remark 3.18. In fact, we can do slightly better: error $O(\frac{\varepsilon D^2}{r^2} \log \varepsilon^{-1})$ and only need $\varepsilon < \frac{r^2}{100}$. This is by taking paths of length $\leq \frac{3D}{r}$ instead of $\leq \frac{3}{r}$ in the ensuing argument. However, when $D = 1$ this is the same bound, and furthermore the theorem can handle when we don’t know the diameter D (but know it lies in $[r, 1]$).

Remark 3.19. In fact, the “geodesic” property for n with 2ε follows from the geodesic property for ε by induction, similarly to (10).

Proof. This is the most annoying part of the argument; there should be simpler ways to do so, but the approach we take has a nice detour to graph algorithms and may be of independent interest, besides potentially improving Proposition A.1 of [13] by allowing precise recovery of all distances (instead of those below a fixed scale).

Define a directed graph G on pairs $(a, b) \in Y \times Y$ where we connect (a, b) to (b, c) by a directed edge if $\infty > A(b, a) > A(b, c)$. Note that G has $|Y|^2$ vertices and $|Y|^3$ edges, so Dijkstra's shortest path algorithm (on the unweighted directed graph G) operates in time $O(|Y|^3)$.

Now, for all $x \in Y$, define $\tau(x) \in Y$ as the minimizer of $A(x, \tau(x))$ subject to for every $z \in Y$, there exists $z' \in Y$ and a directed path from $(\tau(x), x)$ to (z', z) with length at most $\frac{3}{r}$.

Claim. $d(x, \tau(x)) \in [\frac{rD}{9}, \frac{rD}{2}]$.

Proof. Choose any $x' \in Y$ such that $d(x, x') \in [\frac{2rD}{5} - \varepsilon, \frac{2rD}{5} + \varepsilon]$. We show that for every $z \in Y$, there exists $z' \in Y$ and a directed path from (x', x) to (z', z) with length at most $\frac{3}{r}$. Indeed, choose $n = \lfloor \frac{3}{r} \rfloor$ and a sequence $x = x_0, x_1, x_2, \dots, x_n = z$ consisting of points in Y such that

$$|d(x_{i-1}, x_i) - \frac{d(x, z)}{n} + (\frac{n+1}{2} - i) \cdot 3\varepsilon| \leq \varepsilon.$$

Plugging in $i = 1$, we deduce that (since $\frac{3}{r} \geq n \geq \frac{3}{r} - 1 \geq \frac{2.99}{r}$ and $\varepsilon < \frac{r^2 D}{100}$)

$$d(x', x) \geq \frac{2rD}{5} - \varepsilon > \frac{D}{n} + \frac{3n+1}{2}\varepsilon \geq d(x, x_1) + \varepsilon,$$

hence $A(x', x) > A(x, x_1)$ and there is an edge from (x', x) to (x_0, x_1) . Similarly, there exists an edge from (x_{i-1}, x_i) to (x_i, x_{i+1}) for all $1 \leq i \leq n-1$, showing the existence of the directed path from (x', x) to (z', z) .

By minimality, we have $A(x, \tau(x)) < A(x, x')$, so $d(x, \tau(x)) \leq d(x, x') + \varepsilon \leq \frac{rD}{2}$.

Now, we show $d(x, \tau(x)) \geq \frac{rD}{9}$. Choose z such that $d(x, z) \geq \frac{D}{2}$ (this exists since our metric space has diameter D); we know there exists a directed path from $(\tau(x), x)$ to (z', z) with length at most $\frac{3}{r}$, which corresponds to a $x = x_0, x_1, x_2, \dots, x_n = z$ consisting of points in Y such that $A(x_i, x_{i-1}) > A(x_i, x_{i+1})$, hence $d(x_i, x_{i+1}) \leq d(x_{i-1}, x_i) + \varepsilon$. By the triangle inequality, we have

$$nd(x, \tau(x)) + \frac{n(n+1)}{2}\varepsilon \geq \frac{D}{2}.$$

This forces $d(x, \tau(x)) \geq \frac{D}{3n} \geq \frac{rD}{9}$, proving the Claim.

Using this, for any $x, y \in Y$ with $d(x, y) \leq r$, we can find an approximate midpoint $z \in Y$ satisfying $\frac{1}{2}d(x, y) - 4.5\varepsilon \leq d(x, z), d(y, z) \leq \frac{1}{2}d(x, y) + 4.5\varepsilon$ in time $|Y|$ (Theorem 3.14).

Repeatedly using this approximate midpoint oracle and assuming ε is a negative power of 2 for simplicity, we can find a function $\Gamma_x : [0, 1] \cap \varepsilon\mathbb{Z} \rightarrow Y$ such that

$$|d(x, \Gamma_x(\varepsilon k)) - \varepsilon k d(x, \tau(x))| = O(\varepsilon \log \varepsilon^{-1}) \quad (12)$$

for each $0 \leq k \leq \varepsilon^{-1}$ in ε^{-1} time (see proof of Theorem 3.16). Thus, for any $x, y \in Y$, we

can use binary search to compare $A(x, y)$ to each $A(x, \Gamma_x(k\varepsilon))$ for $0 \leq k \leq \varepsilon^{-1}$, so we can find $\min\{\frac{d(x, y)}{d(x, \tau(x))}, 1\}$ up to error $O(\frac{\varepsilon \log \varepsilon^{-1}}{r})$ in time $\log \varepsilon^{-1}$.

Using this approximate distance ratio oracle, for any $x, y \in Y$, we can find $\frac{d(x, y)}{d(x, \tau(x))}$ up to error $O(\frac{\varepsilon \log \varepsilon^{-1}}{rD} \left(1 + \frac{d(x, y)}{d(x, \tau(x))}\right)^2)$. Indeed, if $A(x, y) < A(x, \tau(x))$, then we use the previous oracle; otherwise, this can be approximated by the minimum value of $k\varepsilon \cdot n_k$ over all $k\varepsilon \in [\frac{1}{3}, \frac{2}{3}]$, where n_k is the length of the shortest path from $(\Gamma_x(k\varepsilon), x)$ to some (y', y) . To prove it, we follow a similar thought process as in Claim. Choose n such that $\frac{d(x, y)}{n} \in [\frac{d(x, \tau(x))}{3}, \frac{2d(x, \tau(x))}{3}]$, and choose a sequence $x = x_0, x_1, x_2, \dots, x_n = y$ consisting of points in Y such that

$$|d(x_{i-1}, x_i) - \frac{d(x, y)}{n} + (\frac{n+1}{2} - i) \cdot 3\varepsilon| \leq \varepsilon.$$

Note that for $k\varepsilon \cdot d(x, \tau(x)) > \frac{d(x, y)}{n} + \frac{C\varepsilon \log \varepsilon^{-1}}{rD}$, we have $d(\Gamma_x(k\varepsilon), x) > \frac{d(x, y)}{n} + \frac{n+1}{2}\varepsilon \geq d(x, x_1) + \varepsilon$, hence $A(x', x) > A(x, x_1)$ and there is an edge from (x', x) to (x_0, x_1) . Similarly, there exists an edge from (x_{i-1}, x_i) to (x_i, x_{i+1}) for all $1 \leq i \leq n-1$, showing the existence of the directed path from (x', x) to (y', y) . Thus, by minimality and $n \leq \frac{3d(x, y)}{d(x, \tau(x))}$, we have

$$nk\varepsilon \leq \frac{d(x, y)}{d(x, \tau(x))} + O(\frac{n^2\varepsilon \log \varepsilon^{-1}}{d(x, \tau(x))}) = \frac{d(x, y)}{d(x, \tau(x))} + O(\frac{\varepsilon \log \varepsilon^{-1}}{rD} \left(1 + \frac{d(x, y)}{d(x, \tau(x))}\right)^2).$$

Now, we show $nk\varepsilon \geq \frac{d(x, y)}{d(x, \tau(x))} - O(\frac{\varepsilon \log \varepsilon^{-1}}{rD} \left(1 + \frac{d(x, y)}{d(x, \tau(x))}\right)^2)$. We know there exists a directed path from $(\Gamma_x(k\varepsilon), x)$ to (z', z) with length at most n , which corresponds to a $x = x_0, x_1, x_2, \dots, x_n = y$ consisting of points in Y such that $A(x_i, x_{i-1}) > A(x_i, x_{i+1})$, hence $d(x_i, x_{i+1}) \leq d(x_{i-1}, x_i) + \varepsilon$. By (12) and the triangle inequality, we have

$$nk\varepsilon d(x, \tau(x)) + (O(n \log \varepsilon^{-1}) + \frac{n(n+1)}{2})\varepsilon \geq nd(x, \Gamma_x(k\varepsilon)) + \frac{n(n+1)}{2}\varepsilon \geq d(x, y). \quad (13)$$

If $nk\varepsilon \geq \frac{d(x, y)}{d(x, \tau(x))}$, then we have an even stronger statement than desired; otherwise, if $nk\varepsilon < \frac{d(x, y)}{d(x, \tau(x))}$, then we get from $k\varepsilon \in [\frac{1}{3}, \frac{2}{3}]$ that $n \leq \frac{3d(x, y)}{d(x, \tau(x))}$, so rearranging (13) gives $nk\varepsilon \geq \frac{d(x, y)}{d(x, \tau(x))} - O(\frac{\varepsilon \log \varepsilon^{-1}}{rD} \left(1 + \frac{d(x, y)}{d(x, \tau(x))}\right)^2)$, as desired.

Now, to compute $\frac{d(x, \tau(x))}{d(z, \tau(z))}$ for any x, y, z , we note that if $A(z, x) = \infty$ or $A(z, x) > A(z, \tau(x))$, then $d(z, x) \geq r$ in the first case and $d(z, x) \geq d(z, \tau(x)) - \varepsilon \geq d(x, \tau(x)) - d(z, x) - \varepsilon$ implies $d(z, x) \geq \frac{rD}{20}$ in the second case, and therefore we can compute $\frac{d(x, \tau(x))}{d(z, \tau(z))} = \frac{d(z, x)}{d(z, \tau(z))} / \frac{d(z, x)}{d(x, \tau(x))}$. Otherwise if $A(z, \tau(x)) = \infty$ or $A(z, \tau(x)) \geq A(z, x)$, then $d(z, \tau(x)) \geq \frac{rD}{20}$, and therefore we can compute this as $\frac{d(z, \tau(x))}{d(z, \tau(z))} / \frac{d(z, \tau(x))}{d(x, \tau(x))}$. The additive error is $O(\frac{\varepsilon}{r^2D} \log \varepsilon^{-1})$ since it is the quotient of two terms of the form $n \pm \frac{n^2\varepsilon \log \varepsilon^{-1}}{rD}$ where $n \sim \frac{d(z, x)}{d(z, \tau(z))}$, $\frac{1}{20} \leq n \leq \frac{10}{r}$. Also, note that $\frac{d(x, \tau(x))}{d(z, \tau(z))} \in [\frac{1}{100}, 100]$.

Finally, to compute $d(x, y)$, we use $\frac{d(x, y)}{d(x, \tau(x))} \cdot \frac{d(x, \tau(x))}{d(z, \tau(z))} \cdot d(z, \tau(z))$. For a fixed z , this allows us to compute all $d(x, y)$ up to error $O(\frac{\varepsilon \log \varepsilon^{-1}}{r^2})$ (since $\frac{d(x, y)}{d(x, \tau(x))} \leq \frac{10}{r}$). Use the diameter 1 assumption to normalize. $\square$

3.5 Proof of Theorem 1.1

The basic idea is to combine Theorem 2.5, Theorem 2.6, Theorem 3.1, and either Theorem 3.12 or Theorem 3.17. Finally, use Proposition A.1 from [13] (reproduced in this paper as Theorem 2.8).

We give some more details for the missing case; the non-missing case is easier.

Set $c_1 = \frac{\phi^2}{4} \text{VR}(\frac{r_0}{2})$. If $\varepsilon > \frac{1}{3}\lambda_2 c_1$, replace ε by $\frac{1}{3}\lambda_2 c_1$.

Set $\kappa = \min\{\frac{1}{2}\text{VR}(\frac{1}{8}\lambda_2 c_1), ci_0^d\}$, $c_2 = \lambda_1^4 c_1^4$, $\sigma = \frac{\varepsilon}{C_3}$, $\delta = \frac{\kappa \sigma^2}{8C_2 C_3}$, $\alpha = \frac{1}{2}\text{VR}(\delta) \sim \delta^d \sim \varepsilon^{2d}$, and finally

$$N_0 \sim \varepsilon^{-\max\{d,4\}} \left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\theta} \right), \quad N \sim \frac{CC_1^2}{\alpha \phi \sigma^2} \log \frac{N_0}{\theta} \sim \varepsilon^{-2d-2} (d \log \varepsilon^{-1} + \log \theta^{-1}).$$

where we choose N_0 and N large enough to satisfy

$$\begin{aligned} N_0 \exp(-c\delta^d N/8) &\leq \frac{\theta}{8}, \\ C \left(\frac{1}{\varepsilon} \right)^d \exp(-c\varepsilon^d N_0) &\leq \frac{\theta}{8}, \\ N \exp(-c(r_0 - \varepsilon)^d N_0) &\leq \frac{\theta}{8}, \\ N_0^4 \exp(-\frac{\kappa N_0}{8}) &\leq \frac{\theta}{8}. \end{aligned}$$

Thus, by Theorem 2.5 combined with the volume estimates in Theorem 2.6, with probability $\geq 1 - \frac{\theta}{2}$, we can ensure:

- every ball $B(x, \delta)$ with $x \in Y$ contains at least αN many elements of X ;
- Y is ε -dense in $\mathcal{M}$;
- every ball $B(x, r_0 - \varepsilon)$ with $x \in X$ contains at least $\frac{1}{2}\text{VR}(r_0 - \varepsilon)|Y|$ many elements of Y ;
- every $\Lambda_{x,y}^1 := \{d(x, z) \geq d(y, z) + \frac{d(x,y)}{4}\}$ with $x, y \in X$ contain at least $ci_0^d|Y|$ many elements of Y ;
- if $d(x, y) \geq \frac{r_0}{2}$, then $\Lambda_{x,y}^2 := \{z : d(x, z) - \frac{d(x,y)}{2} \in [\frac{\lambda_2 c_1}{3}, \lambda_2 c_1], d(x, z) - \frac{d(x,y)}{2} \in [-\lambda_2 c_1, \frac{\lambda_2 c_1}{3}]\}$ with $x, y \in X$ contain at least $\kappa|Y|$ many elements of Y .

We want to apply Theorem 3.1 with the above parameters. The key points to note:

- If $x, y \in X$ are distinct and satisfy $d(x, y) < \delta$, then if $z \in B(x, r_0 - \varepsilon)$, then $p(x, z), p(y, z) > \phi$.
- If $d(x, y) \leq \frac{r_0}{2}$, then since $c_2 \leq \phi^2$, we have $\Lambda_{x,y,v,w} \supset \Lambda_{x,y}^1$. If $d(x, y) > \frac{r_0}{2}$ and $\frac{1}{|Y|} \sum_{z \in Y} p(x, z)p(y, z) > c_1$, then $p(x, z)p(y, z) > c_1$ for some $z \in Y$. Without loss of generality let $d(x, z) \geq \frac{d(x,y)}{2}$; then $p(x, z) > c_1$. The condition on p from Theorem 1.1 and $\varepsilon \leq \frac{1}{3}\lambda_2 c_1$ then imply $\Lambda_{x,y,v,w} \supset \Lambda_{x,y}^2$.
- Our choice of N_0 and N satisfy the assumption because of the $\max\{d, 4\}$ in the definition of N_0 .

Finally, the hypothesis regarding Y of Theorem 3.12 and 3.17 is satisfied because Y is ε -dense in $\mathcal{M}$, which is a geodesic space. Also Y has diameter between $1 - 2\varepsilon$ and 1.

4 Algorithm 2: Regularized Optimization

We start with a sketch of our algorithm. We will sample three sets of points uniformly from μ . S_1 will be used for selecting our cluster “centers”, S_2 will be used for running inner products, and S_3 will be used to draw cluster points. Further, we will denote $N_i = |S_i|$. Let $k \in \mathbb{N}_{>0}$. Assume that we have drawn $k - 1$ clusters and have access to functions $D_i(y) = \frac{1}{n_i} \sum_{x \in C_i} d'(x, y)$, where C_i denotes the i -th cluster and $n_i = |C_i| = n$. Consider the objective function

$$g_k(x, y) = \langle f_x, f_y \rangle + \beta \min_{1 \leq i \leq k-1} \mathbb{E} D_i(x) + \beta \min_{1 \leq i \leq k-1} \mathbb{E} D_i(y)$$

where it is assumed that $\beta > 1$, $f_x(z) = \mathbb{E} d'(x, z)$, and $\langle \cdot, \cdot \rangle$ denotes the $L^2(\mu)$ inner product. We will find points $(\hat{x}_k, \hat{y}_k) \in S_1$ such that

$$g_k(\hat{x}_k, \hat{y}_k) \approx \max_{z \in \mathcal{M}} \left(\|f_z\|^2 + 2\beta \min_{1 \leq i \leq k-1} \mathbb{E} D_i(z) \right)$$

and then approximate $g_k(x, \hat{y}_k)$ for all $x \in S_3$, using the results to build C_k by selecting points where the objective is above a certain threshold. We will stop the algorithm when it can be detected that

$$\max_{x \in \mathcal{M}} \min_{1 \leq i \leq k-1} d(x, \hat{x}_i) \leq \varepsilon \quad (14)$$

We will force all cluster centers to be at least distance $c\varepsilon$ from each other, for some small constant $c > 0$. Thus, the maximal number of clusters k_{max} for which the algorithm still runs is upper bounded by a constant depending on ε . Finally, we use the resultant clusters to estimate the true distances between the centers, which are the ε -net our algorithm returns.

The algorithm works under general conditions on $\mathcal{M}$ and $d'(x, y)$. As such, we prove the following general statement (Theorem 4.1) below, which when combined with Proposition 2.6 applies to the manifold case where $(\mathcal{M}, \mu, d)$ is a d -dimensional Riemannian manifold with diameter 1, $|\text{Sec}_{\mathcal{M}}| \leq \Lambda^2$, injectivity radius $\geq i_0$, and μ is a probability measure on $\mathcal{M}$ which is ρ -mutually absolutely continuous with the uniform volume measure, with Radon-Nikodym derivative bounded in $[\rho, \frac{1}{\rho}]$.

Theorem 4.1 (Algorithm 2). *We assume the following*

- $(\mathcal{M}, \mu, d)$ is a diameter 1 geodesic probability space
- (d -regular at scale C_1) There exists a strictly monotonically increasing function $\text{VR} : [0, C_1] \rightarrow \mathbb{R}$ such that $\text{VR}(r) \leq \mu(B(x, r))$ for all $x \in \mathcal{M}$ and $\text{VR}(r) \geq c_2 r^d$
- for all $x, y \in \mathcal{M}$, $\mu(\Lambda_{x,y}) \geq \kappa$, where $\Lambda_{x,y} = \{z \in \mathcal{M} : d(x, z) \geq d(y, z) + \frac{d(x,y)}{4}\}$

We draw N points X μ -uniformly at random from $\mathcal{M}$. Suppose we have a noisy distance function $d'(x, y)$ satisfying the following properties:

- (Symmetry) $d'(x, y) = d'(y, x)$
- (Subgaussian) $d'(x, y)$ is sub-Gaussian with variance proxy $\|d'(x, y)\|_{\psi_2} \leq C_3$
- (“Independence”) For every $x \in S$, the variables $\{d'(x, y)\}_{y \in S}$ are independent conditioned on any subset of the other variables $\{d'(y, z)\}_{y, z \neq x}$

and furthermore whose expectation $f(x, y) = \mathbb{E}d'(x, y)$ satisfies the following properties:

- (Bilipschitz) $f(x, y) = F(d(x, y))$ with $F : [0, \infty) \rightarrow \mathbb{R}$ both C -bilipschitz and monotonically increasing
- (Intercept) $L = F(0) \geq 0$, although we don’t a priori know L

For a pick of $\theta > 0$ and small enough ε , there exists an algorithm that with probability at least $1 - \theta$ returns a ε -net of $\mathcal{M}$ and recovers the true distances between elements of this net up to an additive error of $O(\varepsilon \log \varepsilon^{-1})$. It requires $N = C_0 Q_0 \varepsilon^{-(3d+6)} (\log \varepsilon^{-1} + \log \theta^{-1})^{2(1+\frac{1}{d})}$, with runtime $\tilde{C}_0 Q_1(d) \varepsilon^{-Q_2(d)} (\log \varepsilon^{-1} + \log \theta^{-1})^{Q_3(d)}$ where
with

$$\begin{aligned}
Q_0 &= c_2^{-(1+\frac{2}{d})} (\kappa^{-1} C^{15} (L + C))^{d+2} \\
Q_1(d) &= \begin{cases} C_3 \kappa^{-(d+4)} c_2^{-2(1+\frac{1}{d})} C^{19d+58} (L + C)^{d+4} & d \leq 4 \\ \kappa^{-(2d+2)} c_2^{-1} C^{30d+20} (L + C)^{2d} (C_3^4 + (C_3^2 + C^2)(L + C)^2) & \text{otherwise} \end{cases} \\
Q_2(d) &= \begin{cases} 4d + 12 & d \leq 4 \\ 6d + 4 & \text{otherwise} \end{cases} \\
Q_3(d) &= \begin{cases} 3 + \frac{2}{d} & d \leq 4 \\ 3 & \text{otherwise} \end{cases}
\end{aligned}$$

and where the constants C_0 and $\tilde{C}_0$ do not depend on the measure space and noise properties.

Remark 4.2. We assume that we have access to C , C_1 , c_2 , and C_3 , but not L , which will show up in calculations but can be estimated. For instance, a rough upper estimate of L can be achieved by averaging the noisy distances from a given point to a large set of other points. This initial upper bound can be used to approximate the number of draws from μ required. We can then achieve a tighter estimate as our algorithm progresses. For details, see the proof of Lemma 4.10.

Remark 4.3. The restriction $L \geq 0$ can be removed. Observe that the initial estimate made in Remark 4.2 can be used to determine a L' such that $L + L' \geq 0$ and we can in turn replace our $d'(x, y)$ by $d'(x, y) + L'$ to shift $\mathbb{E}K_0$. Also note that this method can be used to ensure that $L = O(C)$.

We now move on to our analysis of Algorithm 2. Sections 4.1 and 4.2 describe how we approximate the objective function and how these approximations are used to obtain cluster centers. Sections 4.3 and 4.4 analyze the internal structure of the clusters and their relative locations. Sections 4.5 and 4.6 determine constraints on our constants which ensure that the set of centers is an ε -net. Section 4.7 describes how to distances between cluster centers are approximated. Finally, Theorem 4.1 is proven in Section 4.8.

4.1 Accessing Objective Functions

We have access to the objective function via

$$\tilde{g}_k(x, y) = L_{x,y} + \beta \min_{1 \leq i \leq k-1} D_i(x) + \beta \min_{1 \leq i \leq k-1} D_i(y)$$

where $L_{x,y} = \frac{1}{N_2} \sum_{z \in S_2} d'(x, z) d'(z, y)$. We want to achieve

$$|\tilde{g}_k(x, y) - g_k(x, y)| \leq \gamma$$

For this it is enough to have

$$\forall x, y \in S_1, x \neq y, |L_{x,y} - \langle f_x, f_y \rangle| \leq \frac{\gamma}{2} \quad (15)$$

$$\forall x \in S_1, i \in \{1, \dots, k-1\}, |D_i(x) - \mathbb{E} D_i(x)| \leq \frac{\gamma}{4\beta} \quad (16)$$

We first concern ourselves with achieving (15) with probability at least $1 - \frac{\theta_1}{2}$. By union bound, it is enough to have (for arbitrary $x, y \in \mathcal{M}$)

$$\Pr[|L_{x,y} - \langle f_x, f_y \rangle| > \frac{\gamma}{2}] < \frac{\theta_1}{2N_1^2} \quad (17)$$

For which it is enough to achieve (recall that $f_x(z) = \mathbb{E} d'(x, z)$)

$$\Pr[|L_{x,y} - \frac{1}{N_2} \sum_{z \in S_2} f_x(z) f_y(z)| > \frac{\gamma}{4}] < \frac{\theta_1}{4N_1^2} \quad (18)$$

$$\Pr[\frac{1}{N_2} \sum_{z \in S_2} f_x(z) f_y(z) - \langle f_x, f_y \rangle > \frac{\gamma}{4}] < \frac{\theta_1}{4N_1^2} \quad (19)$$

where it is worth noting that $\mathbb{E}_{z \sim \mu}(\mathbb{E} d'(x, z) \mathbb{E} d'(z, y)) = \langle f_x, f_y \rangle$. Using Proposition 3.4 for $t = \frac{N_2 \gamma}{4}$ and noting that $|\mathbb{E} d'(x, y)| < L + C$ and $\|d'(x, z)\|_{\psi_2} \leq C_3$ based on our assumptions, we have

$$\Pr[|L_{x,y} - \frac{1}{N_2} \sum_{z \in S_2} f_x(z) f_y(z)| > \frac{\gamma}{4}] \leq 5 \exp \left(- \frac{c N_2 \gamma^2}{256 C_3^2 (C_3^2 + (L + C)^2 + \frac{1}{4} \gamma)} \right)$$

Hence, the following is sufficient to achieve (18)

$$5 \exp \left(-\frac{cN_2\gamma^2}{256C_3^2(C_3^2 + (L+C)^2 + \frac{1}{4}\gamma)} \right) < \frac{\theta_1}{4N_1^2} \iff \frac{256C_3^2(C_3^2 + (L+C)^2 + \frac{1}{4}\gamma)}{c\gamma^2} \log \frac{20N_1^2}{\theta_1} < N_2$$

Noting that $f_x(z)f_y(z)$ varies between L^2 and $(L+C)^2$ for choices of $z \in \mathcal{M}$, Hoeffding's inequality for bounded random variables (Theorem 2.3) gives

$$\Pr \left[\left| \frac{1}{N_2} \sum_{z \in S_2} f_x(z)f_y(z) - \langle f_x, f_y \rangle \right| > \frac{\gamma}{4} \right] \leq 2 \exp \left(-\frac{N_2\gamma^2}{8((L+C)^2 - L^2)^2} \right) = 2 \exp \left(-\frac{N_2\gamma^2}{8C^2(2L+C)^2} \right)$$

Thus, in order to achieve (19) it is sufficient to have

$$2 \exp \left(-\frac{N_2\gamma^2}{8C^2(2L+C)^2} \right) < \frac{\theta_1}{4N_1^2} \iff \frac{8C^2(2L+C)^2}{\gamma^2} \log \frac{8N_1^2}{\theta_1} < N_2$$

Hence, in the case of small γ it is enough to have $N_2 > A_1(C_3^4 + (C_3^2 + C^2)(L+C)^2)\gamma^{-2} \log \frac{A_2N_1^2}{\theta_1}$ for absolute constants A_1 and A_2 . We now turn to obtaining (16). Assume that we have $\forall x \in S_1, i \in \{1, \dots, k-2\}, |D_i(x) - \mathbb{E}D_i(x)| \leq \frac{\gamma}{4\beta}$. We focus on achieving $\forall x \in S_1, |D_{k-1}(x) - \mathbb{E}D_{k-1}(x)| \leq \frac{\gamma}{4\beta}$ with probability at least $1 - \frac{\theta_1}{2k_{max}}$. By union bound, it is enough to achieve (for arbitrary $x \in \mathcal{M}$)

$$\Pr[|D_{k-1}(x) - \mathbb{E}D_{k-1}(x)| > \frac{\gamma}{4\beta}] < \frac{\theta_1}{2N_1k_{max}} \quad (20)$$

Using Theorem 2.2, we have

$$\Pr[|D_{k-1}(x) - \mathbb{E}D_{k-1}(x)| > \frac{\gamma}{4\beta}] < 2 \exp \left(-\frac{n\gamma^2}{32C_3^2\beta^2} \right)$$

Hence, achieving (20) comes down to

$$2 \exp \left(-\frac{n\gamma^2}{32C_3^2\beta^2} \right) < \frac{\theta_1}{2N_1k_{max}} \iff \frac{32C_3\beta^2}{\gamma^2} \log \frac{4N_1k_{max}}{\theta_1} < n$$

Thus, it is enough to have $n > A_3C_3\beta^2\gamma^{-2} \log \frac{A_4N_1k_{max}}{\theta_1}$ for absolute constants A_3 and A_4 . The logic behind our probability bounds will become clear in Section 4.8.

4.2 Obtaining a Cluster Center

Define $X_k^* = \arg \max_{x \in \mathcal{M}} g_k(x, x)$ and $\hat{X}_k = \arg \max_{(x,y) \in S_1^2, x \neq y} \tilde{g}_k(x, y)$. We want to ensure that any $(\hat{x}_k, \hat{y}_k) \in \hat{X}_k$ has $\max_{z \in \mathcal{M}} g_k(z, z) - g_k(\hat{x}_k, \hat{y}_k) < 4\gamma$. Suppose that S_1 is an η -net of $\mathcal{M}$ and select $x_k^* \in X_k^*$. Then there exists a pair $x, y \in S_1 \cap B_{2\eta}(x_k^*)$. Note that (for $x, y, z \in \mathcal{M}$)

$$\begin{aligned}
\langle f_z, f_z \rangle - \langle f_x, f_y \rangle &= \langle f_z, f_z \rangle - \langle f_x, f_z \rangle + \langle f_x, f_z \rangle - \langle f_y, f_z \rangle + \langle f_y, f_z \rangle - \langle f_x, f_y \rangle \\
&= \langle f_z, f_z - f_x \rangle + \langle f_y, f_z - f_x \rangle + \langle f_z, f_x - f_y \rangle \\
&\leq (L + C)C(2d(x, z) + d(x, y))
\end{aligned} \tag{21}$$

and that $y \mapsto \min_{1 \leq i \leq k-1} \mathbb{E}D_i(y)$ is a C -Lipschitz function. These observations together tell us

$$\begin{aligned}
g_k(x_k^*, x_k^*) - g_k(x, y) &\leq (L + C)C(2d(x, x_k^*) + d(x, y)) \\
&\quad + C\beta(d(x, x_k^*) + d(y, x_k^*)) \\
&\leq (L + C)C(8\eta) + C\beta(4\eta) = \eta(8(L + C)C + 4C\beta)
\end{aligned} \tag{22}$$

In order to achieve $\max_{z \in \mathcal{M}} g_k(z, z) - g_k(x, y) < 2\gamma$ it is enough to have

$$\eta(8(L + C)C + 4C\beta) < 2\gamma \iff \eta < \frac{\gamma}{4(L + C)C + 2C\beta}$$

Now we estimate N_1 such that S_1 is an η -net of $\mathcal{M}$ with probability $\geq 1 - \theta_2$. This can be done by Theorem 2.5 but here's another take which gives roughly the same answer. Take a maximal $\frac{\eta}{3}$ -packing of $\mathcal{M}$. This set is a $\frac{2\eta}{3}$ net of $\mathcal{M}$ (as otherwise you could add another ball to the packing). Hence, if you obtain at least one point in each ball, then you have a η -net of $\mathcal{M}$. Achieving this condition is an instance of the coupon collector problem and the chance that this doesn't happen after N_1 draws from μ is upper bounded by

$$\frac{1}{\text{VR}(\frac{\eta}{3})} \left(1 - \text{VR}(\frac{\eta}{3})\right)^{N_1}$$

where $\frac{1}{\text{VR}(\frac{\eta}{3})}$ is an upper bound on the number of balls and $(1 - \text{VR}(\frac{\eta}{3}))$ is an upper bound on the probability that a given ball isn't sampled on each independent draw. Hence, in order to bound this error probability by θ_2 , it is enough to have

$$N_1 \geq \frac{\log(\theta_2 \text{VR}(\frac{\eta}{3}))}{\log(1 - \text{VR}(\frac{\eta}{3}))}$$

Now, let $(\hat{x}_k, \hat{y}_k) \in \arg \max_{x, y \in S_1, x \neq y} \tilde{g}_k(x, y)$. Note that $\max_{z \in \mathcal{M}} g_k(z, z) - g(\hat{x}_k, \hat{y}_k) < 4\gamma$ because there exist $x, y \in S_1$ satisfying $\max_{z \in \mathcal{M}} g_k(z, z) - g_k(x, y) < 2\gamma$ and the only pairs (x', y') that could have risen to be in $\arg \max_{(x, y) \in S_1^2, x \neq y} \tilde{g}_k(x, y)$ are those for which $\max_{z \in \mathcal{M}} g_k(z, z) - g(x', y') < 4\gamma$. Now, observe that

$$\begin{aligned}
\max_{z \in \mathcal{M}} g_k(z, z) - g_k(\hat{x}_k, \hat{x}_k) &\geq 0 \\
\max_{z \in \mathcal{M}} g_k(z, z) - g_k(\hat{y}_k, \hat{y}_k) &\geq 0
\end{aligned} \tag{23}$$

and also that

$$\begin{aligned} \max_{z \in \mathcal{M}} g_k(z, z) - g_k(\hat{x}_k, \hat{y}_k) &= \frac{1}{2} \left(\max_{z \in \mathcal{M}} g_k(z, z) - g_k(\hat{x}_k, \hat{x}_k) \right) + \frac{1}{2} \left(\max_{z \in \mathcal{M}} g_k(z, z) - g_k(\hat{y}_k, \hat{y}_k) \right) \\ &\quad + \frac{1}{2} \|f_{\hat{x}_k} - f_{\hat{y}_k}\|_2^2 < 4\gamma \end{aligned}$$

Hence we must have

$$\|f_{\hat{x}_k} - f_{\hat{y}_k}\|_2^2 < 8\gamma \quad (24)$$

4.3 Cluster Structure Constants

We now begin an analysis of the relative locations and inner structures of our clusters. Take $\hat{x}_k$ as our k -th cluster center. We will first propose a test $x \mapsto I[\tilde{g}_k(x, \hat{y}_k) > 3\gamma]$ and show that points x within a distance of δ to $\hat{x}_k$ are accepted and those whose distance to $\hat{x}_k$ is above 5σ are rejected, both with high probability. We will then determine how to force our set $\{\hat{x}_i\}_{1 \leq i \leq k_{max}}$ to be ε -close to $\mathcal{M}$ and r -separated. The values δ , σ , and r will then be expressed in terms of ε . We start by determining δ and σ . Note that in the context of the below lemmas we already have $|\tilde{g}_k(\hat{x}_k, \hat{y}_k) - g_k(\hat{x}_k, \hat{y}_k)| < \gamma$.

Lemma 4.4. *There exists an $\delta(\gamma)$ such that for all $x \in \mathcal{M}$, if $|\tilde{g}_k(x, \hat{y}_k) - g_k(x, \hat{y}_k)| < \gamma$ and $d(\hat{x}_k, x) < \delta$ then $|\tilde{g}_k(\hat{x}_k, \hat{y}_k) - \tilde{g}_k(x, \hat{y}_k)| < 3\gamma$.*

Proof. It is enough to show

$$|g_k(\hat{x}_k, \hat{y}_k) - g_k(x, \hat{y}_k)| \leq |\langle f_{\hat{x}_k} - f_x, f_{\hat{y}_k} \rangle| + \beta |\mathbb{E}D_i(\hat{x}_k) - \mathbb{E}D_i(x)| < \gamma$$

For which it is enough to have

$$(L + C)C\delta + \beta C\delta < \gamma \iff \delta < \frac{\gamma}{C(L + C + \beta)}$$

□

Before determining σ , we need the following auxiliary lemma

Lemma 4.5. *Let $x, y \in \mathcal{M}$. If $d(x, y) > 4\xi$, then $\|f_x - f_y\|_2^2 > \frac{\xi^2}{C^2}\kappa$*

Proof. Consider the function $|f_x - f_y|^2$ on $\Lambda_{x,y}$. For any $z \in \Lambda_{x,y}$ we have $d(x, z) \geq d(y, z) + \frac{d(x,y)}{4} > d(y, z) + \xi$. Hence, $|f_x(z) - f_y(z)|^2 \geq \frac{\xi^2}{C^2}$ by the bi-lipschitzness of $f(x, y)$. Noting that $\mu(\Lambda_{x,y}) \geq \kappa$ gives the lemma. □

Lemma 4.6. *There exists a $\sigma(\gamma)$ such that for all $x \in \mathcal{M}$, if $|\tilde{g}_k(x, \hat{y}_k) - g_k(x, \hat{y}_k)| < \gamma$ and $d(\hat{x}_k, x) > 5\sigma$ then $\tilde{g}_k(\hat{x}_k, \hat{y}_k) - \tilde{g}_k(x, \hat{y}_k) > 3\gamma$.*

Proof. We want to achieve $\tilde{g}_k(\hat{x}_k, \hat{y}_k) - \tilde{g}_k(x, \hat{y}_k) > 3\gamma$. For this it is enough to have $g_k(\hat{x}_k, \hat{y}_k) - g_k(x, \hat{y}_k) > 5\gamma$. To achieve this, it is enough to have $(\max_{z \in \mathcal{M}} g_k(z, z) - 4\gamma) - g_k(x, \hat{y}_k) > 5\gamma$, which re-arranges to $\max_{z \in \mathcal{M}} g_k(z, z) - g_k(x, \hat{y}_k) > 9\gamma$. Note that this can be rewritten as

$$\frac{1}{2} \left(\max_{z \in \mathcal{M}} g_k(z, z) - g_k(x, x) \right) + \frac{1}{2} \left(\max_{z \in \mathcal{M}} g_k(z, z) - g_k(\hat{y}_k, \hat{y}_k) \right) + \frac{1}{2} \|f_x - f_{\hat{y}_k}\|_2^2 > 9\gamma$$

for which, based on (23), is it enough to achieve $\|f_x - f_{\hat{y}_k}\|_2^2 > 18\gamma$. From here we will assume that σ is large enough to satisfy

$$\left(\frac{\sigma}{4}\right)^2 \kappa > 8\gamma C^2 \quad (25)$$

such that if $d(\hat{x}_k, \hat{y}_k) > \sigma$ we have $\|f_{\hat{x}_k} - f_{\hat{y}_k}\|_2^2 > \frac{1}{C^2} \left(\frac{\sigma}{4}\right)^2 \kappa > 8\gamma$ by Lemma 4.5, which contradicts (24). Hence $d(\hat{x}_k, \hat{y}_k) < \sigma \implies d(\hat{y}_k, x) > 5\sigma - \sigma = 4\sigma$. This allows us to determine $\|f_x - f_{\hat{y}_k}\|_2^2 > \frac{\sigma^2}{C^2} \kappa$ (again using Lemma 4.5) and hence it is enough to achieve

$$\frac{\sigma^2}{C^2} \kappa > 18\gamma \iff \sigma^2 > \frac{18C^2}{\kappa} \gamma \quad (26)$$

Observe that (25) supersedes (26). Hence, it is sufficient to have

$$\sigma > \sqrt{\frac{128C^2\gamma}{\kappa}} \quad (27)$$

□

4.4 Separability and Closeness Constants

We now move on to our separability and closeness constants r and ε . We will need the following lemma and assumption.

Lemma 4.7. *If $\sigma < \frac{128C^2}{\kappa}$, then $\gamma < \sigma$*

Proof. Based on Lemma 4.6, we have $\sigma^2 > \frac{128C^2}{\kappa} \gamma \implies \frac{128C^2}{\kappa} \sigma > \frac{128C^2}{\kappa} \gamma$ and the lemma follows. □

Assumption 4.8. For all $i \in \{1, \dots, k-1\}$, $|\mathbb{E}D_i(x) - \mathbb{E}d'(x, \hat{x}_i)| < 8C\sigma$

The next lemma shows that if our clusters haven't seen the entire manifold, then we can still choose the next center to be far away from any existing center.

Lemma 4.9. *There exists an $r(\beta, \sigma)$ such that if $\max_{x \in S_1} \min_{1 \leq i \leq k-1} d(\hat{x}_i, x) > 2C^2r$, then for all $i \in \{1, \dots, k-1\}$ we have $d(\hat{x}_k, \hat{x}_i) > r$.*

Proof. Let $(\hat{x}_k^{(1)}, \hat{y}_k^{(1)})$ denote an argmax pair for $\tilde{g}$. Suppose that $\min_{1 \leq i \leq k-1} d(\hat{x}_k^{(1)}, \hat{x}_i) < r$; then

$\min_{1 \leq i \leq k-1} d(\hat{y}_k^{(1)}, \hat{x}_i) < r + \sigma$. So we have

$$\begin{aligned} \min_{1 \leq i \leq k-1} \mathbb{E} d'(\hat{x}_k^{(1)}, \hat{x}_i) &< L + Cr \\ \min_{1 \leq i \leq k-1} \mathbb{E} d'(\hat{y}_k^{(1)}, \hat{x}_i) &< L + C(r + \sigma) \end{aligned}$$

Which, noting that $|\mathbb{E} D_i(\hat{x}_k^{(1)}) - \mathbb{E} d'(\hat{x}_k^{(1)}, \hat{x}_i)| < 8C\sigma$, gives

$$\begin{aligned} \min_{1 \leq i \leq k-1} \mathbb{E} D_i(\hat{x}_k^{(1)}) &< L + C(r + 8\sigma) \\ \min_{1 \leq i \leq k-1} \mathbb{E} D_i(\hat{y}_k^{(1)}) &< L + C(r + 9\sigma) \end{aligned}$$

Observing that $\hat{x}_k^{(1)}, \hat{y}_k^{(1)} \in S_1$, we have $|\tilde{g}_k(\hat{x}_k^{(1)}, \hat{y}_k^{(1)}) - g_k(\hat{x}_k^{(1)}, \hat{y}_k^{(1)})| < \gamma$. Hence

$$\tilde{g}_k(\hat{x}_k^{(1)}, \hat{y}_k^{(1)}) < (L + C)^2 + 2L\beta + \beta C(2r + 17\sigma) + \gamma < (L + C)^2 + 2L\beta + \beta C(2r + 18\sigma)$$

Now by lemma assumption, we may choose $\hat{x}_k^{(2)}$ to satisfy

$$\min_{1 \leq i \leq k-1} d(\hat{x}_k^{(2)}, \hat{x}_i) > 2C^2r$$

and then for any $\hat{y}_k^{(2)} \in B(\hat{x}_k^{(2)}, \sigma)$, we have

$$\min_{1 \leq i \leq k-1} d(\hat{y}_k^{(2)}, \hat{x}_i) > 2C^2r - \sigma.$$

Thus, we have

$$\begin{aligned} \min_{1 \leq i \leq k-1} \mathbb{E} d'(\hat{x}_k^{(2)}, \hat{x}_i) &> L + \frac{1}{C} (2C^2r) \\ \min_{1 \leq i \leq k-1} \mathbb{E} d'(\hat{y}_k^{(2)}, \hat{x}_i) &> L + \frac{1}{C} (2C^2r - \sigma) \end{aligned}$$

Which, noting that $|\mathbb{E} D_i(\hat{x}_k^{(2)}) - \mathbb{E} d'(\hat{x}_k^{(2)}, \hat{x}_i)| < 8C\sigma$, gives

$$\begin{aligned} \min_{1 \leq i \leq k-1} \mathbb{E} D_i(\hat{x}_k^{(2)}) &> L + C(2r - 8\sigma) \\ \min_{1 \leq i \leq k-1} \mathbb{E} D_i(\hat{y}_k^{(2)}) &> L + \frac{1}{C} (2C^2r - \sigma - 8C^2\sigma) > L + C(2r - 9\sigma) \end{aligned}$$

Observing that $\hat{x}_k^{(2)}, \hat{y}_k^{(2)} \in S_1$, we have $|\tilde{g}_k(\hat{x}_k^{(2)}, \hat{y}_k^{(2)}) - g_k(\hat{x}_k^{(2)}, \hat{y}_k^{(2)})| < \gamma$. Hence

$$\tilde{g}_k(\hat{x}_k^{(2)}, \hat{y}_k^{(2)}) > L^2 + 2L\beta + \beta C(4r - 17\sigma) - \gamma > L^2 + 2L\beta + \beta C(4r - 18\sigma)$$

In order for there to be a gap between our two cases (to contradict maximality of $(\hat{x}_k^{(1)}, \hat{y}_k^{(1)})$)

we need

$$L^2 + 2L\beta + \beta C(4r - 18\sigma) > (L + C)^2 + 2L\beta + \beta C(2r + 18\sigma)$$

Re-arranging, this gives us

$$2C\beta r > (L + C)^2 - L^2 + 36C\beta\sigma$$

and then dividing by $2C\beta$ gives

$$r > \frac{1}{2\beta}(2L + C) + 18\sigma \quad (28)$$

□

Lemma 4.10. *Under Assumption 4.8 and with lower bounds on β , it is possible to design a test for (14) which forces the ultimate set of cluster points $\{\hat{x}_i\}_{1 \leq i \leq k_{max}}$ to be ε -close and ensures that $\max_{x \in S_1} \min_{1 \leq i \leq k-1} d(x, \hat{x}_i) > 2C^2r$ for all $1 \leq k < k_{max}$, allowing Lemma 4.9 to be applied.*

Proof. First, we need an estimate $\tilde{L}$ of L . Assume we have at least one cluster, which satisfies $|\mathbb{E}D_1(x) - \mathbb{E}d'(x, \hat{x}_1)| < 8C\sigma$. Set $\tilde{L} = \min_{x \in S_1} D_1(x)$. On account of S_1 being an η -net of $\mathcal{M}$, $\min_{x \in S_1} d(x, \hat{x}_1) \leq \eta < \gamma < \sigma$, with the last inequality following from Lemma 4.7. Hence $L \leq \min_{x \in S_1} \mathbb{E}d'(x, \hat{x}_1) \leq L + C\sigma$ and so $L \leq \min_{x \in S_1} \mathbb{E}D_i(x) \leq L + 9C\sigma$. Noting that $x \in S_1$, so that we have $|D_1(x) - \mathbb{E}D_1(x)| < \frac{\gamma}{4\beta} < \sigma$, we obtain $L - C\sigma \leq \tilde{L} \leq L + 10C\sigma$. Suppose, for all $x \in S_1$

$$\min_{1 \leq i \leq k-1} D_i(x) \leq \tilde{\varepsilon} = \tilde{L} + \frac{1}{C}\varepsilon - 18C\sigma - \frac{\gamma}{4\beta} \leq L + \frac{1}{C}\varepsilon - 8C\sigma - \frac{\gamma}{4\beta}$$

This implies $\min_{1 \leq i \leq k-1} \mathbb{E}D_i(x) \leq L + \frac{1}{C}\varepsilon - 8C\sigma$, which in turn gives $\min_{1 \leq i \leq k-1} \mathbb{E}d'(\hat{x}_i, x) \leq L + \frac{1}{C}\varepsilon$, which yields $\min_{1 \leq i \leq k-1} d(\hat{x}_i, x) \leq \varepsilon$ for all $x \in S_1$. Now, suppose for some $x \in S_1$

$$\min_{1 \leq i \leq k-1} D_i(x) > \tilde{\varepsilon} > L + \frac{1}{C}\varepsilon - 19C\sigma - \frac{\gamma}{4\beta}.$$

This implies $\min_{1 \leq i \leq k-1} \mathbb{E}D_i(x) > L + \frac{1}{C}\varepsilon - 19C\sigma - \frac{\gamma}{2\beta}$, which gives $\min_{1 \leq i \leq k-1} \mathbb{E}d'(\hat{x}_i, x) > L + \frac{1}{C}\varepsilon - 27C\sigma - \frac{\gamma}{2\beta}$, which yields $\min_{1 \leq i \leq k-1} d(\hat{x}_i, x) > \frac{1}{C^2}\varepsilon - 27\sigma - \frac{\gamma}{2\beta C} > \frac{1}{C^2}\varepsilon - 28\sigma$ for some $x \in S_1$.

From here we will set $2C^2r = \frac{1}{C^2}\varepsilon - 28\sigma$. To achieve this, it is enough to have $\frac{1}{2C^4}\varepsilon - \frac{14}{C^2}\sigma > \frac{1}{2\beta}(2L + C) + 18\sigma$ by (28). Re-arranging gives

$$\beta > \frac{C^4(2L + C)}{(\varepsilon - (28C^2 + 36C^4)\sigma)}$$

Hence, under the above assumption on β , the test $\max_{x \in S_1} \min_{1 \leq i \leq k-1} D_i(x) \leq \tilde{\varepsilon}$ achieves ε -closeness of the cluster centers to $\mathcal{M}$ as well as $\max_{x \in S_1} \min_{1 \leq i \leq k-1} d(x, \hat{x}_i) > 2C^2r$. □

4.5 Relationship Between Constants

In this section we make clear the relationship between δ , σ , r , and ε . We will make ε our independent variable and express δ , σ , and r in terms of it. Based on Lemma 4.10, we can set $r = \frac{1}{2C^4}\varepsilon - \frac{14}{C^2}\sigma$. If

we also set $\sigma = \frac{\varepsilon}{2(28C^2+36C^4)}$, then we obtain $\frac{14}{C^2}\sigma < \frac{14}{72C^4}\varepsilon < \frac{1}{4C^4}\varepsilon$ and thus have $\frac{1}{4C^4}\varepsilon < r < \frac{1}{2C^4}\varepsilon$.

Recall that we derived $\sigma > \sqrt{\frac{128C^2\gamma}{\kappa}} \iff \gamma < \frac{\kappa}{128C^2}\sigma^2$ in Lemma 4.6. Hence, we can set $\gamma = \frac{\kappa}{256C^2}\sigma^2$. Next, we can set $\delta = \frac{\gamma}{C(L+C+\beta)} = \frac{\kappa}{256C^3(L+C+\beta)}$ based on Lemma 4.4. If we then set $\beta = \frac{2C^4(2L+C)}{(\varepsilon-(28C^2+36C^4)\sigma)} > \frac{4C^4(2L+C)}{\varepsilon}$ based on Lemma 4.10, we obtain

$$\delta > \frac{\kappa}{256C^3\beta(L+C+1)}\sigma^2 = \frac{\kappa\varepsilon}{1024C^7(2L+C)(L+C+1)}\sigma^2 = \frac{\kappa\varepsilon^3}{4096C^7(2L+C)(L+C+1)(28C^2+36C^4)^2}$$

as well as

$$\delta < \frac{\kappa}{256C^3\beta}\sigma^2 = \frac{\kappa\varepsilon}{1024C^7(2L+C)}\sigma^2 = \frac{\kappa\varepsilon^3}{4096C^7(2L+C)(28C^2+36C^4)^2}$$

In the analysis that follows, we will assume that ε is sufficiently small. In that case, we have $r \sim \varepsilon C^{-4}$, $\sigma \sim \varepsilon C^{-4}$, $\delta \sim \varepsilon^3 \kappa C^{-15}(L+C)^{-2}$, $\gamma \sim \kappa C^{-10}\varepsilon^2$, and $\beta \sim C^4(L+C)\varepsilon^{-1}$.

We can also determine the values of N_1 , N_2 , and n , which will be used in the next section. Recall that we can set $N_1 = \frac{\log(\theta_2 \text{VR}(\frac{\eta}{3}))}{\log(1-\text{VR}(\frac{\eta}{3}))}$ where, based on Section 4.2, we can set

$$\eta = \frac{\gamma}{4(L+C)C+2C\beta} = \frac{\kappa}{128C^2(4(L+C)C+2C\beta)}\sigma^2 \sim \kappa C^{-3}\beta^{-1}\sigma^2 \sim \kappa C^{-11}\beta^{-1}\varepsilon^2 \sim \kappa C^{-15}(L+C)^{-1}\varepsilon^3$$

in the case of small ε . The assumption of d -regularity then gives us $N_1 \sim c_2^{-1}\eta^{-d}\log \frac{\eta}{\theta_2} \sim c_2^{-1}(\kappa^{-1}C^{15}(L+C))^d\varepsilon^{-3d}(\log \varepsilon^{-1} + \log \theta_2^{-1})$. Further, we have $N_2 = A_1(C_3^4 + (C_3^2 + C^2)(L+C)^2)\gamma^{-2}\log \frac{A_2N_1^2}{\theta_1} \sim \kappa^{-2}C^{20}(C_3^4 + (C_3^2 + C^2)(L+C)^2)\varepsilon^{-4}(\log \varepsilon^{-1} + \log \theta_1^{-1})$ and $n = A_3C_3\beta^2\gamma^{-2}\log \frac{A_4N_1k_{max}}{\theta_1} \sim C_3\kappa^{-2}C^{28}(L+C)^2\varepsilon^{-6}(\log \varepsilon^{-1} + \log \theta_1^{-1})$

4.6 Cluster Efficacy

From here we address how to obtain Assumption 4.8. We begin by upper bounding k_{max} . By Lemma 4.9, $\{\hat{x}_i\}_{1 \leq i \leq k_{max}}$ is r -separated. Any r -separated set is also a $\frac{1}{2}r$ -packing. If $\frac{1}{2}r < C_1$ then the volume of a $\frac{1}{2}r$ -ball in $\mathcal{M}$ is $\geq \text{VR}(\frac{1}{2}r)$. Hence, the number of balls in our r -separated set is at most $\frac{1}{\text{VR}(\frac{1}{2}r)}$. Assuming d -regularity then gives $k_{max} \lesssim c_2^{-1}r^{-d} \sim c_2^{-1}C^{4d}\varepsilon^{-d}$ in the case of small ε . We now use this bound to obtain Assumption 4.8.

Lemma 4.11. *Assume $\mathcal{M}$ is d -regular at scale C_1 . For small enough ε and large enough N_3 , it is possible to ensure $n_k = n$ and $|\mathbb{E}D_k(x) - \mathbb{E}d'(x, \hat{x}_k)| < 8C\sigma$*

Proof. Recall that S_3 is an i.i.d. sampled set of points from μ . We create auxiliary clusters $\tilde{C}_i = \{z \in S_3 : |\tilde{g}(\hat{x}_i, \hat{y}_i) - \tilde{g}(z, \hat{y}_i)| < 3\gamma\}$. Let $\tilde{C}_{i,in} = \tilde{C}_i \cap B(\hat{x}_i, 5\sigma)$ and $\tilde{C}_{i,out} = \tilde{C}_i \setminus B(\hat{x}_i, 5\sigma)$. Observe that, by a standard Chernoff bound (Theorem 2.3 in [18]),

$$\Pr(|B(x, \delta) \cap S_3| \leq \frac{1}{2}\text{VR}(\delta)N_3) \leq \exp(-\frac{1}{8}\text{VR}(\delta)N_3)$$

Consider the finite collection of balls $\{B(\hat{x}_i, \delta)\}$. The probability that there is a ball in this collection with $|B(\hat{x}_i, \delta) \cap S_3| \leq \frac{1}{2}\text{VR}(\delta)N_3$ is upper bounded by $k_{max} \exp(-\frac{1}{8}\text{VR}(\delta)N_3)$. Hence, this event happens with probability $< \theta_3$ if we have

$$k_{max} \exp(-\frac{1}{8}\text{VR}(\delta)N_3) < \theta_3 \iff \frac{8}{\text{VR}(\delta)} \log \frac{k_{max}}{\theta_3} < N_3$$

Assume that we have $|B(\hat{x}_i, \delta) \cap S_3| > \frac{1}{2}\text{VR}(\delta)N_3 \implies |B(\hat{x}_i, 5\sigma) \cap S_3| > \frac{1}{2}\text{VR}(\delta)N_3$ for all $i \in \{1, \dots, k_{max}\}$. By replicating our analysis in Section 4.1 (specifically the lines following (20)) it can be shown that, for arbitrary $x \in \mathcal{M}$, $\Pr(|D_i(x) - \mathbb{E}D_i(x)| > \frac{\gamma}{4\beta}) < \frac{\theta_1}{2N_1^\ell}$ if we set $n > \frac{32C_3\beta^2}{\gamma^2} \log \frac{4N_1^2}{\theta_1}$. Combining this with (15) allows us to bound $\Pr(|\tilde{g}_k(x, \hat{y}_k) - g_k(x, \hat{y}_k)| > \gamma) < \frac{\theta_1}{N_1^\ell}$ (noting that $D_i(\hat{y}_k)$ has already been calculated). Thus, the probability that our test fails for a given $x \in S_3$ (and hence it isn't included in $\tilde{C}_{k,in}$ when it should be or is included in $\tilde{C}_{k,out}$ when it shouldn't be) is upper bounded by $\frac{\theta_1}{N_1^\ell}$, where we will later pick $\ell \in [0, 2]$ to minimize the size of N_3 .

Note that by the same Chernoff bound we previously used, we have $\Pr(|\tilde{C}_{k,in}| \leq \frac{1}{4}(1 - \frac{\theta_1}{N_1^2})\text{VR}(\delta)N_3) \leq \exp(-\frac{1}{16}(1 - \frac{\theta_1}{N_1^2})\text{VR}(\delta)N_3)$. Thus, bounding the probability that this happens for at least one of our clusters by θ_3 comes down to

$$k_{max} \exp(-\frac{1}{16}(1 - \frac{\theta_1}{N_1^2})\text{VR}(\delta)N_3) < \theta_3 \iff \frac{16}{(1 - \frac{\theta_1}{N_1^2})\text{VR}(\delta)} \log \frac{k_{max}}{\theta_3} < N_3$$

Now, assume that we have $|\tilde{C}_{k,in}| \geq \frac{1}{4}(1 - \frac{\theta_1}{N_1^2})\text{VR}(\delta)N_3$. In order to ensure that our clusters have enough points, we assume that

$$\frac{1}{4}(1 - \frac{\theta_1}{N_1^2})\text{VR}(\delta)N_3 > n \iff N_3 > \frac{4n}{(1 - \frac{\theta_1}{N_1^2})\text{VR}(\delta)} \quad (29)$$

Using another Chernoff bound (also Theorem 2.3 of [18]), we have

$$\Pr(|\tilde{C}_{k,out}| \geq \frac{3\theta_1}{2N_1^\ell}N_3) \leq \exp\left(-\frac{\theta_1 N_3}{10N_1^\ell}\right)$$

(where we pick the parameter $0 \leq \ell \leq 2$ later) so ensuring that this does not occur for any of our clusters with probability θ_3 comes down to

$$k_{max} \exp\left(-\frac{\theta_1 N_3}{10N_1^\ell}\right) < \theta_3 \iff \frac{10N_1^\ell}{\theta_1} \log \frac{k_{max}}{\theta_3} < N_3 \quad (30)$$

Assume that $|\tilde{C}_{k,out}| < \frac{3\theta_1}{2N_1^\ell}N_3$. We produce each cluster C_i by picking n points from the associated $\tilde{C}_i$. This can be done in whatever way we want. Now we can finally estimate the approximation error of our clusters. Let $n_{k,in} = |C_i \cap \tilde{C}_{i,in}|$ and $n_{k,out} = |C_i \cap \tilde{C}_{i,out}|$ and observe that

$$|\mathbb{E}D_k(x) - \mathbb{E}d'(x, \hat{x}_k)| < \frac{n_{k,in}}{n}5C\sigma + \frac{n_{k,out}}{n}C < 5C\sigma + C\frac{3\theta_1 N_3}{2N_1^\ell n}$$

To achieve the desired upper bound of $8C\sigma$ it is enough to have (assuming our truncation of $\tilde{C}_i$ to C_i is the worst case, where $n_{k,out} = |\tilde{C}_{i,out}|$ and $|\tilde{C}_{i,out}|$ attains its max value)

$$\frac{3\theta_1 N_3}{2N_1^\ell n} < 3\sigma \iff N_3 < \frac{2\sigma N_1^\ell n}{\theta_1} \quad (31)$$

Since $\mathcal{M}$ is d-regular at scale C_1 we have

$$\varepsilon^{-\max(3d+6, 3\ell d)} (\log \varepsilon^{-1})^{1+\ell} \lesssim N_3 \lesssim \varepsilon^{-3\ell d-5} (\log \varepsilon^{-1})^{1+\ell}$$

If we pick $\ell = 1 + \frac{2}{d}$ such that $3\ell d = 3d + 6$ then we have

$$\varepsilon^{-(3d+6)} (\log \varepsilon^{-1})^{1+\ell} \lesssim N_3 \lesssim \varepsilon^{-(3d+11)} (\log \varepsilon^{-1})^{1+\ell}$$

Hence, noting that the lower bound on N_3 comes from (30) which supersedes (29) on account of the log factor being raised to a higher power, for small enough ε we can pick

$$N_3 \sim c_2^{-(1+\frac{2}{d})} (\kappa^{-1} C^{15} (L+C))^{d+2} \varepsilon^{-(3d+6)} (\log \varepsilon^{-1} + \log \theta_1^{-1})^{2(1+\frac{1}{d})}$$

□

Remark 4.12. Having both upper and lower bounds on N_3 reflects the tradeoff of producing clusters that have a low ratio of points outside $B(\hat{x}_i, 5\sigma)$ while having as little points as possible, so that calls to the D_i don't balloon in cost. Note that since we have no knowledge of the position of selected cluster points we can't naively remove errant points and must therefore assume that our truncation from $\tilde{C}_i$ to C_i is the worst case.

4.7 Cluster-Cluster Distances

We will now approximate the true distances between our cluster centers $\{\hat{x}_i\}$. For this we will use cluster-cluster distances given by

$$A(\hat{x}_i, \hat{x}_j) = \frac{1}{n^2} \sum_{x \in C_i} \sum_{y \in C_j} d'(x, y)$$

Observe that

$$|\mathbb{E}A(\hat{x}_i, \hat{x}_j) - \mathbb{E}d'(\hat{x}_i, \hat{x}_j)| < \frac{n_{i,in}n_{j,in}}{n^2} 10C\sigma + \frac{n_{i,out}n_{j,in} + n_{i,in}n_{j,out} + n_{i,out}n_{j,out}}{n^2} C$$

Noting $n_{i,in} \leq n$ gives

$$|\mathbb{E}A(\hat{x}_i, \hat{x}_j) - \mathbb{E}d'(\hat{x}_i, \hat{x}_j)| < 10C\sigma + C \frac{n_{i,out} + n_{j,out}}{n} + C \frac{n_{i,out}n_{j,out}}{n^2}$$

Recalling that $n_{i,out} < \frac{3\theta_1 N_3}{2N_1^\ell}$ and using (31) gives

$$C \frac{n_{i,out} + n_{j,out}}{n} < 6C\sigma$$

and

$$C \frac{n_{i,out} n_{j,out}}{n^2} < \frac{1}{C} (6C\sigma)^2$$

so that we obtain

$$|\mathbb{E}A(\hat{x}_i, \hat{x}_j) - \mathbb{E}d'(\hat{x}_i, \hat{x}_j)| < (16 + 36\sigma)C\sigma$$

By Theorem 2.2 we have $\Pr(|A(\hat{x}_i, \hat{x}_j) - \mathbb{E}A(\hat{x}_i, \hat{x}_j)| > \omega) \leq 2 \exp\left(-\frac{n^2 \omega^2}{2C_3^2}\right)$. In order to bound this probability by θ_4 for all pairs (i, j) with $i \neq j$ it is enough to have

$$2k_{max}^2 \exp\left(-\frac{n^2 \omega^2}{2C_3^2}\right) \leq \theta_4 \iff n^2 \geq \frac{2C_3^2}{\omega^2} \log \frac{2k_{max}^2}{\theta_4}$$

Picking $\omega = C\sigma$ then yields $n \geq \sqrt{\frac{2C_3^2}{C^2} \sigma^{-2} \log \frac{2k_{max}^2}{\theta_4}} = O(\varepsilon^{-1} \log \varepsilon^{-1})$. For small enough ε , our value of n satisfies this inequality and so with probability at least $1 - \theta_4$ we have $|A(\hat{x}_i, \hat{x}_j) - \mathbb{E}A(\hat{x}_i, \hat{x}_j)| < C\sigma$ which in turn gives

$$|A(\hat{x}_i, \hat{x}_j) - \mathbb{E}d'(\hat{x}_i, \hat{x}_j)| < (17 + 36\sigma)C\sigma < 18C\sigma$$

for small enough ε (and hence small enough σ). We now show that we can use our cluster-cluster averages to produce Oracle 3.11.

Lemma 4.13. *For a pick of $\theta_4 > 0$ and small enough ε , it is possible to produce $\mathcal{O}_1(x, y)$ that satisfies Oracle 3.11 with $Y = \{\hat{x}_i\}$.*

Proof. For $x, y, z \in Y$, observe that

$$d(x, y) - d(x, z) > 36C^2\sigma \implies \mathbb{E}d'(x, y) - \mathbb{E}d'(x, z) > 36C\sigma \implies \mathcal{O}_1(x, y) > \mathcal{O}_1(x, z)$$

Noting that $36C\sigma = 36C \frac{\varepsilon}{2(28C^2 + 36C^4)} < \varepsilon$, $\mathcal{O}_1(x, y)$ satisfies Oracle 3.11. $\square$

4.8 Proof of Theorem 4.1

Proof. Assuming $0 < \varepsilon \leq 2^{-20}$ and applying Theorem 3.12, we recover $d(x, y)$ for all $x, y \in Y = \{\hat{x}_i\}$ up to $O(\varepsilon \log \varepsilon^{-1})$. Noting that Y is a ε -net of $\mathcal{M}$, we have satisfied the distance-recovery part of the theorem. Now we just need to verify the sample complexity, runtime, and probability bounds.

We now proceed with a probabilistic analysis of our algorithm, which will allow us to set the θ_i in terms of θ . There is a set of events which need to occur in order for our algorithm to function. These events are

$$\begin{aligned}
E_1 &= \{S_1 \text{ is an } \eta\text{-net of } \mathcal{M}\} \\
E_2 &= \{\forall x, y \in S_1, x \neq y, |L_{x,y} - \langle f_x, f_y \rangle| < \frac{\gamma}{2}\} \\
E_{3,k} &= \{|B(\hat{x}_k, \delta) \cap S_3| > \frac{1}{2} \text{VR}(\delta) N_3\} \\
E_{4,k} &= \{|\tilde{C}_{k,in}| \geq \frac{1}{4} (1 - \frac{\theta_1}{N_1^2}) \text{VR}(\delta) N_3\} \\
E_{5,k} &= \{|\tilde{C}_{k,out}| < \frac{3\theta_1}{2N_1^\ell} N_3\} \\
E_{6,k} &= \{\forall x \in S_1, |D_k(x) - \mathbb{E}D_k(x)| < \frac{\gamma}{4}\} \\
E_7 &= \{\forall \{x, y\} \subset Y, |A(x, y) - \mathbb{E}A(x, y)| < C\sigma\}
\end{aligned}$$

There is an order in which we encounter these events. We first ensure that S_1 is sufficiently dense (E_1), then set up S_2 such that our inner products work out (S_2). We then proceed on to find our first cluster point and build the cluster, first ensuring that the “inner ball” around $\hat{x}_1$ is sufficiently populated in S_3 ($E_{3,1}$) and then that there are sufficiently many points in $\tilde{C}_{1,in}$ ($E_{4,1}$) and a far smaller proportion in $\tilde{C}_{1,out}$ ($E_{5,1}$) before confirming that the resultant $D_i(x)$ approximate their expectations well for $x \in S_1$ ($E_{6,1}$), which ensures that the $\tilde{g}_2(x, y)$ are good approximations of $g_2(x, y)$. We then repeat this process for the next cluster, moving to $E_{3,2}$ then $E_{4,2}$ and so on until $E_{6,k_{max}}$. Finally, we ensure that E_7 occurs to round out the algorithm. Observe that

$$\begin{aligned}
\Pr(E_1) &\geq 1 - \theta_2 \\
\Pr(E_2|E_1) &\geq 1 - \frac{\theta_1}{2} \\
\Pr(E_{3,k}|E_1, \dots, E_{6,k-1}) &\geq 1 - \frac{\theta_3}{k_{max}} \\
\Pr(E_{4,k}|E_1, \dots, E_{3,k}) &\geq 1 - \frac{\theta_3}{k_{max}} \\
\Pr(E_{5,k}|E_1, \dots, E_{4,k}) &\geq 1 - \frac{\theta_3}{k_{max}} \\
\Pr(E_{6,k}|E_1, \dots, \cap E_{5,k}) &\geq 1 - \frac{\theta_1}{2k_{max}} \\
\Pr(E_7|E_1, \dots, E_{6,k_{max}}) &\geq 1 - \theta_4
\end{aligned}$$

Hence we obtain

$$\begin{aligned}
\Pr(E_1 \cap E_2 \cap \dots \cap E_{6,k_{max}} \cap E_7) &\geq (1 - \frac{\theta_1}{2})(1 - \frac{\theta_1}{2k_{max}})^{k_{max}}(1 - \theta_2)(1 - \frac{\theta_3}{k_{max}})^{3k_{max}}(1 - \theta_4) \\
&\geq 1 - (\theta_1 + \theta_2 + 3\theta_3 + \theta_4)
\end{aligned}$$

Thus, setting $\theta_1 = \theta_2 = \frac{\theta_3}{3} = \theta_4 = \frac{\theta}{4}$ gives our desired probability bound.

We now move on to our sample size and runtime bounds. Our sample size is given by $N = N_1 + N_2 + N_3$. By d-regularity, in the case of small ε we have

$$N \sim N_3 \sim c_2^{-(1+\frac{2}{d})} (\kappa^{-1} C^{15} (L+C))^{d+2} \varepsilon^{-(3d+6)} (\log \varepsilon^{-1} + \log \theta^{-1})^{2(1+\frac{1}{d})}$$

Our runtime is $N_2(N_1^2 + k_{max}N_3) + nk_{max}(N_1 + N_3)$. The term $N_2(N_1^2 + N_3)$ results from running approximate inner products $L_{x,y}$ for all $x, y \in S_1, x \neq y$ and between all $x \in S_3$ and all $y \in \{\hat{y}_i\}$ while $nk_{max}(N_1 + N_3)$ follows from running our $D_i(x)$ over all $x \in S_1 \cup S_3$.

By d-regularity with small ε , we have

$$N_2N_1^2 \sim \kappa^{-(2d+2)} c_2^{-1} C^{30d+20} (L+C)^{2d} (C_3^4 + (C_3^2 + C^2)(L+C)^2) \varepsilon^{-6d+4} (\log \varepsilon^{-1} + \log \theta^{-1})^3 \quad (32)$$

$$nk_{max}N_3 \sim C_3 \kappa^{-(d+4)} c_2^{-2(1+\frac{1}{d})} C^{19d+58} (L+C)^{d+4} \varepsilon^{-(4d+12)} (\log \varepsilon^{-1} + \log \theta^{-1})^{3+\frac{2}{d}} \quad (33)$$

where $\lesssim$ is used instead of $\sim$ because we're using an upper bound on k_{max} . Observe that when $6d+4 \leq 4d+12 \iff d \leq 4$, (33) dominates (32), whereas when $d \leq 4$ it is the other way around. Noting that $N_2k_{max}N_3 + nk_{max}N_1$ is lower order than $N_2N_1^2 + nk_{max}N_3$, we have

$$N_{runtime} \sim Q_1(d) \varepsilon^{-Q_2(d)} (\log \varepsilon^{-1} + \log \theta^{-1})^{Q_3(d)}$$

with

$$Q_1(d) = \begin{cases} C_3 \kappa^{-(d+4)} c_2^{-2(1+\frac{1}{d})} C^{19d+58} (L+C)^{d+4} & d \leq 4 \\ \kappa^{-(2d+2)} c_2^{-1} C^{30d+20} (L+C)^{2d} (C_3^4 + (C_3^2 + C^2)(L+C)^2) & \text{otherwise} \end{cases}$$

$$Q_2(d) = \begin{cases} 4d+12 & d \leq 4 \\ 6d+4 & \text{otherwise} \end{cases}$$

$$Q_3(d) = \begin{cases} 3 + \frac{2}{d} & d \leq 4 \\ 3 & \text{otherwise} \end{cases}$$

This completes our proof □

5 Conclusion and Discussion

In this paper we develop several methods which allow us to recover the pairwise distances between points in a δ -net of a diameter 1 geodesic probability space $(\mathcal{M}, \mu, d)$ satisfying certain key constraints when one only has access to noisy distances of a very general form. Further, in the case that $\mathcal{M}$ is a d-manifold, this allow us to construct a Lipschitz approximation of $\mathcal{M}$ by applying results from [11]. If $\mathcal{M}$ is instead of finite diameter $\neq 1$, our methods allow us to approximate $\mathcal{M}$ up to some scale. So for example, if $\text{diam } \mathcal{M} = 3$ then our reconstructed pairwise distances will be those between the corresponding points on $\frac{1}{3}\mathcal{M}$. Modulo scaling, our methods achieve sample complexities and runtimes comparable to those in [13].

5.1 Discussion of Theorem 1.1 and comparison with [13]

Using Algorithm 1 and a sample size of $N \sim \varepsilon^{-2d-2} \log \varepsilon^{-1}$, we can recover the manifold up to $(\varepsilon \log \varepsilon^{-1})^{2/3}$ in time N^3 in the case of no missing data and $N^{4.5}$ in the case of missing data. This complexity and runtime is comparable to [13] (which uses $N \sim \varepsilon^{-2d} (\log \varepsilon^{-1})^5$ samples and takes time $N^{2.2}$), but we make almost no assumptions on the noise (whereas [13] needed fixed variance).

As promised in Theorem 1.2, Theorem 1.1 admits some further extensions. For instance, the mutual independence assumption can be relaxed to the one in the statement of Theorem 1.1. If we want to deal with general sub-exponential variables, we just need a sub-exponential Hoeffding inequality instead of Theorem 2.2. For variables with heavier tails, there is a trick: we can replace $d'(x, y)$ with $\max\{-L, \min\{L, d'(x, y)\}\}$, which is a bounded (hence sub-Gaussian) random variable whose expectation is close to that of $d'(x, y)$ when L is sufficiently large. And you just observe that even if f satisfies the listed properties up to a small enough additive loss, the argument will still go through. Also, we may relax $|f(x, y)| \leq C_2$ to $\max f - \min f \leq C_2$ by subtracting off the average value of $d'(x, y)$ over all $(x, y) \in X$ from d' ; as this average value belongs to $[\min f - C_2, \max f + C_2]$ with high probability, this change will enable $|f(x, y)| \leq 2C_2$. As these modifications will only add technical inconvenience to the argument while not being illustrative, we leave the details to the interested reader.

We now show the setting of [13, Theorem 1] is covered by Theorem 1.1. Recall that [13, Theorem 1] postulates a noisy distance

$$d'(x, y) = d(x, y) + \eta(x, y),$$

where $\eta(x, y) \sim \eta$ is i.i.d. In particular, we have $f(x, y) = d(x, y) + \mathbb{E}\eta$. By subtracting off $\mathbb{E}\eta$ from each noisy distance, we can assume $\mathbb{E}\eta = 0$, in which case $f(x, y) = d(x, y)$. Then we can easily check the necessary properties for f hold for $C_2 = C_3 = 1$. In addition, d' is a subexponential variable, not subgaussian as required by Theorem 1.1, but this can be fixed as mentioned in the previous paragraph. We also note that $\rho = \max\{\rho_{\max}, \frac{1}{\rho_{\min}}\}$. The diameter 1 assumption is needed to fix scaling, but in the special case $f(x, y) = d(x, y)$ we can already observe that $|A(x, y) - d(x, y)| \leq 9\varepsilon$ by a similar (in fact, more direct argument than) Theorem 3.10 and then finish the argument without appealing to Theorem 3.12; this allows us to end with $O(\varepsilon)$ loss instead of $O(\varepsilon \log \varepsilon^{-1})$ loss in Theorem 1.1. Our value of $N \sim \varepsilon^{-2d-2}$ is only slightly worse than $N \sim \varepsilon^{-2d}$ in [13, Theorem 1]. Finally, our missing data model is more general than that in [13]; indeed, in the notation of [13], we have

1. $p(x, y) \geq \frac{\lambda_1}{2} \phi_0$ whenever $d(x, y) \leq \frac{\phi_0}{2H}$;
2. we have $p(z, w) > \frac{\lambda_1}{2\lambda_2} p(x, y)$ whenever $d(z, w) < d(x, y) + \frac{1}{2\lambda_2 H} p(x, y)$.

But while our Theorem 1.1 generalizes [13, Theorem 1], we can in fact handle much more complicated distributions for $d'(x, y)$. We only need bi-Lipschitz type conditions on $f(x, y) = \mathbb{E}d'(x, y)$, with no a priori assumption on $\text{Var} d'(x, y)$ or higher moments (other than sub-Gaussian/sub-exponential). Even bi-Lipschitz type conditions on $f(x, y)$ are not too restrictive: $f(x, y)$ can be,

for example, a bi-Lipschitz function of $d(x, y)$, or it could equal $d(x, y) + q(x)$ for some $\frac{1}{2}$ -Lipschitz function of the manifold (which corresponds to observing biased distances with bias being a Lipschitz function on the manifold). Thus, we vastly improve the applicability of manifold learning techniques to a broad class of situations while not losing much in the quantitative estimates.

Our starting point was the approach used in [13] to make clusters. Specifically [13], which considered the function

$$k(y, z) := \int |d(y, x) - d(z, x)|^2 d\mu(x)$$

which can be represented as $\|r_x - r_y\|^2$, where $r_x(y) := d(x, y)$. Using properties intrinsic to the manifold, [13] showed that $k(y, z) \sim d(y, z)$, so it suffices to estimate $k(y, z)$. If we have a good estimate for $k(y, z)$, then for each $x \in X$ we can form a cluster $\mathcal{C}(x)$ of close enough points to x and use this to estimate $d(x, y)$ for all pairs (x, y) (see Figure 1 in [13]). However, the method in [13], which estimates $\int |\mathbb{E}d'(y, x) - \mathbb{E}d'(z, x)|^2 d\mu(x)$ by $\int \mathbb{E}|d'(y, x) - d'(z, x)|^2 d\mu(x)$, only works *in the case of constant variance*. To estimate $k(y, z)$, we develop a new estimator based on an argmax of inner products as in Section 3.2.1. Although norms are intimately connected to argmax (e.g. the norm in a dual Banach space is defined as an argmax), using the argmax to estimate the L_2 norm given inner product info has not been used before in our setting to the best of the authors' knowledge and could lead to new algorithms for other problems.

Once we have estimated $k(y, z)$, we can proceed as in the case of missing data of [13], which makes use of estimates for the integrals of the form

$$\begin{aligned} k_\Phi(y, z) &= \int_{\mathcal{M}} |d(y, x) - d(z, x)|^2 \Phi(y, x) \Phi(x, z) d\nu(x) \\ A_\Phi(y, z) &= \int_{\mathcal{M}} \Phi(y, x) \Phi(x, z) d\nu(x) \\ V_\Phi(y, z) &= \int_{\mathcal{M}} \psi(A_\Phi(y, x), k_\Phi(y, x)) \Phi(z, x) d\nu(x) \\ W_\Phi(y, z) &= \int_{\mathcal{M}} \psi(A_\Phi(y, x), k_\Phi(y, x)) \Phi(z, x) d\nu(x) \end{aligned}$$

where $\Phi(x, y)$ denotes the probability that $d'(x, y)$ is not missing and $x \mapsto \psi(A_\Phi(y, x), k_\Phi(y, x))$ is a smooth approximation of the indicator function for a ball around y . The techniques developed in proving Theorem 3.1 produce approximations of integrals that are related to these (with the $d(x, y)$ replaced by $f(x, y)$). The construction of the clusters $\mathcal{C}(x)$ involve estimating functions related to k_Φ and A_Φ . As opposed to integration against smooth functions $y \mapsto \psi(A_\Phi(y, x), k_\Phi(y, x))$, sums are taken over the clusters to produce $\sum_{y' \in \mathcal{C}(y)} d'(x', y') m(x', y')$ and $\sum_{y' \in \mathcal{C}(y)} m(x', y')$, which correspond with V_Φ and W_Φ . Finally, the $A(x, y)$ used in Theorem 3.1 corresponds with $\frac{1}{|Y|} m(x, z) m(y, z)$ used to define $\mathcal{C}(x)$ in Section 3.3.

Another special feature of our arguments is that we don't work with integrals. [13] needed integrals and smooth functions, whereas we only need functions defined on the sample points (i.e.

our f is only defined on $X \times X$ where X denotes the sample), so we don't need to extend the function to the entire manifold (the classical Whitney extension problem). Thus, our method can be generalized to measure metric spaces that obey key assumptions characteristic of manifolds with bounded sectional curvature, but that are not quite manifolds.

We also point out that our manifold recovery method in the case of general $f(x, y)$ is new; in particular, we hope the binary search paradigm in Theorem 3.16 can find future applications in the realm of estimating distance ratios given a comparison oracle which tells you given two distances $d(x, y)$ and $d(x, z)$, which of them is larger.

Having discussed Theorem 1.1, we now turn towards discussing the second algorithm Theorem 4.1.

5.2 Discussion of Theorem 4.1

In order to obtain well-spaced clusters in the proof of Theorem 4.1, we make use of a regularized optimization that reflects the tradeoff between maximizing $\langle f_x, f_y \rangle$ and staying away from existing clusters, which is manifest in the $\min_{1 \leq i \leq k-1} D_i(x)$ term. A more intuitive approach to avoiding existing clusters would be to exclude all points in S_1 within a τ -ball of an existing cluster from our argmax. Although we don't have access to metric balls around a given $\hat{x}_i$, we can use $\tilde{B}_i = \{D_i(x) < \tau'\}$ in place of $B(\hat{x}_i, \tau)$ for some threshold τ' . However, this approach runs into the issue that our selected argmax might lie on or close to the boundary of $\mathcal{M} \setminus \cup_{1 \leq i \leq k-1} \tilde{B}_i$ (where it is worth noting that $\mathcal{M}$ has no boundary, but this choice of optimization restriction creates one). In this case, we might be able to increase our objective function by moving into $\cup_{1 \leq i \leq k-1} \tilde{B}_i$, undermining our arguments. We could circumvent this issue by only adding points to the cluster that are also outside of the $\tilde{B}_i$. However, the $\tilde{B}_i$ are not metric balls. On account of $D_i(x)$ being the average noisy distance to points that are highly concentrated in the ball $B(\hat{x}, 5\sigma)$ but which can have highly varying distributions within this ball, the set $\tilde{B}_i$ macroscopically looks like a metric ball, but can have "swells" of order σ on its boundary. If $\hat{x}_k$ is chosen to be near the boundary of a $\tilde{B}_i$, it is difficult to lower bound $\text{vol} B(\hat{x}_i, \epsilon) \setminus \cup_{1 \leq i \leq k-1} \tilde{B}_i$ and hence ensure that we obtain cluster points.

It's also worth noting that, given the form of our $g_k(x, y)$, we need to take $\beta = O(\epsilon^{-1})$ for the analysis to work out, which forces $\delta = O(\epsilon^{-3d})$ instead of $O(\epsilon^{-2d})$, propagating to a higher sample complexity and runtime. This begs the question if a different form of regularization could circumvent this issue.

5.3 Recovering the Manifold Scale

We now return to the topic of scaling and ask: can we guess the diameter? With certain a priori information about scale this is possible. For example, if we have access to the true distance $d(x, y)$ for a one pair of points in our δ -net then the true diameter can be approximated by $\frac{d(x, y)}{\tilde{d}(x, y)}$, where $\tilde{d}(x, y)$ denotes our approximation under the assumption $\text{diam } \mathcal{M} = 1$. This is reminiscent of Eratosthenes' method for measuring the circumference of the Earth by measuring the angle of the

sun's rays in two cities at the same time of day, hence obtaining the angle between them along the surface of the Earth. This could be compared to the known distance between them on foot. In Eratosthenes' case he was modelling the Earth as a sphere, whereas in our case we can use our scale-invariant approximation of $\mathcal{M}$ as a model. In either case, we're using the shape of our space to determine its scale.

5.4 More Open Questions

1. What are the optimal sample complexity and runtime we can achieve? A theoretical lower bound on complexity is ε^{-d} , the size of an ε -net on the sphere.
2. Can we work with partial observations of the manifold, a potential large-noise version of [12]?
3. Besides manifolds, what spaces are amenable to reconstruction from noisy distances using an extension of the framework in this paper?
4. Can we extend the framework to work adaptively? A possible idea is to figure out coarse clusters, and use coarse cluster information to deduce finer clusters.

6 Omitted Proofs

Proof of Theorem 2.5. We tackle the first bullet point. By a standard Chernoff bound, the probability that $B(x, \delta)$ contains fewer than $\frac{1}{2}\text{VR}(\delta)N$ points of $X \setminus \{x\}$ is bounded by

$$\mathbb{P}\left(|B(x, \delta) \cap X| < \frac{1}{2}\text{VR}(\delta)N\right) \leq e^{-\frac{1}{8}\text{VR}(\delta)N}.$$

The ‘‘Furthermore’’ point is similar: By a standard Chernoff bound, the probability that $|\Lambda|$ contains fewer than $\frac{1}{2}\mu(\Lambda)|Y|$ points of Y is bounded by

$$\mathbb{P}\left(|\Lambda \cap Y| < \frac{1}{2}\mu(\Lambda)|Y|\right) \leq e^{-\frac{1}{8}\mu(\Lambda)|Y|}.$$

Now, we tackle the second bullet point. Let Z be a maximal $\frac{\varepsilon}{2}$ -separated set in $\mathcal{M}$; note that $|Z| \leq \frac{1}{\text{VR}(\varepsilon/4)}$ by a packing argument. Then every $B(x, r)$ contains some $B(y, \frac{r}{2})$ for $y \in Z$. By a standard Chernoff bound, the probability that $B(y, \frac{r}{2})$ contains fewer than $\frac{1}{2}\text{VR}(\frac{r}{2})|Y|$ points of Y is bounded by

$$\mathbb{P}\left(|B(y, \frac{\varepsilon}{2}) \cap X| < \frac{1}{2}\text{VR}(\frac{\varepsilon}{2})|Y|\right) \leq e^{-\frac{1}{8}\text{VR}(\frac{\varepsilon}{2})|Y|}.$$

Use union bound to finish. □

Proof of Theorem 2.6. (a) The volume of an r -ball is bounded below by $c(\Lambda, i_0, d)r^d$ by Bishop-Günther [5, Theorem III.4.2].

(b) If $d(x, y) \geq \frac{i_0}{4}$, then the set $W = B(y, \frac{i_0}{16})$ works. Otherwise, the set $W = \{x \in M; \frac{i_0}{4} < d(x, z) < \frac{i_0}{2}, \angle xyz < \pi/4\}$ works, by inspecting the proof of Case 1 in Proof of Inequality (3.8) [13].

(c) Let x^* be a point along the geodesic from x to y such that $d(x, z) = \frac{d(x, y)}{2} + u$ and $d(y, z) = \frac{d(x, y)}{2} - u$. Then $B(x^*, \frac{u}{2})$ works. $\square$

Proof of Theorem 3.4. By Theorem 2.2 applied to the independent random variables $\mathbb{E}[X_i]Y_i$, we know that

$$\Pr\left(\left|\sum_{i=1}^n (\mathbb{E}[X_i]Y_i - \mathbb{E}[X_i]\mathbb{E}[Y_i])\right| \geq \frac{t}{2}\right) \leq 2 \exp\left(-\frac{ct^2}{8 \sum_{i=1}^n \mathbb{E}[X_i]^2 K^2}\right) \leq 2 \exp\left(-\frac{ct^2}{8nK^2L^2}\right). \quad (34)$$

By independence and definition of Orlicz norm, we have

$$\mathbb{E}[\exp(\frac{\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)^2}{K^2})] \leq \prod_{i=1}^n \mathbb{E}[\exp(\frac{(Y_i - \mathbb{E}Y_i)^2}{K^2})] \leq 2^n.$$

Thus, by Markov's inequality, we have

$$\Pr\left(\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)^2 \geq K^2n + t\right) \leq e^{-\frac{t}{K^2} - n} 2^n \leq e^{-t/K^2}. \quad (35)$$

Now by Minkowski's inequality (or triangle inequality for L_2 norm), we have

$$\begin{aligned} \sum_{i=1}^n Y_i^2 &\leq \left(\sqrt{\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)^2} + \sqrt{\sum_{i=1}^n (\mathbb{E}Y_i)^2}\right)^2 \\ &\leq 2\left(\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)^2 + \sum_{i=1}^n (\mathbb{E}Y_i)^2\right) \\ &\leq 2\left(\sqrt{\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)^2} + nL\right)^2. \end{aligned}$$

Combining this with (35), we get

$$\Pr\left(\sum_{i=1}^n Y_i^2 \geq 2((K^2 + L^2)n + t)\right) \leq e^{-t/K^2}. \quad (36)$$

Finally, conditioned on a choice of $Y_1, \dots, Y_n$ satisfying $|\sum_{i=1}^n (\mathbb{E}[X_i]Y_i - \mathbb{E}[X_i]\mathbb{E}[Y_i])| \leq \frac{t}{2}$ and $\sum_{i=1}^n Y_i^2 \leq 2((K^2 + L^2)n + t)$, we have by Theorem 2.2 applied to the independent random variables

$X_i Y_i$,

$$\begin{aligned} \Pr\left(\left|\sum_{i=1}^n (X_i Y_i - \mathbb{E}[X_i] \mathbb{E}[Y_i])\right| \geq t \middle| Y_1, \dots, Y_n\right) &\leq \Pr\left(\left|\sum_{i=1}^n (X_i Y_i - \mathbb{E}[X_i] Y_i)\right| \geq \frac{t}{2} \middle| Y_1, \dots, Y_n\right) \\ &\leq 2 \exp\left(-\frac{ct^2}{8 \sum_{i=1}^n Y_i^2 K^2}\right) \\ &\leq 2 \exp\left(-\frac{ct^2}{16K^2((K^2 + L^2)n + t)}\right). \end{aligned}$$

Combining this with (34) and (36), we get

$$\begin{aligned} \Pr\left(\left|\sum_{i=1}^n (X_i Y_i - \mathbb{E}[X_i] \mathbb{E}[Y_i])\right| \geq t\right) &\leq 2 \exp\left(-\frac{ct^2}{8nK^2L^2}\right) + e^{-t/K^2} + 2 \exp\left(-\frac{ct^2}{16K^2(2K^2n + t)}\right) \\ &\leq 5 \exp\left(-\frac{ct^2}{16K^2((K^2 + L^2)n + t)}\right). \end{aligned}$$

□

References

- [1] Michael Anderson, Atsushi Katsuda, Yaroslav Kurylev, Matti Lassas, and Michael Taylor. Boundary regularity for the ricci equation, geometric convergence, and gel’fand’s inverse boundary problem. *Inventiones mathematicae*, 158(2):261–321, 2004.
- [2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.
- [3] S. N. Bernstein. On a modification of chebyshev’s inequality and of the error formula of laplace. In *Collected Works*. Academy of Sciences of the USSR, 1924. English translation available in various probability anthologies.
- [4] Leandra L Brickson, Dongwoon Hyun, Marko Jakovljevic, and Jeremy J Dahl. Reverberation noise suppression in ultrasound channel signals using a 3d fully convolutional neural network. *IEEE transactions on medical imaging*, 40(4):1184–1195, 2021.
- [5] Isaac Chavel. *Riemannian Geometry: A Modern Introduction*, volume 98. Cambridge University Press, 2006.
- [6] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.
- [7] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and

- structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.
- [8] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proceedings of the National Academy of Sciences*, 102(21):7432–7437, 2005.
 - [9] Kenneth C Creager. Anisotropy of the inner core from differential travel times of the phases pkp and pkikp. *Nature*, 356(6367):309–314, 1992.
 - [10] Maarten V de Hoop, Paul Kepley, and Lauri Oksanen. On the construction of virtual interior point source travel time distances from the hyperbolic neumann-to-dirichlet map. *SIAM Journal on Applied Mathematics*, 76(2):805–825, 2016.
 - [11] Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Reconstruction and interpolation of manifolds. i: The geometric whitney problem. *Found. Comput. Math.*, 20(5):1035–1133, October 2020.
 - [12] Charles Fefferman, Sergei Ivanov, Matti Lassas, Jinpeng Lu, and Hariharan Narayanan. Reconstruction and interpolation of manifolds ii: Inverse problems with partial data for distances observations and for the heat kernel. *American Journal of Mathematics*, 147(5):1331–1382, 2025.
 - [13] Charles Fefferman, Sergei Ivanov, Matti Lassas, and Hariharan Narayanan. Reconstruction of a riemannian manifold from noisy intrinsic distances. *SIAM Journal on Mathematics of Data Science*, 2(3):770–808, 2020.
 - [14] Peter R Hoskins. Principles of ultrasound elastography. *Ultrasound*, 20(1):8–15, 2012.
 - [15] Marko Jakovljevic, Scott Hsieh, Rehman Ali, Gustavo Chau Loo Kung, Dongwoon Hyun, and Jeremy J Dahl. Local speed of sound estimation in tissue using pulse-echo ultrasound: Model-based approach. *The Journal of the Acoustical Society of America*, 144(1):254–266, 2018.
 - [16] Alexander Kachalov, Yaroslav Kurylev, and Matti Lassas. *Inverse boundary spectral problems*. Chapman and Hall/CRC, 2001.
 - [17] Yaroslav Kurylev, Lauri Oksanen, and Gabriel P Paternain. Inverse problems for the connection laplacian. *Journal of Differential Geometry*, 110(3):457–494, 2018.
 - [18] Colin McDiarmid. *Concentration*, pages 195–248. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

- [19] Dominique Perraul-Joncas and Marina Meila. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery. *arXiv preprint arXiv:1305.7255*, 2013.
- [20] Gianmarco F Pinton, Jeremy J Dahl, and Gregg E Trahey. Rapid tracking of small displacements with ultrasound. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 53(6):1103–1117, 2006.
- [21] Jiaming Qiu and Xiongtao Dai. Estimating riemannian metric with noise-contaminated intrinsic distance. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 73983–73994. Curran Associates, Inc., 2023.
- [22] Xiaolei Qu, Takashi Azuma, Hirofumi Nakamura, Haruka Imoto, Satoshi Tamano, Shu Takagi, Shin-Ichiro Umemura, Ichiro Sakuma, and Yoichiro Matsumoto. Bent ray ultrasound tomography reconstruction using virtual receivers for reducing time cost. In *Medical Imaging 2015: Ultrasonic Imaging and Tomography*, volume 9419, pages 75–80. SPIE, 2015.
- [23] Plamen Stefanov, Gunther Uhlmann, and Andras Vasy. Boundary rigidity with partial data. *Journal of the American Mathematical Society*, 29(2):299–332, 2016.
- [24] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.