Tangent Space and Dimension Estimation with the Wasserstein Distance

Uzu Lim, Harald Oberhauser, and Vidit Nanda

Abstract. Consider a set of points sampled independently near a smooth compact submanifold of Euclidean space. We provide mathematically rigorous bounds on the number of sample points required to estimate both the dimension and the tangent spaces of that manifold with high confidence. The algorithm for this estimation is Local PCA, a local version of principal component analysis. Our results accommodate for noisy non-uniform data distribution with the noise that may vary across the manifold, and allow simultaneous estimation at multiple points. Crucially, all of the constants appearing in our bound are explicitly described. The proof uses a matrix concentration inequality to estimate covariance matrices and a Wasserstein distance bound for quantifying nonlinearity of the underlying manifold and non-uniformity of the probability measure.

1. Introduction

In this paper, we study the problem of estimating tangent spaces and the intrinsic dimension of a data manifold with high confidence. Our goal is to provide mathematically rigorous, explicit and practical bounds on the number of sample points required for such estimations. In data science terms, a tangent space gives the optimal local linear regression and the intrinsic dimension is the degree of freedom of data. Our estimators are standard applications of Local PCA, a local version of *principal component analysis* (PCA). Locally computed principal components approximate tangent spaces, and their eigenvalues allow inference of the intrinsic dimension.

To the best our knowledge, our results on *both* tangent space and dimension estimation are the first ones which simultaneously: (1) apply to noisy non-uniform distribution concentrated near a manifold, with the noise term allowed to vary across the manifold, (2) accommodate multiple data points, and (3) explicitly compute all constants appearing in the bounds, including dependence on dimension. Our proofs clearly separate the geometric and probabilistic aspects of the estimation process into modular components; we hope that the reader will find this convenient when attempting to use, build upon or improve our results. We begin by defining our estimators.

Mathematical Institute, University of Oxford, Radcliffe Observatory, Andrew Wiles Building, Woodstock Rd, Oxford OX2 6GG

E-mail addresses: lims@maths.ox.ac.uk, oberhauser@maths.ox.ac.uk, nanda@maths.ox.ac.uk.

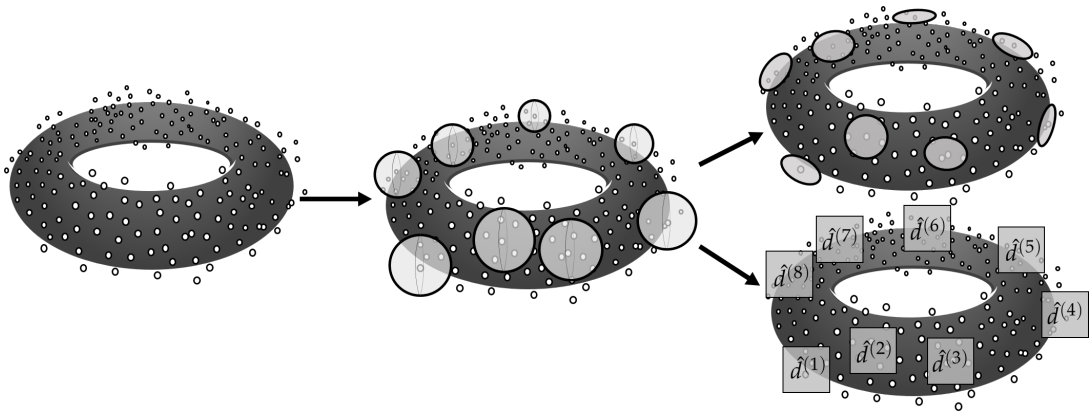

FIGURE 1. An illustration of Local PCA. Left: Dataset concentrated near a torus. Middle: Local neighborhood selection. Top Right: Tangent space estimation. Top bottom: Dimension estimation.

Estimators from Local PCA. Given m points $\mathbf{x} = \{x_1, \dots x_m\} \subset \mathbb{R}^D$, denote by $\bar{x} = \frac{1}{m}\sum_i x_i$ the mean and denote by $\hat{\Sigma}[\mathbf{x}] = \frac{1}{m}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$ the empirical covariance matrix. By PCA we mean the diagonalisation $\hat{\Sigma}[\mathbf{x}] = U\Lambda U^\top$, where U is an orthogonal matrix and Λ is a diagonal matrix. Writing $U = [v_1, \dots v_D]$ and letting diagonal entries of Λ be $\lambda_1 \geq \dots \geq \lambda_D \geq 0$, we define lower-dimensional subspaces and eigenvalues as:

$$\Pi_k[\mathbf{x}] := \text{span}(v_1, \dots, v_k)$$
$$\vec{\lambda}\hat{\Sigma}[\mathbf{x}] := (\lambda_1, \dots, \lambda_D)$$

Local PCA at an open set $W \subseteq \mathbb{R}^D$ performs PCA on points of $\mathbf{x}$ that lie in W. We are interested in W given by an open ball. Given a radius parameter $r > 0$, let $\mathbf{x}_i := \{x_j \mid j \neq i\} \cap \{y \mid \|y - x_i\| < r\}$. Define the *$k$-dimensional tangent space estimator* and the *intrinsic dimension estimator* with threshold η:

$$\begin{aligned} \hat{\Pi}(\mathbf{x}, r, i, k) :=& \Pi_k[\mathbf{x}_i] \\ \hat{d}(\mathbf{x}, r, i, \eta) :=& \text{Thr}\left(\vec{\lambda}\hat{\Sigma}[\mathbf{x}_i],\ \eta\right) \end{aligned} \tag{1.1}$$

where $\text{Thr}\left((\lambda_1, \dots \lambda_D), \eta\right)$ is the smallest k such that $(\lambda_{k+1} + \dots + \lambda_D) \leq \eta \cdot (\lambda_1 + \dots + \lambda_D)$.

When we calculate $\hat{\Pi}$ and $\hat{d}$ for a sample drawn near a d-dimensional manifold, we will get accurate estimations of tangent spaces and the intrinsic dimension d. Intuitively, this is because when a manifold is zoomed in closely enough at each point, its curvature flattens out and we essentially get a d-dimensional disk. Let's translate this intuition to precise mathematics. To do this, we precisely describe how we draw a random sample near a manifold.

Setup. Let $M \subset \mathbb{R}^D$ be a smoothly embedded d-dimensional compact manifold. Let μ_0 be a Borel probability measure on $\mathbb{R}^D$ with a probability density function $\varphi : M \to \mathbb{R}_{\geq 0}$: for each open $U \subseteq \mathbb{R}^D$, define

$$\mu_0(U) := \int_{U \cap M} \varphi \ \mathrm{d}\,\mathcal{H}^d$$

where $\mathcal{H}^d$ is the d-dimensional Hausdorff measure. Let $X \sim \mu_0$. Let Y be a $\mathbb{R}^D$-valued random variable representing noise, with bounded norm $\|Y\| \leq s$. Now our random sample $\mathbf{X} = \{X_1, \dots X_m\}$ is drawn i.i.d. from μ:

$$\mu := \mathrm{Law}(X + Y)$$

Here we emphasise that X and Y are *not assumed to be independent*. Assume that φ satisfies the Lipschitz condition $\|\varphi(x) - \varphi(y)\| \leq \alpha \cdot \mathrm{d}_M(x, y)$ for every $x, y \in M$, where d_M is the geodesic distance on M. Assume that $s < \tau$, where τ is the reach of M, defined as the maximum length to which M can be thickened normally without self-intersection.

Additionally, denote by $\omega_d = \pi^{d/2}/\Gamma(\frac{d}{2} + 1)$ the volume of the unit d-dimensional ball. Denote by $\measuredangle(\Pi_1, \Pi_2)$ the principal angle between subspaces Π_1, Π_2 (Definition 5.4). Denote by $\mathbb{P}(E)$ the probability of event E. Denote by $\varphi_{\max}, \varphi_{\min}$ the maximum and the minimum of the function φ. Our main results ensure accurate estimations if:

(1) r is small enough to ignore curvature
(2) r is big enough to ignore noise
(3) mr^d is big enough to ensure dense sampling

Main Results.

THEOREM A (Tangent Space Estimation). Let $\mathbf{X} = \{X_1, \dots X_m\}$ be a random sample as above. Given $\theta, \delta, \varrho > 0$, the following holds:

$$\sqrt{2\tau s} \leq r \leq S_1 \quad \text{and} \quad \frac{m(r - 2s)^d}{\log m} \geq S_2 \implies \mathbb{P}\left(\max_{i \leq \varrho m} \measuredangle\left(\widehat{T}_i, T_i\right) \leq \theta \right) \geq 1 - \delta$$

Here T_i is the tangent space of M at $X_i^{\perp}$, the orthogonal projection of X_i to M. $\widehat{T}_i = \hat{\Pi}(\mathbf{X}, r, i, d)$ is the tangent space estimator defined in (1.1). S_1, S_2 are defined as:

$$\begin{aligned} S_1(\tau, d, \varphi, \theta) &= \frac{c_1 \tau \sin\theta}{(d+2)} \cdot \frac{\varphi_{\min}}{3\varphi_{\min} + 8d\varphi_{\max} + 5\alpha\tau} \\ S_2(\varrho, D, d, \varphi, \theta) &= \frac{c_2 (d+2)^2}{\omega_d \varphi_{\min} \sin^2\theta} \log\left(\frac{c_3 D \varrho}{\delta}\right) \end{aligned}$$

and $c_1 = 1/16$, $c_2 = 4642, c_3 = 14$.

THEOREM B (Intrinsic Dimension Estimation). Let $\mathbf{X} = \{X_1, \dots X_m\}$ be a random sample as above. Given $\eta, \delta, \varrho > 0$ with $\eta < (2D)^{-1}$, the following holds:

$$\sqrt{2\tau s} \le r \le S_1 \quad \text{and} \quad \frac{m(r-2s)^d}{\log m} \ge S_2 \implies \mathbb{P}\Big(\hat{d}_i = d \text{ for } i \le \varrho m\Big) \ge 1 - \delta$$

where $\hat{d}_i = \hat{d}(\mathbf{X}, r, i, \eta)$ is the dimesnion estimator defined in (1.1). Here S_1, S_2 are:

$$S_1(\tau, d, \varphi, \eta) = \frac{c_1 \tau}{(d+2)D(1+\eta^{-1})} \cdot \frac{\varphi_{\min}}{3\varphi_{\min} + 8d\varphi_{\max} + 5\alpha\tau}$$

$$S_2(\varrho, D, d, \varphi, \eta) = \frac{c_2(d+2)^2 D^2 (1+\eta^{-1})^2}{\omega_d \varphi_{\min}} \log\left(\frac{c_3 D \varrho}{\delta}\right)$$

and $c_1 = 1/48$, $c_2 = 41778, c_3 = 14$.

Remarks. If φ vanishes in a small region, we may avoid division by zero by replacing $\varphi_{\min}$ by $1.04 \cdot \Phi(r-2s)$[1], where $\Phi(r) = \inf_{x \in M} \mu_0\big(U_{x,r}\big)/(\omega_d r^d)$ and $U_{x,r}$ is the set of points on M within geodesic distance r from x. Also, conditions for r given by two inequalities can be collectively replaced by one upper bound on a function Q, defined in Proposition 4.4. Lastly, we may set $r = (c \log m/m)^{1/d}$ with $c = 1.01^d S_2$ to recover the situation in [3], where in our case c is fully calculated in the main theorems[2].

1.1. Structure of the paper. Theorems A and B follow easily from Theorem 5.3 in Section 5, which is about estimating covariance matrices locally. Ingredients for its proof span Sections 2, 3, 4. In Section 2, we modify the matrix Hoeffding's inequality to show that Local PCA correctly estimates covariance (Proposition 2.6). In Section 3, we show that given two compactly supported probability measures μ, ν valued in $\mathbb{R}^D$, there is a Lipschitz relation of the form $\|\Sigma[\mu] - \Sigma[\nu]\| \le C \cdot \mathrm{W}_1(\mu, \nu)$ where $\Sigma[\mu]$ is the covariance matrix of μ (Proposition 3.3). In Section 4, we show that if a well-behaved measure on a manifold is restricted to a tiny ball, then its Wasserstein distance to the uniform measure over the unit tangential disk is small (Proposition 4.4). The Lipschitz relation in Section 3 then translates the Wasserstein bound to the bound on matrix norms.

We summarize the notations and conventions of this article in the Appendix (page 27).

1.2. Related works. The task of estimating geometric and topological quantities of manifolds from finitely many sample points lies at the crux of statistical inference, and as such the literature surrounding these topics is vast. Below we have described some of the techniques of which we are aware, and direct the reader to [36, 21, 7] for a more comprehensive survey.

[1] Φ quantifies local concentration of the measure μ_0. See Theorem 5.3.

[2] The constant 1.01 arises due to noise considerations; it comes from ensuring that $r - 2s \ge r/1.01$.

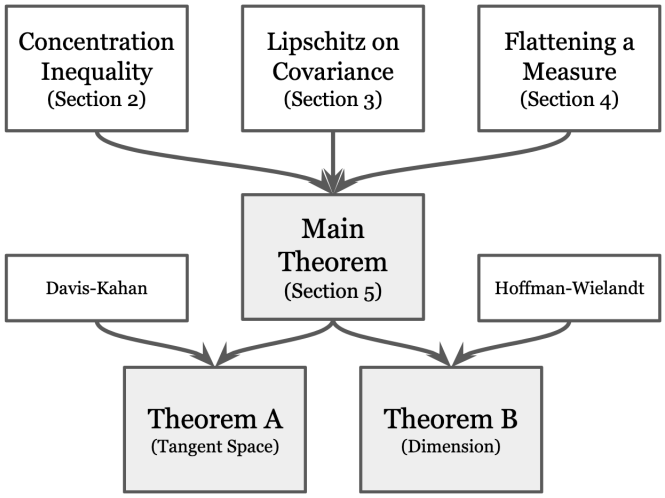

FIGURE 2. Summary of the relations between the main results.

Tangent space estimation. Probabilistic bounds on tangent space estimation using Local PCA have been studied in considerable detail, for example in [3, 32, 17, 29]. To the best of our knowledge, our work is the first in which the tangent space estimation applies to:

(1) Noisy non-uniform distribution with noise allowed to vary across the manifold,
(2) Deals with multiple data points simultaneously, and
(3) Explicitly computes all constants in bounds, including dimensional dependence.

The dimensional dependence, for example, reflects the fact that covariance of the uniform distribution over the d-dimensional unit disk have $O(1/d)$ terms (see Lemma 6.1).

In [17] and [32], the underlying probability measure is assumed to be uniform, and only estimation at a single point is considered. In [29], various constants have not been explicitly computed, and there is no consideration of noise in data distribution. In [3], various constants have not been computed explicitly, thus not specifying the minimum sample size requirement and scaling factor c for their prescription $r = (c \log m/m)^{1/d}$. Furthermore, their noise model is assumed to be orthogonal to the manifold.

Dimension estimation. The idea to use local principal component analysis for estimating intrinsic dimension is ancient, dating back at least to [12]. As such, there is a plethora of literature on the problem of estimating intrinsic dimensions. The work of [23] provides a practical and widely-used maximum likelihood estimator, but there are no known theoretical guarantees of its correctness even for synthetic data. The minimax-based estimator of [18] does come with such guarantees, but in order to compute it one is compelled to solve minimisation problems over the symmetric group on m elements (with m being the total size of the input dataset); thus, this estimator becomes intractable in practice. The recent work

of [6] introduces a far more efficient Wasserstein-based estimator with guarantees[3], but does not adapt to noise. Our efforts in this paper were motivated by the desire to find a suitable balance between practical efficiency, theoretical soundness and compatibility with noise.

Concentration inequality. Our concentration inequality for covariance matrices, Proposition 2.4, is directly derived from the matrix Hoeffding inequality in [31]. A more sophisticated approach, such as the one from [19], may be used to improve our concentration inequality. For instance, the constants appearing in Proposition 2.4 may be improved. Similar methods for analyzing (non-local, non-manifold) PCA are also studied in [20, 27].

Other Techniques. We also list related techniques that appear in other papers. A cubic bound of the form $\|\Sigma[\mu]-\Sigma[\nu]\| \leq Cr^3$, where μ, ν are probability measures supported on a ball of radius r in $\mathbb{R}^D$, is derived for uniform measures in [5]. We also obtain a similar inequality (Proposition 3.3 and Corollary 4.5). The key difference in the two derivations is that our approach uses the Wasserstein distance rather than the total variation distance from [5] to quantify similarity of measures. Our inequality has the advantage of allowing non-uniformity and of having explicit constants.

We use a transportation plan in Proposition 4.4 to quantify how much a measure supported near a manifold locally deviates from the uniform measure on a tangential disk. This transportation plan is executed with a similar idea as the proof of Proposition 3.1 in [30]. However, their transportation plan does not involve noise and applies to different types of local covariance matrices.

In [4], local polynomial regression were used to estimate manifolds and their tangent spaces from uniform point samples lying on tubular neighbourhoods. Compared to this work, our results have the advantage of not requiring the noise to be uniformly distributed. Our result only estimates tangent spaces and not higher-order information like curvature. However, the Wasserstein bound could potentially be leveraged to produce bounds on polynomial approximations.

Local PCA has been extensively used in contexts independent of the manifold hypothesis [12, 16, 33, 25], although the theoretical analysis is either heuristic or makes strong assumptions on the underlying distribution (e.g. Gaussian). Theoretical analysis in manifold learning is a flourishing field, with many significant examples including [14, 13, 2, 3, 11, 10, 18, 4, 30] and many others.

Acknowledgements.

We are grateful to Eddie Aamari, Yariv Aizenbud, Barak Sober and Hemant Tyagi for valuable discussions.

UL is supported by the Korea Foundation for Advanced Studies.

VN is supported by the EPSRC Grant EP/R018472/1.

[3]We note in passing that the number of points we require to ensure a $1-\delta$ probability of correct dimension estimation in our result is $m \sim \log(1/\delta)$, which improves on the rate $m \sim \log(1/\delta)^3$ of [6].

HO is supported by the EPSRC grant "Datasig" [EP/S026347/1], The Alan Turing Institute, and the Oxford-Man Institute.

2. Local estimation of covariance matrices

The main result of this section is Proposition 2.6, where we establish bounds for local covariance estimation. Our main tool is the *matrix Hoeffding inequality* [31, Theorem 1.3][4]. Here onwards, we will use $\|A\|$ to denote the operator norm of a given matrix A: $\|A\| := \sup_{\|x\|=1} \|Ax\|$.

THEOREM 2.1 (Matrix Hoeffding). *Let $Y_1, \dots Y_m$ be independent Hermitian random $D \times D$ matrices so that for each i we have both $\mathbb{E}Y_i = 0$ and $\|Y_i\| \leq \alpha_i$ for some real number $\alpha_i \geq 0$. Write $\sigma^2 = \sum_{i=1}^m \alpha_k^2$. Then for every $\epsilon \geq 0$,*

$$\mathbb{P}\big(\|Y_1 + \cdots + Y_m\| \geq \epsilon \big) \leq 2D \cdot \exp\left(\frac{-\epsilon^2}{8\sigma^2} \right)$$

This inequality can be used to establish concentration of vectors.[5]

COROLLARY 2.2. *Let $X_1, \dots X_m$ be independent random vectors in $\mathbb{R}^D$ satisfying $\mathbb{E}X_i = 0$, and $\|X_i\| \leq \alpha_i$ for some real number α_i. Write $\sigma^2 = \sum_1^m \alpha_i^2$. Then for every $\epsilon \geq 0$,*

$$\mathbb{P}\big(\|Y_1 + \cdots + Y_m\| \geq \epsilon \big) \leq 2(D+1) \cdot \exp\left(\frac{-\epsilon^2}{8\sigma^2} \right)$$

Throughout the remainder of this section, we fix a Borel probability measure μ on $\mathbb{R}^D$. We define some probabilistic notions.

DEFINITION 2.3. Given $X \sim \mu$, the *covariance matrix* of μ is the following $D \times D$ matrix:

$$\Sigma[\mu] := \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^\top]$$

Let δ_x be the Dirac delta measure at a point x. Given $\mathbf{x} = \{x_1, \dots x_m\} \subset \mathbb{R}^D$, define the empirical measure $\delta_{\mathbf{x}}$:

$$\delta_{\mathbf{x}} := \frac{1}{m}(\delta_{x_1} + \cdots + \delta_{x_m})$$

Given a Borel set $U \subseteq \mathbb{R}^D$, the *normalised restriction* of μ to U is defined as follows: for each Borel set $V \subset \mathbb{R}^D$,

$$\mu|_U(V) := \frac{\mu(U \cap V)}{\mu(U)}$$

We impose the convention that $\mu|_U = 0$ whenever $\mu(U) = 0$, and note that $\mu|_U$ constitues a Borel probability measure on $\mathbb{R}^D$ whenever $\mu(U) > 0$.

[4] Our version of the matrix Hoeffding inequality follows from the one in [31] by noting that for any matrix A, the operator norm $\|A\|$ equals $\max(\lambda_{\max}(A), \lambda_{\max}(-A))$ where $\lambda_{\max}$ denotes the largest eigenvalue. And moreover, $\|A\| \leq \alpha$ implies that $\alpha^2 \cdot \mathrm{Id} - A^2$ is positive definite.

[5] Apply Hermitian dilation, which takes a rectangular matrix A and produces a Hermitian matrix $A_H = \left[\begin{smallmatrix} 0 & A^\top \\ A & 0 \end{smallmatrix}\right]$. Then $\|A_H\|^2 = \|A_H^2\| = \|A\|^2$ and the result applies.

If $\mathbf{X} = (X_1, \dots X_m)$ is μ-i.i.d. sample, then $\Sigma[\delta_{\mathbf{X}}] = \frac{1}{m}\sum_{i=1}^m (X_i - \bar{X})(X_i - \bar{X})^\top$, where $\bar{X} = \frac{1}{m}\sum_i X_i$ is the sample mean. The expected value of $\Sigma[\delta_{\mathbf{X}}]$ is in fact $\frac{m-1}{m}\Sigma[\mu]$, but the following computation tells us that we may use it to estimate $\Sigma[\mu]$.

PROPOSITION 2.4 (Concentration inequalities for covariance). *Let μ be a Borel probability measure on $\mathbb{R}^D$ and let $\boldsymbol{X} = (X_1, \dots X_m)$ be an i.i.d. sample drawn from μ. Suppose that the support of μ is contained in a ball of radius r. Then for each $\epsilon \geq 0$,*

$$\mathbb{P}\big(\|\hat{\Sigma}_0 - \Sigma[\mu]\| \geq \epsilon\big) \leq 2D \cdot \exp\left(-\frac{m\epsilon^2}{512r^4}\right)$$

$$\mathbb{P}\big(\|\hat{\Sigma} - \Sigma[\mu]\| \geq \epsilon\big) \leq (4D+2) \cdot \exp\left(-\frac{m\epsilon^2}{1152r^4}\right)$$

where, denoting $\bar{X} = \frac{1}{m}\sum_i X_i$,

$$\hat{\Sigma}_0 = \frac{1}{m}\sum_{i=1}^m (X_i - \mathbb{E}X)(X_i - \mathbb{E}X)^\top, \quad \hat{\Sigma} = \frac{1}{m}\sum_{i=1}^m (X_i - \bar{X})(X_i - \bar{X})^\top$$

PROOF. We may assume that $r = 1$ without loss of generality, since for general r we know that $r^2\Sigma$ is the covariance of $r \cdot X$ for all $X \sim \mu$. Thus, we have $\|X - \mathbb{E}X\| \leq 2$ by the triangle inequality and the constraint on the support of μ. The bound for $\hat{\Sigma}_0$ is obtained directly by applying the matrix Hoeffding inequality from Theorem 2 as follows. Writing $\Sigma[\mu] = \Sigma$, set $Y_i = \frac{1}{m}((X_i - \mathbb{E}X)(X_i - \mathbb{E}X)^\top - \Sigma)$. Then $\|Y_i\| \leq (4+4)/m$ and $\sigma^2 = m \cdot (8/m)^2 = 64/m$. Since $\hat{\Sigma}_0 = \hat{\Sigma} + (\bar{X} - \mathbb{E}X)(\bar{X} - \mathbb{E}X)^\top$, we have

$$\mathbb{P}(\|\hat{\Sigma} - \Sigma\| \geq t) = \mathbb{P}(\|\hat{\Sigma}_0 - (\bar{X} - \mathbb{E}X)(\bar{X} - \mathbb{E}X)^\top - \Sigma\| \geq t).$$

Therefore, for any parameter α in $[0, 1]$, we obtain

$$\begin{aligned}
\mathbb{P}\big(\|\hat{\Sigma} - \Sigma\| \geq t\big) &\leq \mathbb{P}\big(\|\hat{\Sigma}_0 - \Sigma\| \geq \alpha t\big) + \mathbb{P}\big(\|\bar{X} - \mathbb{E}X\|^2 \geq (1-\alpha)t\big) \\
&\leq \mathbb{P}\big(\|\hat{\Sigma}_0 - \Sigma\| \geq \alpha t\big) + \mathbb{P}\left(\|\bar{X} - \mathbb{E}X\| \geq \frac{1}{2}(1-\alpha)t\right) \\
&\leq 2D \cdot \exp\left(-\frac{\alpha^2 m t^2}{512}\right) + 2(D+1) \cdot \exp\left(-\frac{(1-\alpha)^2 m t^2}{128}\right).
\end{aligned}$$

In the last inequality, we used the bound for $\hat{\Sigma}_0$ as well as Corollary 2.2, with $\sigma^2 = 4$. Choosing $\alpha = 2/3$ to make the exponents equal, we obtain the second bound. □

We will estimate $\Sigma[\mu|_U]$ with $\Sigma[\delta_{\mathbf{X}}|_U]$ assuming that U is bounded.

PROPOSITION 2.5. *Let $\mathbf{X} = (X_1, \dots X_m)$ be an i.i.d. sample drawn from μ and let $U \subseteq \mathbb{R}^D$ be a Borel set which is contained in a ball of radius r. Denote by $\hat{\Sigma}_U$ the covariance $\Sigma[\delta_{\mathbf{X}}|_U]$, and similarly write $\Sigma_U = \Sigma[\mu|_U]$. Then for any error level $\epsilon > 0$, we have that $\hat{\Sigma}_U$ estimates Σ_U:*

$$\mathbb{P}\big(\|\hat{\Sigma}_U - \Sigma_U\| \leq \epsilon\big) \geq 1 - \delta,$$

where δ is an expression such that $\lim_{m\to\infty}\delta = 0$, defined as:

$$\delta = (4D+2)(1-\mu(U)(1-\xi))^m \quad \textit{with} \quad \xi := \exp(-\epsilon^2/1152r^4).$$

PROOF. The proof follows from conditioning the membership of elements of $\mathbf{X}$ to U. Denoting by $\mathcal{S}_I$ the event $(X_i \in U \iff i \in I)$ and writing $u := \mu(U)$, we have

$$\mathbb{P}\big(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon\big) = \sum_{I\subseteq\{1,\dots m\}} \mathbb{P}\big(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon \mid \mathcal{S}_I\big)\cdot\mathbb{P}(\mathcal{S}_I).$$

Writing $|I|$ for the cardinality of each I, we have

$$\begin{aligned}
\mathbb{P}\big(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon\big) &= \sum_{I\subseteq\{1,\dots m\}} u^{|I|}(1-u)^{m-|I|}\mathbb{P}\big(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon \mid \mathcal{S}_I\big)\\
&= \sum_{k=0}^{m}\binom{m}{k}u^k(1-u)^{m-k}\mathbb{P}\big(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon \mid \mathcal{S}_{\{1,\dots k\}}\big)\\
&\leq \sum_{k=0}^{m}\binom{m}{k}u^k(1-u)^{m-k}\cdot(4D+2)\xi^k\\
&= (4D+2)\cdot(1-u(1-\xi))^m.
\end{aligned}$$

Here Proposition 2.4 was applied in the only inequality above. Note that the possibility $\mathcal{S}_\emptyset$ is correctly accounted for since we included $k=0$ when indexing the sum in the second line above. □

Now we prove the main result of this section, about estimating $\Sigma[\mu|_{U_i}]$ for open balls U_i.

PROPOSITION 2.6. *Let μ be a Borel measure supported on a compact subset $K \subset \mathbb{R}^D$, and let $\mathbf{X} = (X_1, \dots X_m)$ be a μ-i.i.d. sample. Given a radius $r > 0$, consider for $1 \leq i \leq m$ the covariances $\hat{\Sigma}_i := \Sigma[\delta_{\mathbf{X}_i}|_{U_i}]$ and $\Sigma_i = \Sigma[\mu|_{U_i}]$, where $\mathbf{X}_i = \{X_j | j \neq i\}$ and $U_i = \mathcal{B}_r(X_i)$. Let $\epsilon, \delta, \varrho > 0$ where we assume[6] that $\epsilon \leq 2r^2$. Then the following holds:*

$$\frac{m}{\log m} \geq \frac{1156r^4}{u_0\epsilon^2}\log\left(\frac{14D\varrho}{\delta}\right) \implies \mathbb{P}\left(\max_{i\leq\varrho m}\|\hat{\Sigma}_i - \Sigma_i\| \leq \epsilon\right) \geq 1-\delta$$

where $u_0 = \inf_{x\in K}\mu(\mathcal{B}_r(x)) > 0$.

PROOF. Let $k = \lfloor \varrho m \rfloor$. Define the set $E_i \subseteq (\mathbb{R}^D)^m$ as:

$$E_i := \left\{\mathbf{x} = (x_1, \cdots x_m) \mid \left\|\hat{\Sigma}[\delta_{\mathbf{x}_i}|_{U_i}] - \Sigma[\mu|_{U_i}]\right\| > \epsilon\right\}.$$

[6] We lose nothing from this assumption; suppose μ, ν are two measures supported on a single ball of radius r. Then $\|\Sigma[\mu] - \Sigma[\nu]\| \leq 2r^2$ since $\|\Sigma[\mu] - \Sigma[\nu]\| = \sup_{\|x\|=1} x^\top(\mathbb{E}_{X\sim\mu, Y\sim\nu}XX^\top - YY^\top)x = \sup_{\|x\|=1}(\langle X, x\rangle^2 - \langle Y, x\rangle^2 \leq 2r^2) \leq 2r^2$.

where $\mathbf{x}_i = \{x_j | j \neq i\}$. By the union bound, symmetry, and Proposition 2.5, we then have:

$$\begin{aligned}\mu(E_1 \cup \cdots \cup E_k) &\leq \mu(E_1) + \cdots + \mu(E_k) \\ &= k \cdot \int \mu^{k-1}\Big(\{(x_2, \cdots x_m) | (x_1, x_2, \cdots x_m) \in E_1\} \Big) \, \mathrm{d}\, \mu(x_1) \\ &\leq k \cdot \int (4D+2)(1 - u_x(1-\xi))^{m-1} \, \mathrm{d}\, \mu(x)\end{aligned}$$

where $u_x = \mu(\mathcal{B}_r(x))$, $\xi = \exp(-\epsilon^2/1152r^4)$, and μ^{k-1} is the product measure on $(\mathbb{R}^D)^{k-1}$ induced by μ. Since $0 < \xi < 1$ and $0 < u_x \leq 1$ for any x in the support K of μ, we have that $0 < u_x(1-\xi) < 1$ as well. Letting $u_0 := \inf_{x \in K} u_x$, we have:

$$\int (4D+2)k(1 - u_x(1-\xi))^{m-1} \, \mathrm{d}\, \mu(x) \leq (4D+2)k(1 - u_0(1-\xi))^{m-1} \tag{2.1}$$

Letting right hand side of (2.1) to be $\leq \delta$, we get the condition:

$$\begin{aligned}&(4D+2)k(1 - u_0(1-\xi))^{m-1} \leq \delta \\ \iff &\frac{-1}{\log\left(1 - u_0(1-\xi)\right)} \cdot \log\left(\frac{(4D+2)k}{\delta}\right) \leq m - 1 \end{aligned} \tag{2.2}$$

To produce a simpler lower bound for m, we calculate:

$$\frac{-1}{\log\left(1 - u_0(1-\xi)\right)} \leq \frac{1}{u_0}\left(\frac{1152r^4}{\epsilon^2} + 1\right) - \frac{1}{2} \leq \frac{1}{u_0} \cdot \frac{1156r^4}{\epsilon^2} - \frac{1}{2}$$

where the first inequality is due to Lemma 6.8, and the second inequality follows from the assumption that $\epsilon^2 \leq 4r^4$.[7] Using the fact that $\log((4D+2)/\delta) \geq 2$ and Lemma 6.6, we obtain the claimed sufficient condition for (2.2):

$$\frac{1156r^4}{u_0\epsilon^2} \log\left(\frac{14D\varrho}{\delta}\right) \leq \frac{m}{\log m}$$

To establish that $u_0 > 0$, consider the covering of K by balls of radius $r/2$. Since K is compact, it admits a subcover $\{\mathcal{B}_{r/2}(x) \mid x \in J\}$, with J a finite set. Thus, every $x \in K$ admits a $y \in J$ satisfying $x \in \mathcal{B}_{r/2}(y)$. Triangle inequality guarantees that $\mathcal{B}_{r/2}(y) \subseteq \mathcal{B}_r(x)$, so that $\mu(\mathcal{B}_{r/2}(y)) \leq \mu(\mathcal{B}_r(x))$ and hence $\inf_{y \in J} \mu(\mathcal{B}_{r/2}(y)) \leq \inf_{x \in K} \mu(\mathcal{B}_r(x))$. Since the left hand side is an infimum over a finite set of strictly positive numbers, it is also strictly positive and we have $u_0 > 0$ as desired. □

3. Lipschitz property of covariance matrix

Our goal in this section is to outline sufficient conditions under which the assignment $\mu \mapsto \Sigma[\mu]$ becomes a Lipschitz function with respect to the Wasserstein distance [34] on its

[7] By similar reasoning, the left hand side of (2.2) is at least $\frac{1}{u_0}(1150r^4/\epsilon^2)$, so that this sufficient condition doesn't weaken the bound much.

domain, defined as follows. Let (M, d_M) be a Polish metric space equipped with probability measures μ and ν. For each $p \geq 1$, the *p-Wasserstein distance* between μ and ν equals

$$\mathrm{W}_p(\mu, \nu) := \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{M \times M} \mathrm{d}_M(x, y)^p \, \mathrm{d}\, \gamma(x, y) \right)^{1/p}$$

where $\Pi(\mu, \nu)$ is the set of measures on $M \times M$ with marginals equal to μ and ν. Note that whenever $1 \leq p \leq q$, we have $\mathrm{W}_p(\mu, \nu) \leq \mathrm{W}_q(\mu, \nu)$ by the power mean inequality. Throughout this section, we use the notation $X \sim \mu$ and $Y \sim \nu$, whenever probability distributions μ, ν are defined.

LEMMA 3.1. *Given Borel probability measures μ, ν valued in $\mathbb{R}^D$, define $\tilde{\mu} = \mathrm{Law}(X - \mathbb{E}X)$ and similarly $\tilde{\nu}$. Then for each $p \geq 1$,*

(1) $\|\mathbb{E}X - \mathbb{E}Y\| \leq \mathrm{W}_p(\mu, \nu)$
(2) $\mathrm{W}_p(\tilde{\mu}, \tilde{\nu}) \leq 2 \cdot \mathrm{W}_p(\mu, \nu)$

PROOF. Defining $x_0 := \mathbb{E}X$ and $y_0 := \mathbb{E}Y$, we have

$$\begin{aligned}
\|x_0 - y_0\| &= \left\| \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} (x - y) \, \mathrm{d}\, \mu(x) \, \mathrm{d}\, \nu(y) \right\| \\
&= \left\| \int_{\mathbb{R}^D \times \mathbb{R}^D} (x - y) \, \mathrm{d}\, \gamma(x, y) \right\|, \text{ for any } \gamma \in \Pi(\mu, \nu) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \left\| \int_{\mathbb{R}^D \times \mathbb{R}^D} (x - y) \, \mathrm{d}\, \gamma(x, y) \right\| \\
&\leq \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \|x - y\| \, \mathrm{d}\, \gamma(x, y) \\
&= \mathrm{W}_1(\mu, \nu)
\end{aligned}$$

Noting that $W_1(\mu, \nu) \leq W_p(\mu, \nu)$ for any $p \geq 1$, we get the first claim. For the second claim,

$$\begin{aligned}
\mathrm{W}_p(\tilde{\mu}, \tilde{\nu})^p &= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \|(x - x_0) - (y - y_0)\|^p \, \mathrm{d}\, \gamma(x, y) \\
&= 2^p \cdot \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \left(\frac{\|x - y\| + \|x_0 - y_0\|}{2} \right)^p \mathrm{d}\, \gamma(x, y) \\
&\leq 2^p \cdot \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \frac{\|x - y\|^p + \|x_0 - y_0\|^p}{2} \, \mathrm{d}\, \gamma(x, y) \\
&= 2^{p-1}(\mathrm{W}_p(\mu, \nu)^p + \|x_0 - y_0\|^p) \\
&\leq 2^p \cdot \mathrm{W}_p(\mu, \nu)^p
\end{aligned}$$

where the first inequality is the power mean inequality, and the second inequality follows from the first claim. □

LEMMA 3.2. *For probability measures μ, ν defined on $\mathbb{R}$ and supports contained the interval $[-R, +R]$, we have the $2R$-Lipschitz relation for all $p \geq 1$:*

$$\mathbb{E}[X^2] - \mathbb{E}[Y^2] \leq 2R \cdot \mathrm{W}_p(\mu, \nu)$$

PROOF. Since W_p is increasing in p, it suffices to prove the assertion for $p = 1$.

$$\begin{aligned}
\mathbb{E}[X^2] - \mathbb{E}[Y^2] &= \int_{\mathbb{R}} \int_{\mathbb{R}} (x^2 - y^2) \,\mathrm{d}\,\mu(x) \,\mathrm{d}\,\nu(y) \\
&= \int_{\mathbb{R}\times\mathbb{R}} (x^2 - y^2) \,\mathrm{d}\,\gamma(x, y), \text{ for any } \gamma \in \Pi(\mu, \nu) \\
&\leq 2R \cdot \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \,\mathrm{d}\,\gamma(x, y) \\
&= 2R \cdot \mathrm{W}_1(\mu, \nu)
\end{aligned}$$

where the only inequality above follows from the fact that the derivative of $f(x) = x^2$ is bounded by $2R$ if $x \in [-R, +R]$. □

PROPOSITION 3.3. *Suppose μ, ν are probability measures on $\mathbb{R}^D$ such that each measure comes with a ball of radius r that contains the support of the measure. Then for $p \geq 1$, we have the following Lipschitz property:*

$$\|\Sigma[\mu] - \Sigma[\nu]\| \leq 4r \cdot \mathrm{W}_p(\tilde{\mu}, \tilde{\nu}) \leq 8r \cdot \mathrm{W}_p(\mu, \nu)$$

where $\tilde{\mu} = \mathrm{Law}(X - \mathbb{E}X)$.

PROOF. We assume that $r = 1$, since the case for general r follows by scaling: r affects the covariance matrix on the order of r^2 and the Wasserstein distance on the order of r. Also, the second inequality follows from the first by Lemma 3.1, so it suffices to show the first inequality. Since we are then working with $\tilde{\mu}$ and $\tilde{\nu}$ and since covariance matrix is invariant under translation, we may rewrite $\mu = \tilde{\mu}$ and $\nu = \tilde{\nu}$ and assume that μ, ν have zero means. We may also assume that both $\operatorname{supp}\mu$ and $\operatorname{supp}\nu$ are contained within $\mathcal{B}_2(0)$ by the triangle inequality; there is a ball $\mathcal{B}_1(x)$ of radius 1 containing $\operatorname{supp}\mu$, so that by triangle inequality, $\operatorname{supp}\mu \subseteq \mathcal{B}_1(x) \subseteq \mathcal{B}_2(0)$.

Denoting $S := \Sigma[\mu] - \Sigma[\nu]$, it is a real symmetric matrix and we may diagonalise it as $S = U\Lambda U^\top$. $U = [u_1, \dots u_D]$ is orthogonal and Λ is a diagonal matrix with entries $\lambda_1 \geq \cdots \geq \lambda_D$. The operator norm of S is $\max_i |\lambda_i|$, which can be written as:

$$\begin{aligned}
\|S\| &= \max_i |\lambda_i| = \max_i \left|(U^\top S U)_{i,i}\right| \\
&= \max_i \left|\mathbb{E}[U^\top X X^\top U]_{i,i} - \mathbb{E}[U^\top Y Y^\top U]_{i,i}\right| \\
&= \max_i \left|\mathbb{E}(U^\top X)_i^2 - \mathbb{E}(U^\top Y)_i^2\right|
\end{aligned}$$

where $A_{i,i}$ refers to the (i,i)th entry of a matrix A and w_i refers to the ist entry of a vector w. Now we are done by the following that holds for all i:

$$\begin{aligned}
\mathbb{E}(U^\top X)_i^2 - \mathbb{E}(U^\top Y)_i^2 &\leq 4\,\mathrm{W}_1((U^\top \mu)_i, (U^\top \nu)_i) \\
&\leq 4\,\mathrm{W}_1(U^\top \mu, U^\top \nu) \\
&= 4\,\mathrm{W}_1(\mu, \nu)
\end{aligned}$$

where $U^\top \mu = \mathrm{Law}(U^\top X)$ and $(U^\top \mu)_i$ denotes the marginal of $U^\top \mu$ at its ith coordinate. The first inequality is Lemma 3.2 with $2R = 4$. The second inequality is a general fact that applies to the Wasserstein distances between marginals. The last equality follows from the fact that the Wasserstein distance is invariant with respect to isometry applied simultaneously to the two measures. Finally, multiplying by the Lipschitz constant 2 for the non-centered measures, we get the Lipschitz constant 8. The inequality for other p follows since W_p is increasing in p. □

4. Wasserstein bound for Flattening a Measure on Manifold

In this section, we quantify the extent to which a probability distribution valued near a manifold approximates the uniform distribution over a tangential disk, using the Wasserstein distance. We first define the measure of interest using a probability density function, Hausdorff measure, and a noise term.

DEFINITION 4.1. Given a metric space and a positive integer d, denote by $\mathcal{H}^d$ the d-dimensional Hausdorff measure [28] on the metric space:

$$\mathcal{H}^d(U) = \lim_{\delta \downarrow 0} \mathcal{H}^d_\delta(U), \quad \mathcal{H}^d_\delta(U) = \frac{\omega_d}{2^d} \inf_{\substack{\mathrm{diam}(C_j)<\delta \\ U \subseteq \cup C_j}} \left(\sum_{j=1}^{\infty} \mathrm{diam}(C_j)^d \right)$$

where $\omega_d := \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$. Given a Borel set $U \subseteq \mathbb{R}^D$ with a finite, nonzero real d-dimensional Hausdorff measure $\mathcal{H}^d(U) \in (0, \infty)$, denote by $\mathrm{Unif}_d(U)$ the d-dimensional uniform probability measure over U with respect to $\mathcal{H}^d$; for each V,

$$\mathrm{Unif}_d(U) := \mathcal{H}^d|_U, \text{ i.e. } \mathrm{Unif}_d(U)(V) = \frac{\mathcal{H}^d(U \cap V)}{\mathcal{H}^d(U)}$$

DEFINITION 4.2. Suppose M is a d-dimensional smooth compact manifold with a smooth embedding into $\mathbb{R}^D$ and $\varphi : M \to \mathbb{R}^+$ is a continuous function satisfying $\int_M \varphi \, \mathrm{d}\, \mathcal{H}^d = 1$. Let μ_0 be the Borel probability measure given by defining for each open $U \subseteq \mathbb{R}^D$ the following:

$$\mu_0(U) = \int_{U \cap M} \varphi \, \mathrm{d}\, \mathcal{H}^d$$

Let $s \geq 0$ be a constant, $X \sim \mu_0$ and let Y be a random variable valued in $\mathbb{R}^D$ with bounded norm $\|Y\| \leq s$. Here X and Y are *not* assumed to be independent. Define

$$\mu := \mathrm{Law}(X + Y)$$

Then $\mathcal{P}(M, s)$ is defined as the set of all such pairs (μ_0, μ), given M and s.

The following are notions from differential geometry relevant to us.

DEFINITION 4.3. For each compact Riemannian manifold $M \subset \mathbb{R}^D$,

(1) For each $x, y \in M$, let $\mathrm{d}_M(x,y)$ be the length of the shortest geodesic connecting x and y.[8]
(2) The *reach* τ of M is the supremum of $t \geq 0$ satisfying the following: If $x \in \mathbb{R}^D$ satisfies $\mathrm{d}_{\mathbb{R}^D}(x, M) \leq t$, then there is a unique point $x_\perp \in M$ such that $\mathrm{d}_{\mathbb{R}^D}(x, x_\perp) = \mathrm{d}_{\mathbb{R}^D}(x, M)$. Here, $\mathrm{d}_{\mathbb{R}^D}(x,y) = \|x - y\|$ is the Euclidean distance on $\mathbb{R}^D$, and $\mathrm{d}_{\mathbb{R}^D}(x, M) = \inf_{y \in M} \mathrm{d}_{\mathbb{R}^D}(x,y)$.
(3) For each point $x \in M$, we denote by $\mathring{\mathcal{B}}_r \subseteq T_x M$ the open ball of radius r around $0 \in T_x M$, while the notation $\mathcal{B}_r(x) \subseteq \mathbb{R}^D$ is reserved for the (usual) open ball of radius r around $x \in \mathbb{R}^D$.
(4) Given $x \in M$, the *exponential map* $\exp_x$ sends each $v \in T_x M$ to the endpoint of the unique geodesic on M starting at x with the initial velocity of v.

We remark that $1/\tau$ is an upper bound of the acceleration of geodesics on M in the ambient space $\mathbb{R}^D \supset M$. The following is the main result of this section.

PROPOSITION 4.4. *Let $(\mu_0, \mu) \in \mathcal{P}(M, s)$ where $M \subseteq \mathbb{R}^D$ is a compact smoothly embedded d-dimensional manifold with reach τ and $s \geq 0$. Let $x \in \operatorname{supp} \mu$, let $x_\perp$ be any point in $\mathcal{B}_s(x) \cap M$, and let r be a number satisfying $2s \leq r \leq (\sqrt{2}-1)\tau - 2s$. Then there exists a function Q so that the following holds for any $p \geq 1$:*

$$\mathrm{W}_p(\nu, \tilde{\nu}) \leq \tau \cdot Q\left(\frac{r}{\tau}, \frac{s}{\tau}\right)$$
$$\text{where } \nu := \mu|_{\mathcal{B}_r(x)}, \text{ and } \tilde{\nu} := \mathrm{Unif}_d(\mathcal{B}_r(x_\perp) \cap T_{x_\perp} M)$$

Furthermore, we may take:

$$Q(\rho,\sigma) = 3\sigma + (\rho+2\sigma)^2 + \frac{1.2\varphi_{\max}}{\Phi}(2\rho + (\rho+2\sigma)^2)(1-\Omega^d) + \frac{2.2\rho}{\Phi}(\varphi_{\max} - \varphi_{\min}) + 1.4\rho^3$$

where $\varphi_{\max}, \varphi_{\min}$ are extrema of φ taken over $\mathcal{B}_{r+2s}(x_\perp)$ and

$$\Phi = \Phi(x_\perp, r-2s) := \frac{\mu_0(\exp_{x_\perp} \mathring{\mathcal{B}}_{r-2s})}{\omega_d (r-2s)^d}, \text{ and } \Omega := \frac{\rho - 2\sigma}{(\rho+2\sigma) + (\rho+2\sigma)^2}$$

PROOF. We consider the following multi-step transportation plan (see Figure 4), from $\nu_0 := \nu$, going through $\nu_1, \nu_2, \nu_3, \nu_4$ which we define below and finally reaching $\nu_5 := \tilde{\nu}$. Informally, these steps can be summarized as

(1) Perform a naive denoising on ν_0 to get ν_1
(2) Apply inverse exponential map to get ν_2
(3) Fold in the portion of ν_2 on the outer rim to the inside to get ν_3
(4) Flatten out the nonuniformity and get ν_4.
(5) Rescale radius uniformly to get ν_5.

[8] Equivalently, $\mathrm{d}_M(x,y)$ be the infimum of lengths of all piecewise regular curves that connect x and y. This follows from the Hopf-Rinow Theorem; see Corollary 6.21 and 6.22 in [22].

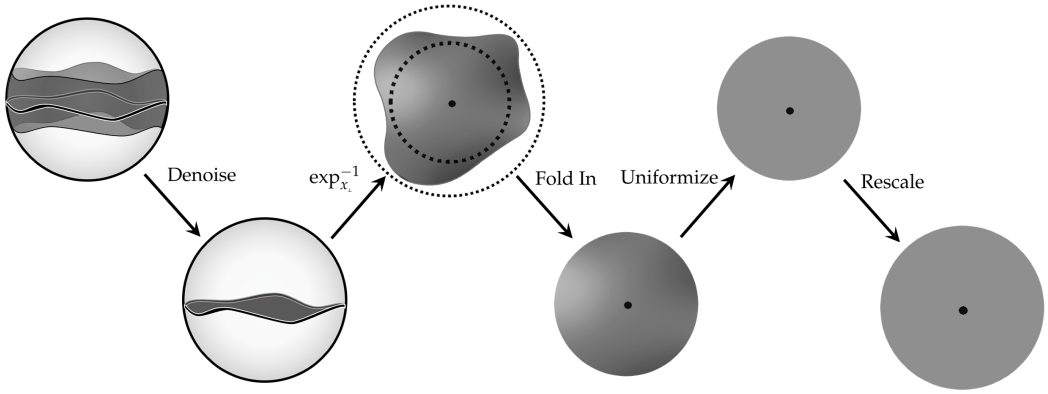

FIGURE 3. An overview of the transportation plan in the proof of Proposition 4.4. The last four sub-diagrams take place on the tangent space. Nonuniform shadings in the 3rd, 4th sub-diagrams indicate nonuniform probability distribution.

Step 1. Suppose that $X \sim \mu_0$ and $(X+Y) \sim \mu$. We define $\nu_1 := \mathrm{Law}(X \mid X+Y \in \mathcal{B}_r(x))$ and define the transportation plan ν_{01} by $\nu_{01} := \mathrm{Law}((X+Y, X) \mid X+Y \in \mathcal{B}_r(x))$, whose marginals are ν_0 and ν_1. Thus for each open $U \subseteq \mathbb{R}^D$, we have

$$\begin{aligned}\nu_1(U) =& \mathbb{P}(X \in U \mid X+Y \in \mathcal{B}_r(x)) \\ =& \frac{1}{\mu(\mathcal{B}_r(x))} \mathbb{P}(X \in U \text{ and } X+Y \in \mathcal{B}_r(x)) \end{aligned} \tag{4.1}$$

where $\mu(\mathcal{B}_r(x)) = \mathbb{P}(X+Y \in \mathcal{B}_r(x))$, which follows by the definition of μ. The transportation cost is bounded as $\mathrm{W}_p(\nu_0, \nu_1) \le \mathbb{E}_{(X+Y,X)\sim\nu_{01}} \|(X+Y)-X\| \le s$. Note that by the assumption $x \in \operatorname{supp} \mu$, we have $\mu(\mathcal{B}_r(x)) > 0$ and thus we are not conditioning on the null event.

By Equation (4.1), ν_1 is well understood in regions where the condition $X+Y \in \mathcal{B}_r(x)$ either always or never holds. If $X \in \mathcal{B}_{r-s}(x)$, then since $\|Y\| \le s$, the triangle inequality implies $X+Y \in \mathcal{B}_r(x)$. Similarly if $X \notin \mathcal{B}_{r+s}(x)$, then $X+Y \notin \mathcal{B}_r(x)$. By also noting that $\|x - x_\perp\| \le s$, the triangle inequality once again implies $\mathcal{B}_{r-2s}(x_\perp) \subseteq \mathcal{B}_{r-s}(x)$ and $\mathcal{B}_{r+s}(x) \subseteq \mathcal{B}_{r+2s}(x_\perp)$. Applying Equation (4.1), we get the following:

$$\begin{aligned} \nu_1(U) &\le \frac{\mu_0(U)}{\mu(\mathcal{B}_r(x))} && \text{for any } U \\ \nu_1(U) &= \frac{\mu_0(U)}{\mu(\mathcal{B}_r(x))} && \text{for } U \subseteq \mathcal{B}_{r-2s}(x_\perp) \\ \nu_1(U) &= 0 && \text{for } U \subseteq \mathcal{B}_{r+2s}(x_\perp)^{\mathrm{c}} \end{aligned} \tag{4.2}$$

where A^{c} denotes the complement of a set A. Note that $\mu(\mathcal{B}_r(x))$ is a constant, since we fixed x.

Step 2. We define ν_2 by pushing forward ν_1 along the inverse of the exponential map $\exp_{x_\perp}$, but we must do it where the exponential is invertible. The injecvitity radius is defined

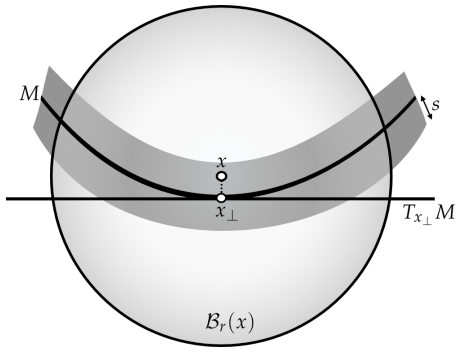

FIGURE 4. Measure μ and its restriction $\mu|_{\mathcal{B}_r(x)}$, where $x \in \mathbb{R}^D$ and $x_\perp \in M$.

as the largest radius ι so that for any $z \in M$, $\exp_z$ is a diffeomorphism (and thus invertible) when restricted to the ball of radius ι centered at $0 \in T_z M$. It is known that the injectivity radius is at least $\pi \cdot \tau$ (Proposition A1 of [1]). Meanwhile, Lemma 6.11 implies the following inclusions, which tell us our domains of interest:

$$\begin{aligned} \exp_{x_\perp}(\mathring{\mathcal{B}}_{r_{\text{in}}}) &\subseteq \mathcal{B}_{r-2s}(x_\perp) \cap M \\ \mathcal{B}_{r+2s}(x_\perp) \cap M &\subseteq \exp_{x_\perp}(\mathring{\mathcal{B}}_{r_{\text{out}}}) \end{aligned} \tag{4.3}$$

where $\mathring{\mathcal{B}}_r$ is the open ball of radius r in $T_{x_\perp}M$ centered at 0, and the radii $r_{\text{in}}, r_{\text{out}}$ are defined as:

$$\begin{aligned} r_{\text{in}} &:= r - 2s \\ r_{\text{out}} &:= (r + 2s) + (r + 2s)^2/\tau \end{aligned} \tag{4.4}$$

Now $r + 2s \leq (\sqrt{2} - 1)\tau$ implies $r_{\text{out}} \leq \pi\tau$, and thus the exponential map is invertible on $\mathcal{B}_{r+2s}(x_\perp) \cap M$. Therefore, noting Equation (4.2), we may define ν_2 as follows:

$$\nu_2 := (F^{-1})_* \nu_1, \text{ where } F = \exp_{x_\perp}|_{\mathring{\mathcal{B}}_{r_{\text{out}}}}$$

Or equivalently,

$$\nu_2(U) = \nu_1(\exp_{x_\perp}(U \cap \mathring{\mathcal{B}}_{r_{\text{out}}}))$$

Note that the support of ν_2 is contained in $F^{-1}(\mathcal{B}_{r+2s}(x_\perp))$ by the definition of ν_2 and Equation (4.2). We also have $F^{-1}(\mathcal{B}_{r+2s}(x_\perp)) \subseteq \mathring{\mathcal{B}}_{r_{\text{out}}}$ by Equation (4.3).

The transportation plan is the application of Lemma 6.3 to the pushforward along $\exp_{x_\perp}^{-1}$. In performing the transportation, we regard the tangent space as embedded: $T_{x_\perp}M \subseteq \mathbb{R}^D$ so that the transportation happens in the ambient space $\mathbb{R}^D$. By the last result mentioned in Lemma 6.10, the transportation cost then is bounded as:

$$\mathrm{W}_p(\nu_1, \nu_2) \leq \frac{(r + 2s)^2}{\tau}$$

Thus by Equations (4.2) and (4.3),

$$\begin{aligned}
\nu_2(U) &\le \frac{\mu_0(\exp_{x_\perp} U)}{\mu(\mathcal{B}_r(x))} && \text{for } U \subseteq \mathring{\mathcal{B}}_{r_{\text{out}}} \\
\nu_2(U) &= \frac{\mu_0(\exp_{x_\perp} U)}{\mu(\mathcal{B}_r(x))} && \text{for } U \subseteq \mathring{\mathcal{B}}_{r_{\text{in}}} \\
\nu_2(U) &= 0 && \text{for } U \subseteq (\mathring{\mathcal{B}}_{r_{\text{out}}})^{\text{c}}
\end{aligned} \tag{4.5}$$

Meanwhile, we can evaluate $\mu_0(U)$ when $U \subseteq \mathring{\mathcal{B}}_{r_{\text{out}}}$ explicitly using the area formula from geometric measure theory[9], which is a generalization of chain rule:

$$\mu_0(\exp_{x_\perp}(U)) = \int_{\exp_{x_\perp}(U)} \varphi \,\mathrm{d}\,\mathcal{H}^d = \int_U \varphi(\exp_{x_\perp} y)\,\mathrm{J}\exp_{x_\perp}(y)\,\mathrm{d}\,y$$

Here, $\mathrm{J}\,f$ denotes the Jacobian of a function f and $\mathrm{d}\,y$ is the d-dimensional Lebesgue measure. Thus,

$$\begin{aligned}
\nu_2(U) &\le \frac{1}{\mu(\mathcal{B}_r(x))} \int_U \varphi(\exp_{x_\perp} y)\,\mathrm{J}\exp_{x_\perp}(y)\,\mathrm{d}\,y && \text{for } U \subseteq \mathring{\mathcal{B}}_{r_{\text{out}}} \\
\nu_2(U) &= \frac{1}{\mu(\mathcal{B}_r(x))} \int_U \varphi(\exp_{x_\perp} y)\,\mathrm{J}\exp_{x_\perp}(y)\,\mathrm{d}\,y && \text{for } U \subseteq \mathring{\mathcal{B}}_{r_{\text{in}}} \\
\nu_2(U) &= 0 && \text{for } U \subseteq (\mathring{\mathcal{B}}_{r_{\text{out}}})^{\text{c}}
\end{aligned} \tag{4.6}$$

Step 3. We saw that ν_2 can be written in terms of μ_0 inside radius r_{in} and vanishes outside radius r_{out}. The annular region between the two radii is harder to understand since it is where curvature and noise interact, as indicated by Equation (4.1). In Step 3 we remove this annular region, so that we only need to deal with ν_2 restricted to $\mathring{\mathcal{B}}_{r_{\text{in}}}$. We decompose ν_2 as $\nu_2 = \nu_2^{\text{in}} + \nu_2^{\text{out}}$, where we define for each Borel set $U \subseteq T_{x_\perp} M$ the following:

$$\begin{aligned}
\nu_2^{\text{in}}(U) &:= \nu_2(U \cap \mathring{\mathcal{B}}_{r_{\text{in}}}) \\
\nu_2^{\text{out}}(U) &:= \nu_2(U \cap (\mathring{\mathcal{B}}_{r_{\text{out}}} - \mathring{\mathcal{B}}_{r_{\text{in}}}))
\end{aligned}$$

Define

$$\nu_3 := \left(\textstyle\int \nu_2^{\text{in}}\right)^{-1} \nu_2^{\text{in}}$$

where $\int \nu_2^{\text{in}} := \nu_2^{\text{in}}(T_{x_\perp} M)$ is the total mass of ν_2^{in}, which is a constant. The transportation plan is to: (a) transport ν_2^{out} to the Dirac delta distribution centered at $0 \in T_x M$ and (b) transport this Dirac delta distribution back to $(\int \nu_2^{\text{out}} / \int \nu_2^{\text{in}})\nu_2^{\text{in}}$. Note that $\int \nu_2^{\text{out}} / \int \nu_2^{\text{in}}$ is just a normalization constant and that $\int \nu_2^{\text{in}} + \int \nu_2^{\text{out}} = 1$. By Lemma 6.4, the transportation cost $\mathrm{W}_p(\nu_2, \nu_3)$ is bounded by $(r_{\text{out}} + r_{\text{in}}) \int \nu_2^{\text{out}}$, since the first part of this transportation moves by distance at most r_{out}, the second part moves by at most r_{in}, and the total mass to move is $\int \nu_2^{\text{out}}$.

[9] See for example [9] for a standard reference in geometric measure theory

Equation (4.6) carries over since ν_3 and ν_2^{in} are proportional; for each open $U \subseteq T_{x_\perp} M$,

$$\nu_3(U) = \frac{1}{\mu(\mathcal{B}_r(x)) \int \nu_2^{\text{in}}} \int_{U \cap \mathring{\mathcal{B}}_{r_{\text{in}}}} \varphi(\exp_{x_\perp} y) \, \mathrm{J} \exp_{x_\perp}(y) \, \mathrm{d}\, y \tag{4.7}$$

Step 4. We flatten out the non-uniformity in ν_3. As in Equation (4.7) above, ν_3 is given by the probability density function $\psi(y) := \varphi(\exp_{x_\perp} y) \, \mathrm{J} \exp_{x_\perp}(y)$ times a constant. Defining $\nu_4 = \mathrm{Unif}_d(\mathring{\mathcal{B}}_{r_{\text{in}}})$, we can directly apply Lemma 6.5:

$$\mathrm{W}_p(\nu_3, \nu_4) \leq \frac{\omega_d r_{\text{in}}^d}{\mu(\mathcal{B}_r(x)) \int \nu_2^{\text{in}}} \cdot (\psi_{\max} - \psi_{\min}) \cdot 2 r_{\text{in}}$$

where the factor $\omega_d r_{\text{in}}^d$ is needed to rescale the Lebesgue measure $\mathrm{d}\, y$ in Equation (4.7) into $\widetilde{\mathrm{d}\, y} = \mathrm{d}\, y / (\omega_d r_{\text{in}}^d)$ so that $\int_{\mathring{\mathcal{B}}_{r_{\text{in}}}} \widetilde{\mathrm{d}\, y} = 1$, so that Lemma 6.5 can be applied. In the above, extrema of ψ are taken over $\mathring{\mathcal{B}}_{r_{\text{in}}}$. The variation $\psi_{\max} - \psi_{\min}$ can be controlled with the triangle inequality as follows[10]:

$$|\psi_{\max} - \psi_{\min}| \leq (\varphi_{\max} - \varphi_{\min}) \left(1 + \frac{r_{\text{in}}^2}{2\tau^2}\right) + \varphi_{\min} \frac{2 r_{\text{in}}^2}{3\tau^2}$$

Here the extrema of φ are taken over the geodesic ball $\exp_{x_\perp}(\mathring{\mathcal{B}}_{r_{\text{in}}})$. We used Corollary 6.13, which tells us that:

$$1 - \frac{\|y\|^2}{6\tau^2} \leq |\, \mathrm{J} \exp_{x_\perp}(y)| \leq 1 + \frac{\|y\|^2}{2\tau^2} \tag{4.8}$$

We furthermore note that, by Equation 4.6,

$$\mu(\mathcal{B}_r(x)) \int \nu_2^{\text{in}} = \int_{\mathring{\mathcal{B}}_{r_{\text{in}}}} \varphi(\exp_{x_\perp} y) \, \mathrm{J} \exp_{x_\perp}(y) \, \mathrm{d}\, y \geq \omega_d r_{\text{in}}^d \left(1 - \frac{r_{\text{in}}^2}{6\tau^2}\right) \varphi_{\min}$$

so that

$$\varphi_{\min} \leq \frac{\mu(\mathcal{B}_r(x)) \int \nu_2^{\text{in}}}{\omega_d r_{\text{in}}^d} \cdot \frac{1}{1 - r_{\text{in}}^2 / 6\tau^2}$$

Thus the transportation cost is bounded as:

$$\mathrm{W}_p(\nu_3, \nu_4) \leq \left(\frac{\omega_d r_{\text{in}}^d}{\mu(\mathcal{B}_r(x)) \int \nu_2^{\text{in}}} (\varphi_{\max} - \varphi_{\min}) (1 + \frac{r_{\text{in}}^2}{2\tau^2}) + \frac{2 r_{\text{in}}^2 / 3\tau^2}{1 - r_{\text{in}}^2 / 6\tau^2} \right) \cdot 2 r_{\text{in}}$$

We note at this point that the extrema of φ may be taken over $\mathcal{B}_{r+2s}(x_\perp)$ instead, since $\mathcal{B}_{r+2s}(x_\perp) \supseteq \exp_{x_\perp}(\mathring{\mathcal{B}}_{r_{\text{in}}})$. This relaxation is done for a compatibility with another extrema of φ taken later.

Step 5. Here we simply rescale $\mathring{\mathcal{B}}_{r_{\text{in}}}$ to $\mathring{\mathcal{B}}_r$ radially, which multiplies the associated probability density function by a constant factor (Lemma 6.9), so that we get another uniform distribution. By Lemma 6.3, the transportation cost is bounded by $r - r_{\text{in}} = 2s$.

[10] Writing $\psi^{(1)} := \varphi \circ \exp_{x_\perp}$ and $\psi^{(2)} := \mathrm{J} \exp_{x_\perp}$ so that $\psi = \psi^{(1)} \psi^{(2)}$, we obtain that $|\psi_{\max} - \psi_{\min}| \leq |\psi^{(1)}_{\max} \psi^{(2)}_{\max} - \psi^{(1)}_{\min} \psi^{(2)}_{\min}| \leq |\psi^{(1)}_{\max} - \psi^{(1)}_{\min}| \cdot |\psi^{(2)}_{\max}| + |\psi^{(1)}_{\min}| \cdot |\psi^{(2)}_{\max} - \psi^{(2)}_{\min}|$

The Total Bound. Collecting the bounds[11], we get:

$$\begin{aligned}
&\mathrm{W}_p(\nu_0,\nu_5)\\
\leq&\,\mathrm{W}_p(\nu_0,\nu_1)+\mathrm{W}_p(\nu_1,\nu_2)+\mathrm{W}_p(\nu_2,\nu_3)+\mathrm{W}_p(\nu_3,\nu_4)+\mathrm{W}_p(\nu_4,\nu_5)\\
\leq&s+\frac{(r+2s)^2}{\tau}+(r_{\mathrm{in}}+r_{\mathrm{out}})\int\nu_2^{\mathrm{out}}\\
&+\left(\frac{\omega_d r_{\mathrm{in}}^d}{\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{in}}}(\varphi_{\max}-\varphi_{\min})\,(1+\frac{r_{\mathrm{in}}^2}{2\tau^2})+\frac{2r_{\mathrm{in}}^2/3\tau^2}{1-r_{\mathrm{in}}^2/6\tau^2}\right)\cdot 2r_{\mathrm{in}}+2s
\end{aligned}\tag{4.9}$$

Using Equations (4.5), (4.6) and (4.8), we obtain the following bounds:

$$\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{in}}=\mu_0(\exp_{x_\perp}\mathring{\mathcal{B}}_{r_{\mathrm{in}}})\leq\varphi_{\max}(1+\frac{r_{\mathrm{out}}^2}{2\tau^2})\omega_d r_{\mathrm{in}}^d$$

$$\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{out}}\leq\mu_0(\exp_{x_\perp}(\mathring{\mathcal{B}}_{r_{\mathrm{out}}}-\mathring{\mathcal{B}}_{r_{\mathrm{in}}}))\leq\varphi_{\max}(1+\frac{r_{\mathrm{out}}^2}{2\tau^2})\omega_d(r_{\mathrm{out}}^d-r_{\mathrm{in}}^d)$$

where $\varphi_{\max}$ is the maximum of φ taken over $\mathcal{B}_{r+2s}(x_\perp)$.[12] Combining these, we get:

$$\frac{\int\nu_2^{\mathrm{out}}}{\int\nu_2^{\mathrm{in}}}=\frac{\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{out}}}{\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{in}}}\leq\frac{\varphi_{\max}(1+r_{\mathrm{out}}^2/2\tau^2)\omega_d(r_{\mathrm{out}}^d-r_{\mathrm{in}}^d)}{\mu_0(\exp_{x_\perp}\mathring{\mathcal{B}}_{r_{\mathrm{in}}})}=\Phi'(\Omega^{-d}-1)$$

$$\text{with }\Omega=\frac{r_{\mathrm{in}}}{r_{\mathrm{out}}},\Phi'=\frac{\varphi_{\max}(1+r_{\mathrm{out}}^2/2\tau^2)\omega_d r_{\mathrm{in}}^d}{\mu_0(\exp_{x_\perp}\mathring{\mathcal{B}}_{r_{\mathrm{in}}})}\geq 1$$

Here the upper bound for $\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{in}}$ was used only to show $\Phi'\geq 1$. We can bound $\int\nu_2^{\mathrm{out}}$ using the above, as follows:

$$\int\nu_2^{\mathrm{out}}=\left(1+\frac{\int\nu_2^{\mathrm{in}}}{\int\nu_2^{\mathrm{out}}}\right)^{-1}\leq\left(1+\frac{1}{\Phi'(\Omega^{-d}-1)}\right)^{-1}\leq\Phi'(1-\Omega^d)$$

where the first inequality holds by plugging in the upper bound for $\int\nu_2^{\mathrm{out}}/\int\nu_2^{\mathrm{in}}$, and the second inequality holds since $\Phi'\geq 1$. Plugging these into Equation (4.9), we get that $\mathrm{W}_p(\nu_0,\nu_5)$ is no larger than

$$\begin{aligned}
&3s+\frac{(r+2s)^2}{\tau}+(r_{\mathrm{in}}+r_{\mathrm{out}})(1-\Omega^d)\varphi_{\max}(1+\frac{r_{\mathrm{out}}^2}{2\tau^2})\frac{\omega_d r_{\mathrm{in}}^d}{\mu_0(\exp_{x_\perp}\mathring{\mathcal{B}}_{r_{\mathrm{in}}})}\\
&+\left(\frac{\omega_d r_{\mathrm{in}}^d}{\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{in}}}(\varphi_{\max}-\varphi_{\min})\left(1+\frac{r_{\mathrm{in}}^2}{2\tau^2}\right)+\frac{2r_{\mathrm{in}}^2/3\tau^2}{1-r_{\mathrm{in}}^2/6\tau^2}\right)\cdot 2r_{\mathrm{in}}
\end{aligned}$$

[11] We use a slight abuse of notation and identify ν_k with $\iota_*\nu_k$ for $k=2,\dots 5$, where $\iota: T_{x_\perp}M\hookrightarrow\mathbb{R}^D$ is the inclusion of tangent space. This is not a problem, since generally $\mathrm{W}_p(\iota_*\mu_1,\iota_*\mu_2)\leq\mathrm{W}_p(\mu_1,\mu_2)$ holds for any measures μ_1,μ_2 on $T_{x_\perp}M$.

[12] It suffices to take maximum of φ over $\mathcal{B}_{r+2s}(x)$ in bounding $\mu(\mathcal{B}_r(x))\int\nu_2^{\mathrm{out}}$, since ν_2 is supported on $\exp_{x_\perp}^{-1}(\mathcal{B}_{r+2s}(x))$.

By the assumption $r + 2s \leq (\sqrt{2}-1)\tau$, we have both $r_{\text{in}} \leq (\sqrt{2}-1)\tau$ and $r_{\text{out}} \leq (2-\sqrt{2})\tau$. These inequalities further imply:

$$1 + \frac{r_{\text{in}}^2}{2\tau^2} \leq 1.09 \text{ and } 1 + \frac{r_{\text{out}}^2}{2\tau^2} \leq 1.18 \text{ and } \frac{2/3}{1 - r_{\text{in}}^2/6\tau^2} \leq 0.69.$$

Plugging these numbers into our bound above for $\mathrm{W}_p(\nu_0, \nu_5)$ yields the desired result. □

We have the following bound, upon further assumptions on the noise radius s and the probability density φ:

COROLLARY 4.5. *In Proposition 4.4, suppose that we additionally assume that there exist α, β satisfying:*

$$\|\varphi(x) - \varphi(y)\| \leq \alpha \cdot \mathrm{d}_M(x, y), \text{ for any } x, y \in M$$
$$\sigma \leq \beta\rho^2, \text{ with } \beta \leq 1.2$$

Then we have the following bound for any $p \geq 1$:

$$\mathrm{W}_p(\nu, \tilde{\nu}) \leq Q_1(\rho, \beta) \cdot \tau\rho^2$$

where $Q_1(\rho, \beta)$ is given by:

$$Q_1(\rho, \beta) = 3\beta + \beta_1^2 + \frac{1.2\varphi_{\max} d}{\Phi}(2 + \beta_1^2\rho)(1 + 4\beta) + \frac{4.4\alpha\tau}{\Phi}(1 + \beta_1\rho)\beta_1 + 1.4\rho$$

where $\beta_1 = 1 + 2\beta\rho$. In particular, for $\beta = 1/2$, we have:

$$Q_2(\rho) := Q_1(\rho, \frac{1}{2}) = (2.5 + 3.4\rho + \rho^2) + \frac{3.6\varphi_{\max} d}{\Phi}(2 + \rho + 2\rho^2 + \rho^3) + \frac{4.4\alpha\tau}{\Phi}(1 + 2\rho + 2\rho^2 + \rho^3)$$

PROOF. We first have:

$$\rho + 2\sigma \leq \beta_1\rho, \text{ and } 1 - \Omega^d \leq d(1 + 4\beta)\rho \tag{4.10}$$

where the first line is by the definition of β_1 and the second line is by Lemma 6.7. This almost derives Q_1, except the bound on $\varphi_{\max} - \varphi_{\min}$. Since geodesic distance is used, the Lipschitz assumption on φ implies: $\varphi_{\max} - \varphi_{\min} \leq 2\alpha r_{\text{out}}$ by using radial segments in the ball $\mathring{\mathcal{B}}_{r_{\text{out}}} \subseteq T_{x_\perp} M$. By the definition of r_{out} and the bound $\rho + 2\sigma \leq \beta_1\rho$, we have:

$$\frac{r_{\text{out}}}{\tau} \leq (\rho + 2\sigma) + (\rho + 2\sigma)^2 \leq (1 + \beta_1\rho)\beta_1\rho$$

and thus:

$$\varphi_{\max} - \varphi_{\min} \leq 2\alpha\tau(1 + \beta_1\rho)\beta_1\rho \tag{4.11}$$

Plugging in Equations 4.10 and 4.11 into the expression for $Q(\rho, \sigma)$ derives the expression for Q_1. Note that the condition $\beta \leq 1.2$ is simply added so that if $\rho \leq \sqrt{2} - 1$, we get $1 - 2\beta\rho \geq 0$ and thus $\rho - 2\beta\rho^2 = \rho(1 - 2\beta\rho) \geq 0$, which is necessary for applying Lemma 6.7. The expression for Q_2 is obtained by direct substitution of $\beta = 1/2$. □

5. Tangent space and dimension estimation

In this section, we combine the Propositions 2.6, 3.3, and 4.4 to prove Theorem 5.3. This in turn implies both Theorem A and B.[13]

DEFINITION 5.1. Given a d-dimensional subspace $\Pi \subseteq \mathbb{R}^D$, denote the $D \times D$ orthogonal projection matrix to Π by P_Π, which is a real symmetric matrix, given concretely as:

$$\mathrm{P}_\Pi = A_\Pi A_\Pi^\top$$

where $A_\Pi \in \mathbb{R}^{D\times d}$ is any matrix whose columns form an orthonormal basis of Π.

DEFINITION 5.2. Let $\mathbf{X} = (X_1, \dots X_m)$ be an i.i.d. sample drawn from μ, a Borel probability measure on $\mathbb{R}^D$. Given $x \in \mathbb{R}^D$ and $r > 0$, define:

$$\hat{\mathrm{P}}_i := \frac{d+2}{r^2}\Sigma[\delta_{\mathbf{X}_i}|_{U_i}], \text{ where } \mathbf{X}_i = \{X_j\}_{j\neq i}, U_i = \mathcal{B}_r(X_i)$$

If $\Pi \subseteq \mathbb{R}^D$ is a d-dimensional subspace, then Lemma 6.1 says that:

$$(d+2)\Sigma[\mathrm{Unif}(\Pi \cap \mathcal{B}_1(0))] = P_\Pi$$

Thus an approximation to this covariance matrix in Proposition 4.4 amounts to the approximation of a projection matrix, and justifies the definition of $\hat{\mathrm{P}}_i$.

THEOREM 5.3. *Let $(\mu, \mu_0) \in \mathcal{P}(M, s)$[14] where M is a smoothly embedded compact d-dimensional manifold $M \subseteq \mathbb{R}^D$ with reach τ and $s \geq 0$ is a real number. Let φ be the probability density function of μ_0 which satisfies $\|\varphi(x) - \varphi(y)\| \leq \alpha \cdot \mathrm{d}_M(x, y)$. Let $X_1, \dots X_m$ be an i.i.d. sample drawn from μ and let $X_1^\perp, \dots X_m^\perp$ be their orthogonal projections to M. Given $\delta, \epsilon, \alpha > 0$ and assuming*[15] *$\epsilon < 2$, suppose r, m satisfy the following:*

$$\sqrt{\frac{2s}{\tau}} < \frac{r}{\tau} < \frac{\epsilon}{16(d+2)Q_2(r/\tau)} \text{ and } \frac{m}{\log m} \geq \frac{4642(d+2)^2}{u_0\epsilon^2}\log\left(\frac{14D\alpha}{\delta}\right)$$

where $u_0 = \inf_{x\in \mathrm{supp}\,\mu} \mu(\mathcal{B}_r(x))$. Then with probability at least $1-\delta$, the following holds:

$$\max_{i\leq \alpha m} \left\|\hat{\mathrm{P}}_i - \mathrm{P}_i\right\| \leq \epsilon$$

where P_i is the projection matrix to the tangent space $T_{X_i^\perp}M$, and Q_2 is defined as:

$$Q_2(\rho) = (2.5 + 3.4\rho + \rho^2) + \frac{3.6\varphi_{\max} d}{\Phi}(2 + \rho + 2\rho^2 + \rho^3) + \frac{4.4\alpha\tau}{\Phi}(1 + 2\rho + 2\rho^2 + \rho^3)$$

where $\Phi = \dfrac{\mu_0(\exp_{x_\perp} \mathring{\mathcal{B}}_{r-2s})}{\omega_d(r-2s)^d}$

[13] Minor technical note: In the special cases discussed in the Introduction, we set $k = m$ in Theorems A and B, use Lemma 6.6, and use $\log(14D) > 1 + \log(4D + 2)$ assuming $D \geq 2$.

[14] See Definition 4.2.

[15] Nothing is lost from this assumption since operator norm of the difference of two projection operators is at most 2.

PROOF. Out of total allowed error ϵ, we will allocate one half $\epsilon/2$ to the concentration inequality (Proposition 2.6) and the other half $\epsilon/2$ to the curvature (Proposition 4.4). Throughout the proof, we use the shorthand $U_i = \mathcal{B}_r(X_i^\perp)$.

Concentration inequality: By Proposition 2.6, we may use k points for local covariance estimation by error level $r^2\epsilon/2(d+2)$:

$$\|\Sigma[\delta_{\mathbf{x}_i}|_{U_i}] - \Sigma[\mu|_{U_i}]\| \le \frac{r^2}{d+2} \cdot \frac{\epsilon}{2}, \text{ for all } i \le k$$

with probability at least $1-\delta$, if m satisfies the inequality in the theorem statement.

Curvature: By combining Corollary 4.5 and Proposition 3.3, the following holds for every $x \in \operatorname{supp}\mu$:

$$\left\|\Sigma[\mu|_{U_i}] - \frac{r^2}{d+2}\,\mathrm{P}_i\right\| \le 8r \cdot \frac{r^2 Q_2}{\tau} \le \frac{8\tau\epsilon}{16(d+2)Q_2} \cdot \frac{r^2 Q_2}{\tau} = \frac{r^2}{d+2} \cdot \frac{\epsilon}{2}$$

where $Q_2 = Q_2(r/\tau)$. In the second inequality, the assumption on r in the theorem statement was used. Note that $\frac{r^2}{d+2}\,\mathrm{P}_{X_i^\perp}$ is the covariance of the uniform measure over the tangential disk of radius r, by Lemma 6.1.

By the triangle inequality, for all $i \le k$ we have

$$\begin{aligned}\left\|\frac{d+2}{r^2}\Sigma[\delta_{\mathbf{x}_i}|_{U_i}] - \mathrm{P}_i\right\| &\le \frac{d+2}{r^2}\left(\|\Sigma[\delta_{\mathbf{x}}|_{U_i}] - \Sigma[\mu|_{U_i}]\| + \left\|\Sigma[\mu|_{U_i}] - \frac{r^2}{d+2}P_i\right\|\right)\\ &\le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,\end{aligned}$$

as desired. We note that the assumptions $2s \le r$ and $r + 2s \le (\sqrt{2}-1)\tau$ of Proposition 4.4 follow from the assumption on r and $\epsilon < 2$. □

5.1. Proof of Theorem A. To use Theorem 5.3, we relate the projection matrices to angular deviation between subspaces using the Davis-Kahan theorem (see [35] or [8]).

DEFINITION 5.4. Suppose $\Pi_1, \Pi_2 \subseteq \mathbb{R}^D$ two subspaces of $\mathbb{R}^D$. The *principal angle* between Π_1 and Π_2 is defined as:

$$\measuredangle(\Pi_1, \Pi_2) := \max_{x\in\Pi_1}\min_{y\in\Pi_2} \measuredangle(x,y)$$

with $\cos\measuredangle(x,y) = \langle x, y\rangle/(\|x\|\cdot\|y\|)$.

DEFINITION 5.5. For a real symmetric matrix A of size $D\times D$, suppose its diagonalization is given by $A = U\Lambda U^\top$, with U being an orthogonal matrix and Λ being a diagonal matrix whose entries are arranged in the decreasing order. Then for an integer $k \le D$, define $\Pi(A,k) \subseteq \mathbb{R}^D$ as the span of the first k columns of U.

THEOREM 5.6 (Davis-Kahan). *Suppose that A is a real symmetric matrix with eigenvalues $\lambda_1^A \ge \lambda_2^A \ge \cdots$. Then for any other real symmetric matrix B and a positive integer k such that $\lambda_k^A \ne \lambda_{k+1}^A$,*

$$\sin\measuredangle\Big(\Pi(A,k), \Pi(B,k)\Big) \le \frac{\|A-B\|}{\lambda_k^A - \lambda_{k+1}^A}$$

Proof of Theorem A. Taking $\epsilon = \sin\theta$ in Theorem 5.3, the following holds for each $i \le k$:

$$\| \mathrm{P}_i - \hat{\mathrm{P}}_i \| \le \sin\theta$$

Since both P_i and $\hat{\mathrm{P}}_i$ are real symmetric matrices and since eigenvalues of P_i are $(1, \dots 1, 0, \dots 0)$, letting $A = \mathrm{P}_i$, $B = \hat{\mathrm{P}}_i$, and $k = d$ in the Davis-Kahan theorem gives:

$$\sin \measuredangle \Big(\Pi(\mathrm{P}_i, d), \Pi(\hat{\mathrm{P}}_i, d) \Big) \le \| \mathrm{P}_i - \hat{\mathrm{P}}_i \| \le \sin\theta$$

Since P_i is the projection matrix to $T_{X_i^\perp} M$, a d-dimensional space, we have $\Pi(\mathrm{P}_i, d) = T_{X_i^\perp} M$. Furthermore, $\Pi(\hat{\mathrm{P}}_i, d) = \Pi(\Sigma[\delta_{\mathbf{X}_i}|_{U_i}], d) = \hat{\Pi}_i$, where $U_i = \mathcal{B}_r(X_i)$.

Finally, the condition $r/\tau < \epsilon/16(d+2)q$ in Theorem A implies $r/\tau \le 1/48$ and thus $Q_2(r/\tau) \le q = 3 + (8\varphi_{\max} d + 5\alpha\tau)/\varphi_{\min}$, so that Theorem 5.3 applies. Also, the condition $r + 2s \le (\sqrt{2}-1)\tau$ is dropped from Theorem A because $r/\tau < 1/48$ and $\sqrt{2s/\tau} < r/\tau$ implies $r + 2s \le (\sqrt{2}-1)\tau$.

5.2. Proof of Theorem B. To relate a perturbation of eigenvalues to a perturbation of covariance matrices, we use the Hoffman-Wielandt theorem [15].

THEOREM 5.7 (Hoffman-Wielandt). *For normal matrices A, A' of dimension $D \times D$, there is an enumeration of eigenvalues $(\lambda_1, \dots \lambda_D)$ of A and $(\lambda'_1, \dots \lambda'_D)$ of A' such that*

$$\sum_{i=1}^{D} |\lambda_i - \lambda'_i|^2 \le \|A - A'\|_{\mathrm{F}}^2$$

where $\|A\|_{\mathrm{F}} := \sqrt{Tr(A^\top A)}$ denotes the Frobenius norm, with $Tr(\bullet)$ denoting the trace. In particular, if A, A' are real symmetric matrices, then[16]

$$\|\vec{\lambda}[A] - \vec{\lambda}[A']\| \le \|A - A'\|_{\mathrm{F}}$$

where $\vec{\lambda}[A] \in \mathbb{R}^D$ is the vector of eigenvalues of A, arranged in the decreasing order.

Now we note the following simple result for dimension estimation using tail sum.

PROPOSITION 5.8. *Let $\vec{\lambda} = (\lambda_1, \dots \lambda_D) \in \mathbb{R}^D$ be such that $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_D \ge 0$. Let $\vec{\lambda}(d, D) = \frac{1}{d+2}(1, \dots 1, 0 \dots 0) \in \mathbb{R}^D$ where there are $D - d$ zeros. Let η be a tolerance parameter such that $0 < \eta < 1/(2d)$.*

$$\left\| \vec{\lambda} - \vec{\lambda}(d, D) \right\|_2 < \frac{1}{3\sqrt{D}(1 + \eta^{-1})} \implies \mathrm{Thr}(\vec{\lambda}, \eta) = d$$

where Thr *is defined in the Introduction.*

PROOF. Writing $\vec{\lambda} - \vec{\lambda}(d, D) = (t_1, \dots t_D)$, let $q_1 = |t_1| + \cdots + |t_d|$, $q_2 = |t_{d+1}| + \cdots + |t_D|$, and $q = q_1 + q_2 = \|\vec{\lambda} - \vec{\lambda}(d, D)\|_1$. Then since generally $D^{-1/2}\|x\|_1 \le \|x\|_2$, we have:

$$q < \sqrt{D} \cdot \frac{\eta}{3\sqrt{D}(1+\eta)} = \frac{\eta}{3(1+\eta)}$$

[16] This special case for real symmetric matrices follows from Lemma 6.14.

A sufficient condition for $\mathrm{Thr}(\vec{\lambda}, \eta) = d$ is:

$$q_2 \leq \eta \|\vec{\lambda}\|_1, \text{ and } q_2 + \left(\frac{1}{d+2} - q_1\right) > \eta \|\vec{\lambda}\|_1$$

Since $\|\vec{\lambda}(d, D)\|_1 = d/(d+2)$, triangle inequality implies that $\frac{d}{d+2} - q \leq \|\vec{\lambda}\|_1 \leq \frac{d}{d+2} + q$. Thus we can formulate the following sufficient conditions:

$$\begin{aligned} & q < \eta \left(\frac{d}{d+2} - q\right), \text{ and } \frac{1}{d+2} - q > \eta \left(\frac{d}{d+2} + q\right) \\ \iff & (1+\eta)q < \frac{\eta d}{d+2}, \text{ and } (1+\eta)q < \frac{1-\eta d}{d+2} \\ \iff & q < \frac{\min(\eta d, 1-\eta d)}{(1+\eta)(d+2)} \end{aligned}$$

By our assumption that $\eta < 1/(2d)$, we have $\min(\eta d, 1-\eta d) = \eta d$. Thus our sufficient condition is $q < \frac{\eta}{1+\eta} \cdot \frac{d}{d+2}$. The right hand side is minimised for $d = 1$, so that this is precisely implied by the assumption. □

Proof of Theorem B.

The proof goes verbatim except we use the Hoffman-Wielandt theorem instead of the Davis-Kahan theorem, and that we use the estimation error for the covariances $\|\hat{\Sigma} - \Sigma\|_2$: $\epsilon^{-1} = 3D(1+\eta^{-1})$. Then the following chain of inequalities hold with probability $\geq 1 - \delta$:

$$\|\vec{\lambda} - \vec{\lambda}(d, D)\|_2 \leq \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}} \leq \sqrt{D} \cdot \|\hat{\Sigma} - \Sigma\|_2 \leq \frac{1}{3\sqrt{D}(1+\eta^{-1})}$$

The proof is then completed by applying Proposition 5.8.

References

[1] E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman. Estimating the reach of a manifold. *Electronic journal of statistics*, 13(1):1359–1399, 2019.

[2] E. Aamari and C. Levrard. Stability and minimax optimality of tangential delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry*, 59(4):923–971, 2018.

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

[4] Y. Aizenbud and B. Sober. Non-parametric estimation of manifolds from noisy data. *arXiv:2105.04754 [math.ST]*, 2021.

[5] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local PCA. *J. Mach. Learn. Res.*, 18:Paper No. 9, 57, 2017.

[6] A. Block, Z. Jia, Y. Polyanskiy, and A. Rakhlin. Intrinsic dimension estimation. *arXiv preprint arXiv:2106.04018*, 2021.

[7] F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4, 2021.

[8] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[9] H. Federer. *Geometric measure theory*. Springer, 2014.

[10] C. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan. Fitting a putative manifold to noisy data. In *Conference On Learning Theory*, pages 688–720. PMLR, 2018.

[11] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[12] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 100(2):176–183, 1971.

[13] C. R. Genovese, M. P. Pacifico, I. Verdinelli, L. Wasserman, et al. Minimax manifold estimation. *Journal of machine learning research*, 13:1263–1291, 2012.

[14] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012.

[15] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 118–120. World Scientific, 2003.

[16] N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural computation*, 9(7):1493–1516, 1997.

[17] D. N. Kaslovsky and F. G. Meyer. Non-asymptotic analysis of tangent space perturbation. *Inf. Inference*, 3(2):134–187, 2014.

[18] J. Kim, A. Rinaldo, and L. Wasserman. Minimax rates for estimating the dimension of a manifold. *arXiv preprint arXiv:1605.01011*, 2016.

[19] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

[20] V. Koltchinskii and K. Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157, 2017.

[21] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*, volume 1. Springer.

[22] J. M. Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

[23] E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

[24] M. Lezcano-Casado. Geometric optimisation on manifolds with applications to deep learning. *DPhil Thesis, University of Oxford*, 2021.

[25] T. Minka. Automatic choice of dimensionality for pca. *Advances in neural information processing systems*, 13:598–604, 2000.

[26] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[27] M. Reiß and M. Wahl. Nonasymptotic upper bounds for the reconstruction error of pca. *The Annals of Statistics*, 48(2):1098–1123, 2020.

[28] L. Simon. *Lectures on geometric measure theory*. The Australian National University, Mathematical Sciences Institute, Centre . . . , 1983.

[29] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.*, 65(8):1067–1144, 2012.

[30] R. Tinarrage. Recovering the homology of immersed manifolds. *arXiv preprint arXiv:1912.03033*, 2019.

[31] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[32] H. Tyagi, E. Vural, and P. Frossard. Tangent space estimation for smooth embeddings of Riemannian manifolds. *Inf. Inference*, 2(1):69–114, 2013.

[33] S. Valle, W. Li, and S. J. Qin. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*, 38(11):4389–4401, 1999.

[34] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[35] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[36] L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.

6. Appendix

6.1. Notations and conventions. Here are some conventions we use.

- The word 'dimension' and 'intrinsic dimension' are used interchangeably, where 'intrinsic' simply distinguishes it from the 'ambient' dimension.
- All manifolds are connected.
- All vectors are by default column vectors.
- $\|v\| = \sqrt{v^\top v}$ denotes the Euclidean norm of a vector $v \in \mathbb{R}^D$.
- $\|A\|$ denotes the operator norm of a matrix $A \in \mathbb{R}^{m\times n}$, seen as a map $\mathbb{R}^n \to \mathbb{R}^m$. $\|A\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(A^\top A)}$ denotes its Frobenius norm.
- I_d denotes the $d \times d$ identity matrix.
- $\mathbb{E}[X]$ denotes the expected value of a random variable X.
- $\Sigma[\mu]$ denotes the covariance matrix of a Borel probability measure μ over $\mathbb{R}^D$.
- $\mathcal{B}_r(x) \subseteq \mathbb{R}^D$ denotes the open ball of radius r centered at $x \in \mathbb{R}^D$.
- Given a smoothly embedded manifold $M \subseteq \mathbb{R}^D$ and a point $x \in M$, $\mathring{\mathcal{B}}_r \subseteq T_xM$ denotes the open ball of radius r centered at $0 \in T_xM$, assuming that the choice of x is clear from the context.
- $\vec{\lambda}[A] \in \mathbb{R}^D$ denotes the eigenvalues of a real symmetric matrix A of size $D \times D$, arranged in the decreasing order.
- $\omega_d = \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ denotes the volume of the d-dimensional unit ball.

Additionally, the following letters have specific meanings if not stated otherwise:

Notation	Meaning
M	A compact manifold smoothly embedded in $\mathbb{R}^D$
d	(Intrinsic) dimension of M
D	Ambient dimension
τ	Reach of M
μ	Main distribution of interest with noise
μ_0	μ before adding noise
φ	Probability density function on M used to define μ_0
α	Lipschitz constant for φ
m	Sample size
r	Local detection radius
s	Noise radius
ϱ	Probabilistic guarantees hold for $\lfloor \varrho m \rfloor$ out of m points
δ	Probabilistic guarantees hold with probability $\geq 1-\delta$
ρ	Normalized local detection radius $\rho = r/\tau$
σ	Normalized noise radius $\sigma = s/\tau$

6.2. Technical lemmas.

LEMMA 6.1. *(Lemma 13 from [5]) Given a d-dimensional subspace Π of $\mathbb{R}^D$, the covariance matrix of the uniform distribution over the disk $\Pi \cap \mathcal{B}_1(0)$ is given by:*

$$\Sigma[\mathrm{Unif}_d(\Pi \cap \mathcal{B}_1(0))] = \frac{1}{d+2}\,\mathrm{P}_\Pi$$

where P_Π is the $D \times D$ projection matrix to Π. Eigenvalues of this matrix are:

$$\frac{1}{d+2}(\underbrace{1,\dots 1}_{d},\underbrace{0,\dots 0}_{D-d})$$

PROOF. Denote by $\Pi_{d,D}$ the d-dimensional subspace of $\mathbb{R}^D$ spanned by the first d canonical basis vectors. The only nontrivial covariance between the marginals of $\mathrm{Unif}_d(\Pi_{d,D} \cap \mathcal{B}_1(0))$ is:

$$\frac{1}{\omega_d}\int_{\|x\|\le 1} x_1^2\,\mathrm{d}\,x = \frac{1}{\omega_d \cdot d}\int_{\|x\|\le 1}\|x\|^2\,\mathrm{d}\,x = \frac{1}{d}\int_0^1 r^2 \cdot dr^{d-1}\,\mathrm{d}\,r = \int_0^1 r^{d+1}\,\mathrm{d}\,r = \frac{1}{d+2}$$

where $1/\omega_d$ is multiplied so that the distribution is uniform over the unit disk. This yields the calculation for the vector of eigenvalues. Thus,

$$\Sigma[\mathrm{Unif}_d(\Pi_{d,D} \cap \mathcal{B}_1(0)] = \frac{1}{d+2}\begin{bmatrix} I_d & 0 \\ 0 & \mathbf{0}_{D-d} \end{bmatrix}$$

Given any d-dimensional subspace $\Pi \subseteq \mathbb{R}^D$, consider an orthonormal basis $A = [v_1, \dots v_D]$ such that the first d vectors $[v_1, \dots v_d]$ span Π. If $X \sim \mathrm{Unif}(\Pi \cap \mathcal{B}_1(0))$, then $A^{-1}X \sim \mathrm{Unif}(\Pi_{d,D} \cap \mathcal{B}_1(0))$. Thus the covariance matrix of X is

$$\frac{1}{d+2}A\begin{bmatrix} I_d & 0 \\ 0 & \mathbf{0}_{D-d} \end{bmatrix}A^\top = \frac{1}{d+2}[v_1, \dots v_d][v_1, \dots v_d]^\top = \frac{1}{d+2}\,\mathrm{P}_\Pi$$

□

LEMMA 6.2. *Suppose*

$$\vec{\lambda}(d,D) := \frac{1}{d+2}(\underbrace{1,\dots 1}_{d},\underbrace{0,\dots 0}_{D-d})$$

If $d \le d'$, then

$$\|\vec{\lambda}(d,D) - \vec{\lambda}(d',D)\|^2 = \frac{(d'-d)(dd'+4d+4)}{(d+2)^2(d'+2)^2}$$

Also for any $k \ne d$, we have:

$$\|\vec{\lambda}(k,D) - \vec{\lambda}(d,D)\| \ge \|\vec{\lambda}(d,D) - \vec{\lambda}(d+1,D)\| = \frac{\sqrt{(d+1)(d+4)}}{(d+2)(d+3)}$$

PROOF. The norm of difference is given by direct computation:

$$\|\vec{\lambda}(d,D)-\vec{\lambda}(d',D)\|^2 = d\cdot\left(\frac{1}{d+2}-\frac{1}{d'+2}\right)^2+\frac{d'-d}{(d'+2)^2}=\frac{(d'-d)(dd'+4d+4)}{(d+2)^2(d'+2)^2}$$

The partial derivative of the above expression with respect to d and d' are strictly negative and positive respectively, whenever $0<d<d'$. Thus for each $d\geq 2$,

$$\begin{aligned}
&\min_{d'\neq d}\|\vec{\lambda}(d,D)-\vec{\lambda}(d',D)\| \\
=&\min(\|\vec{\lambda}(d,D)-\vec{\lambda}(d+1,D)\|,\|\vec{\lambda}(d,D)-\vec{\lambda}(d-1,D)\|) \\
=&\min\left(\frac{\sqrt{(d+1)(d+4)}}{(d+2)(d+3)},\frac{\sqrt{d(d+3)}}{(d+1)(d+2)}\right) \\
=&\frac{\sqrt{(d+1)(d+4)}}{(d+2)(d+3)}
\end{aligned}$$

where we use the fact that $\frac{\sqrt{(d+1)(d+4)}}{(d+2)(d+3)}$ is decreasing in d for $d\geq 0$ (directly checked by computing the derivative of its square). □

Let's prove simple inequalities associated to optimal transport, constituting the main tools to obtain the necessary bounds for covariance matrices.

LEMMA 6.3. *Let M be a Polish metric space with metric d_M. Suppose $A,B\subseteq M$ are Borel measurable, with inclusion maps $\iota^A: A\hookrightarrow M, \iota^B: B\hookrightarrow M$. Suppose that there is a continuous bijection $f: A\to B$ with a $L\geq 0$ with $d_M(x,f(x))<L$ for any x. Let μ be a Borel probability measure on A. Then for any $p\geq 1$, the Wasserstein distance between pushforwards of μ and $f_*\mu$ along inclusions are bounded by L:*

$$\mathrm{W}_p(\iota^A_*\mu,\iota^B_* f_*\mu)\leq L$$

PROOF. If $X\sim\iota^A_*\mu$, then $f(X)\sim\iota^B_* f_*\mu$. Therefore, by using the coupling $(X,f(X))$, we obtain the claim:

$$W_p(\iota^A_*\mu,\iota^B_* f_*\mu)\leq(\mathbb{E}_X\,\mathrm{d}_M(X,f(X))^p)^{1/p}\leq L$$

□

LEMMA 6.4. *Let M be a Polish metric space with metric d_M and a finite diameter $L:=\sup_{x,y\in M}\mathrm{d}_M(x,y)$. For a Borel probability measure μ on M and a Dirac delta measure δ_x centered at $x\in M$, we have:*

$$\mathrm{W}_p(\mu,\delta_x)\leq L$$

PROOF. Define the transportation plan ν on $M\times M$ by

$$\nu(U\times V)=\begin{cases}\mu(U) & \text{if } x\in V\\ 0 & \text{otherwise}\end{cases}$$

whose marginals are μ and δ_x. The transportation cost is bounded by L. □

LEMMA 6.5. *Let M be a Polish metric space with metric d_M and a finite diameter $L := \sup_{x,y\in M} \mathrm{d}_M(x,y)$. Fix a Borel probability measure μ on M. Let f be a non-negative continuous function on M with $\sup_{x\in M} f(x) - \inf_{x\in M} f(x) \leq C$ and $\int_M f(x)d\mu(x) = 1$. Let μ_f be the Borel probability measure on M given by taking f as the probability density function. Then for any $p \geq 1$,*

$$\mathrm{W}_p(\mu_f, \mu) \leq CL$$

PROOF. For any real number a, we have $a = \max(0,a) - \max(0,-a)$. Applying this to $a = f(x) - 1$, we may write:

$$\begin{aligned} \mu_f &= \mu + \mu_f^+ - \mu_f^- \\ \text{where } \mu_f^+(U) &= \int_U \max(0, f(x) - 1)\,\mathrm{d}\,\mu(x) \\ \mu_f^-(U) &= \int_U \max(0, 1 - f(x))\,\mathrm{d}\,\mu(x) \end{aligned}$$

As such, for any point $x \in M$,

$$\mathrm{W}_p(\mu_f, \mu) = \mathrm{W}_p(\mu + \mu_f^+ - \mu_f^-, \mu) \leq \mathrm{W}_p(\mu_f^+, \mu_f^-)$$

The inequality holds since generally $\mathrm{W}_p(\mu + \nu_1, \mu + \nu_2) \leq \mathrm{W}_p(\nu_1, \nu_2)$. Since $\mu(M) = \mu_f(M)$, we have $A := \mu_f^+(M) = \mu_f^-(M)$. Then

$$\mathrm{W}_p(\mu_f^+, \mu_f^-) \leq \mathrm{W}_p(\mu_f^+, A \cdot \delta_x) + \mathrm{W}_p(A \cdot \delta_x, \mu_f^-) \leq 2AL$$

The second inequality is by the previous lemma. By definition of μ_f^+, μ_f^-,

$$\begin{aligned} A = \mu_f^+(M) &\leq \sup_{x\in M} f(x) - 1 \\ A = \mu_f^-(M) &\leq 1 - \inf_{x\in M} f(x) \end{aligned}$$

Thus $2A \leq C$, and $2AL \leq CL$. □

LEMMA 6.6. *Suppose a, b, x are real where $b > 1$ and $x > e$. Then we have that*

$$\frac{x}{\log x} > a(1 + \log b) \implies x > a \log bx \implies \frac{x}{\log x} > a$$

PROOF. Writing $y = \log x > 1$ and $c = \log b > 0$, the assertion follows trivially:

$$x/y > a(1 + c) \implies x > a(y + c) \implies x/y > a$$

□

LEMMA 6.7. *For the following function*

$$f(x) = \frac{1 - ax}{(1 + ax)(1 + x + ax^2)}$$

the following holds whenever $a > 0, k \geq 1$ and $x \in [0, 1/a]$:

$$f(x)^k \geq 1 - k(1 + 2a)x$$

PROOF. Let's always assume $x \in [0, 1/a]$ here. By direct evaluation, $f'(0) = -(1+2a)$ and thus the claim is equivalent to $f(x)^k \geq 1 + kf'(0)x$. Since $f(0) = 1$, it's sufficient to show that $(f^k)'(x) \geq kf'(0)$ for any x. We have $f' < 0$ since f is decreasing, and we can also directly check that $0 \leq f \leq 1$. Thus $(f^k)' = kf^{k-1}f' \geq kf'$. Thus it suffices to show that $f' \geq f'(0)$. By direct computation, we have:

$$f'(x) = \frac{2a^3x^3 - (a^2x^2 + 4ax + 2a + 1)}{(1+ax)^2(1+x+ax^2)^2}$$

We want $f' \geq f(0) = -(1+2a)$, which is equivalent to:

$$2a^3x^3 - (a^2x^2 + 4ax + 2a + 1) + (1+2a)(1+ax)^2(1+x+ax^2)^2 \geq 0$$

which holds since all of the coefficients are positive, upon expanding the brackets. □

LEMMA 6.8. *For every $t > 0$ and $s > 1$, the following hold:*

$$\frac{1}{1-e^{-1/t}} - t \in [\frac{1}{2}, 1]$$
$$\frac{1}{\log(1-s^{-1})} + s \in [\frac{1}{2}, 1]$$

Furthermore, both functions are increasing.

PROOF. The function $s(t) = 1/(1-e^{-1/t})$ is an increasing bijection from $(0, \infty)$ to $(1, \infty)$ and we have $t = -1/\log(1 - s(t)^{-1})$. Thus it suffices to prove the properties regarding the function:

$$f(t) = \frac{1}{1-e^{-1/t}} - t = \frac{e^u}{e^u - 1} - \frac{1}{u} = \frac{ue^u - e^u + 1}{u(e^u - 1)}, \text{ where } u = \frac{1}{t}$$

Then the claim that this quantity falls in the interval $[1/2, 1]$ is equivalent to:

$$ue^u - u \leq 2ue^u - 2e^u + 2, \text{ and } ue^u - e^u + 1 \leq ue^u - u$$

or equivalently,

$$0 \leq (u-2)e^u + (u+2), \text{ and } 1 + u \leq e^u$$

The second inequality is a standard fact, and plugging it into the first inequality shows it easily. To show that $f(t)$ is increasing, we evaluate the derivative:

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\left(\frac{1}{1-e^{-1/t}} - t\right) = \frac{e^{1/t}}{(e^{1/t}-1)^2t^2} - 1$$

The derivative is positive iff:

$$\frac{1}{t^2} \leq \frac{(e^{1/t}-1)^2}{e^{1/t}}$$

which follows from the following:

$$u \leq u\sum_{k=0}^{\infty} \frac{(u/2)^{2k}}{(2k+1)!} = e^{u/2} - e^{-u/2}, \text{ where } u = \frac{1}{t}$$

□

LEMMA 6.9. *Let $f_0 : \mathbb{R}^d \to \mathbb{R}^+$ be a function such that $f_0(x) = f_0(\lambda x)$ for any $\lambda > 0$, and that f_0 is differentiable when restricted to the unit sphere S^{d-1}. Define the scaling map $f(x) = f_0(x)x$ for $x \neq 0$. Then the Jacobian determinant of f is given by:*

$$\mathrm{J}\, f(x) = f_0(x)$$

PROOF. We have that $\frac{\partial}{\partial x_j}(f_0(x)x_i) = \delta_{ij}\varphi + \frac{\partial f_0}{\partial x_j}x_i$ where δ_{ij} is the Kronecker delta. Then

$$\mathrm{J}\, f = \det(f_0 I_d + (\nabla g)x^\top) = f_0 + (\nabla f_0)^\top x = f_0$$

by the matrix determinant lemma and the fact that the directional derivative of $f_0(x)$ along x is zero. □

The following lemma, which is a simple extension of Proposition 6.3 of [26], controls the deviation of geodesic from the first order approximation:

LEMMA 6.10. *Let M be a smooth compact n-manifold embedded in $\mathbb{R}^D$ with reach τ. Suppose that x, y are connected by a (unit speed) geodesic $\gamma : [0, \tilde{r}] \to M$ of length $\tilde{r}$ with $\gamma(0) = x, \gamma(\tilde{r}) = y$, and denote $r = \|x - y\|$. Then the following inequalities hold:*

$$\tilde{r} - \frac{\tilde{r}^2}{2\tau} \leq r \leq \tilde{r}$$

If $r \leq 0.5\tau$, then the following hold:

$$\frac{\tilde{r}}{\tau} \leq 1 - \sqrt{1 - \frac{2r}{\tau}},\ \textit{and}\ \|y - (x + \tilde{r}\dot{\gamma}(0))\| \leq \frac{\tilde{r}^2}{2\tau}$$

If $r \leq (\sqrt{2} - 1)\tau \approx 0.4\tau$, then the following also hold:

$$\tilde{r} \leq r + \frac{r^2}{\tau},\ \textit{and}\ \|y - (x + \tilde{r}\dot{\gamma}(0))\| \leq \frac{r^2}{\tau}$$

PROOF. Since straight lines are geodesics in $\mathbb{R}^D$, we have $r \leq \tilde{r}$. Meanwhile by the triangle inequality,

$$r = \|\gamma(\tilde{r}) - \gamma(0)\| \geq \|\tilde{r}\dot{\gamma}(0)\| - \left\| \int_0^{\tilde{r}} \int_0^{t_1} \ddot{\gamma}(t_2)\, \mathrm{d}\, t_2\, \mathrm{d}\, t_1 \right\| \geq \tilde{r} - \frac{\tilde{r}^2}{2\tau}$$

When $r \leq \tau/2$, this is equivalent to $\tilde{r} \notin (\tau - \tau\sqrt{1 - 2\tau^{-1}r}, \tau + \tau\sqrt{1 - 2\tau^{-1}r})$. Since $\tilde{r} = 0$ when $r = 0$, by continuity we must have $\tilde{r} \leq \tau - \tau\sqrt{1 - 2\tau^{-1}r}$.

To get the error bound of first-order approximation, we calculate by basic calculus the following:

$$\gamma(\tilde{r}) - \gamma(0) = \int_0^{\tilde{r}} \dot{\gamma}(t_1)\, \mathrm{d}\, t_1 = \int_0^{\tilde{r}} \left(\dot{\gamma}(0) + \int_0^{t_1} \ddot{\gamma}(t_2)\, \mathrm{d}\, t_2 \right) \mathrm{d}\, t_1 = \tilde{r}\dot{\gamma}(0) + \int_0^{\tilde{r}} \int_0^{t_1} \ddot{\gamma}(t_2)\, \mathrm{d}\, t_2\, \mathrm{d}\, t_1$$

and thus

$$\|\gamma(\tilde{r}) - (\gamma(0) + \tilde{r}\dot{\gamma}(0))\| = \left\| \int_0^{\tilde{r}} \int_0^{t_1} \ddot{\gamma}(t_2)\, \mathrm{d}\, t_2\, \mathrm{d}\, t_1 \right\| \leq \int_0^{\tilde{r}} \int_0^{t_1} \frac{1}{\tau}\, \mathrm{d}\, t_2\, \mathrm{d}\, t_1 = \frac{\tilde{r}^2}{2\tau}$$

where the last inequality holds because for any t, $\|\ddot{\gamma}(t)\| \leq \tau^{-1}$ (the norm of the second fundamental form is bounded above by τ^{-1}. See Proposition 6.1 of [26]).

To get simpler bounds, now suppose that $r \leq (\sqrt{2}-1)\tau$. We note that $x \in [0, \sqrt{2}-1]$ implies[17] $1 - \sqrt{1-2x} \leq x + x^2$. Thus

$$\tilde{r} \leq \tau - \tau\sqrt{1-2\tau^{-1}r} \leq r + \frac{r^2}{\tau}$$
$$\|\gamma(\tilde{r}) - (\gamma(0) + \tilde{r}\dot{\gamma}(0))\| \leq \frac{\tilde{r}^2}{2\tau} \leq \frac{r^2}{2\tau^3}(r+\tau)^2 \leq \frac{r^2}{\tau}$$

□

LEMMA 6.11. *Let $M \subseteq \mathbb{R}^D$ be a compact smoothly embedded d-dimensional manifold with reach τ. Let $x \in M$ and let $0 \leq r \leq (\sqrt{2}-1)\tau$ be a radius parameter. Then*

$$\exp_x(\mathring{\mathcal{B}}_r) \subseteq \mathcal{B}_r(x) \cap M \subseteq \exp_x(\mathring{\mathcal{B}}_{r+r^2/\tau})$$

PROOF. The first inclusion $\exp_x(\mathring{\mathcal{B}}_r) \subseteq \mathcal{B}_r(x) \cap M$ holds because a straight line is a geodesic in the ambient space $\mathbb{R}^D$. To see the second inclusion, suppose that $\|x-y\| = s \leq (\sqrt{2}-1)\tau$. Then Lemma 6.10 tells us that any geodesic connecting (x, y) has length at most $s + s^2/\tau$. Applyig this to every $s \leq r$, we get the inclusion. □

Sectional curvature may be used to bound the Jacobian of the exponential map, as follows[24]:

THEOREM 6.12. *Let M be a Riemannian manifold with sectional curvature bounded below and above by κ_- and κ_+. Then for $x \in M$ and $v \in T_xM$, the following holds:*

$$\min\left(1, \frac{\sin\sqrt{\kappa_+}\|v\|}{\sqrt{\kappa_+}\|v\|}\right) \leq \|(\mathrm{d}\exp_x)_v\| \leq \max\left(1, \frac{\sin\sqrt{\kappa_-}\|v\|}{\sqrt{\kappa_-}\|v\|}\right)$$

for all $\|v\|$ if $\kappa_+ \leq 0$, and for $\|v\| \leq \pi/\sqrt{\kappa_+}$ otherwise. The quantity $\frac{\sin x}{x}$ is taken to be 1 when $x = 0$.

This implies a weaker bound given in terms of the reach:

COROLLARY 6.13. *Let $M \subseteq \mathbb{R}^D$ be a smoothly embedded compact Riemannian manifold with reach τ. Then for $x \in M$ and $v \in T_xM$ satisfying $r := \|v\| \leq \pi\tau$, we have:*

$$\frac{\sinh\sqrt{2}\tau^{-1}r}{\sqrt{2}\tau^{-1}r} \leq \|(\mathrm{d}\exp_x)_v\| \leq \frac{\sin\tau^{-1}r}{\tau^{-1}r}$$

In particular, if $r \leq 2\tau$, then

$$1 - \frac{r^2}{6\tau^2} \leq \|(\mathrm{d}\exp_x)_v\| \leq 1 + \frac{r^2}{2\tau^2}$$

[17] Since $(x+x^2)/(1-\sqrt{1-2x}) \in [1, 1.07]$ when $x \in [0, \sqrt{2}-1]$, this relaxation overestimates by at most 7 percent.

PROOF. Norm of the second fundamental form is bounded above by τ^{-1} [26], and thus by the Gauss equation applied to sectional curvature (i.e. $K(u,v) = \langle R(u,v)u, v\rangle = \langle \mathbb{I}(u,u), \mathbb{I}(v,v)\rangle - \|\mathbb{I}(u,v)\|^2$ for orthonormal u, v), we may take $\kappa_- = -2\tau^{-2}$ and $\kappa_+ = \tau^{-2}$ for the curvature bounds. Thus the radius condition reads $r \le \pi\tau$. Then we have:

$$\frac{\sin\sqrt{\kappa_+}r}{\sqrt{\kappa_+}r} = \frac{\sin\tau^{-1}r}{\tau^{-1}r} = 1 - \frac{r^2}{6\tau^2} + O(r^4) \ge 1 - \frac{r^2}{6\tau^2}$$
$$\frac{\sin\sqrt{\kappa_-}r}{\sqrt{\kappa_-}r} = \frac{\sinh\sqrt{2}\tau^{-1}r}{\sqrt{2}\tau^{-1}r} = 1 + \frac{r^2}{3\tau^2} + O(r^4) \le 1 + \frac{r^2}{2\tau^2} \text{ for } r \le 2\tau$$

where in the end we used $\sinh x \le x + \frac{x^3}{4}$ for $x \in [0, 2\sqrt{2}]$[18]. □

LEMMA 6.14. *For a metric space M and its n-fold product space M^n, the following function is a metric on M^n:*

$$\mathrm{d}_\circ(x,y) := \min_{\sigma,\tau\in S_n} \mathrm{d}_M(\sigma\cdot x, \tau\cdot y) = \min_{\sigma\in S_n} \mathrm{d}_M(x, \sigma\cdot y)$$

where S_n is the permutation group on n elements and $\sigma\cdot(y_1,\dots y_n) = (y_{\sigma(1)},\dots y_{\sigma(n)})$ permutes the coordinates. If $M = \mathbb{R}$, $x, y \in M$, and if entries of x, y are arranged in the decreasing order, then

$$\mathrm{d}_\circ(x,y) = \|x - y\|$$

PROOF. Reflexivity and symmetry of $d_\circ$ hold obviously. To see the triangle inequality, suppose that $x, y, z \in M^D$ and define σ_{xy} by the relation $d_\circ(x,y) = \mathrm{d}_M(x, \sigma_{xy}\cdot y)$ (similarly for σ_{yz}, σ_{xz}). Then

$$\begin{aligned}\mathrm{d}_\circ(x,y) + \mathrm{d}_\circ(y,z) &= \mathrm{d}_M(x,\sigma_{xy}\cdot y) + \mathrm{d}_M(y,\sigma_{yz}\cdot z)\\ &= \mathrm{d}_M(x,\sigma_{xy}\cdot y) + \mathrm{d}_M(\sigma_{xy}\cdot y, \sigma_{xy}\cdot\sigma_{yz}\cdot z)\\ &\ge \mathrm{d}_M(x, \sigma_{xy}\cdot\sigma_{yz}\cdot z)\\ &\ge \mathrm{d}_\circ(x,z)\end{aligned}$$

This shows that $\mathrm{d}_\circ$ is indeed a metric.

Consider $M = \mathbb{R}$. Suppose that $x_1 \le \cdots \le x_n, y_1 \le \cdots \le y_n$. Then we claim that for any $\sigma \in S_n$, $\|x - y\| \le \|x - \sigma\cdot y\|$. Suppose $z \in \mathbb{R}^n$ doesn't necessarily have its entries ordered in a decreasing order. If there exists a pair $i < j$ with $z_i > z_j$, then we have: $\|x - \tau_{ij}\cdot z\| < \|x - z\|$, where $\tau_{ij} \in S_n$ is the transposition that swaps i and j. This is because whenever $a < b, a' < b'$, we have $(a-a')^2 + (b-b')^2 < (a-b')^2 + (b-a')^2$. By repeatedly applying this sorting process to $z = \sigma\cdot y$, we get the claim. The sorting process ends in finite time because one can recursively take the smallest unsorted element and swap it all the way down, i.e. perform a bubble sort. □

[18] This can be manually checked by computing the first and the second derivative of $x + x^3/4 - \sinh x$.