Tangent Space and Dimension Estimation with the Wasserstein Distance

Uzu Lim, Harald Oberhauser, and Vidit Nanda

Abstract. We provide explicit bounds on the number of sample points required to estimate tangent spaces and intrinsic dimensions of (smooth, compact) Euclidean submanifolds via local principal component analysis. Our approach directly estimates covariance matrices locally, which simultaneously allows estimating both the tangent spaces and the intrinsic dimension of a manifold. The key arguments involve a matrix concentration inequality, a Wasserstein bound for flattening a manifold, and a Lipschitz relation for the covariance matrix with respect to the Wasserstein distance.

1. Introduction

Much of modern data science relies on the assumption that the true distribution underlying a given data set concentrates near a manifold. Here, 'manifold' refers to a smoothly embedded compact submanifold M of an ambient Euclidean space $\mathbb{R}^D$, with $\dim M$ considerably smaller than D. This *manifold hypothesis* has been the subject of active mathematical study; investigations include topological inference from finite samples [20], hypothesis testing [9], and fitting manifolds to data [8], among many others. Furthermore, many dimensionality reduction techniques rely on the manifold hypothesis for their success — see for instance Local Tangent Space Alignment [30] and Uniform Manifold Approximation and Projection (UMAP) [18].

Our goal in this paper is to provide rigorous and explicit bounds on the number of sample points required to estimate tangent spaces and intrinsic dimensions of smooth manifolds with high confidence. The estimators arise from a local version of *principal component analysis* (PCA). The local principal components approximate the tangent space at the given point, whereas the associated eigenvalues allow us to infer the intrinsic dimension. Our main contribution here is the derivation of probabilistic bounds for Local PCA-based tangent space and dimension estimation. Crucially, these bounds adapt to noisy non-uniform distributions concentrated near a manifold, and all the relevant constants have been computed explicitly.

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.

Estimation with Local PCA. Let $\mathbf{x} = \{x_1, \dots x_m\}$ be points in $\mathbb{R}^D$ and let $\bar{x} = \frac{1}{m}\sum_i x_i$, both written as column vectors. PCA refers to the following diagonalization[1]:

$$\hat{\Sigma}[\mathbf{x}] = U\Lambda U^\top, \text{ where } \hat{\Sigma}[\mathbf{x}] = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})(x_i - \bar{x})^\top.$$

Here U is an orthogonal matrix and Λ is a diagonal matrix with entries $\lambda_1 \geq \dots \geq \lambda_D \geq 0$. We are interested in the following quantities obtained from PCA:

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

where k lies in $\{1, 2, \dots, D\}$, while $v_1, \dots v_k$ are the first k columns of U and $\vec{\lambda}A$ denotes the eigenvalues of a real symmetric matrix A, arranged in the decreasing order. We call v_k the *k-th principal component* of $\mathbf{x}$.

Local PCA at an open set $U \subseteq \mathbb{R}^D$ refers to performing PCA on points of $\mathbf{x}$ which lie in U. Given a radius parameter $r > 0$, we may perform Local PCA at the open balls $\mathcal{B}_r(x_i)$ of radius r. The points of interest are:

$$\mathbf{x}^{(i)} = \{x_j\}_{j \neq i} \cap \mathcal{B}_r(x_i)$$

We then define the *k-dimensional tangent space estimator* and *intrinsic dimension estimator* at point x_i as follows:

$$\hat{\Pi}_k^{(i)} := \Pi_k[\mathbf{x}^{(i)}]$$
$$\hat{d}^{(i)} := \operatorname{argmin}_k \left\| \frac{1}{r^2}\vec{\lambda}\hat{\Sigma}[\mathbf{x}^{(i)}] - \vec{\lambda}(k, D) \right\|$$
$$\text{where } \vec{\lambda}(k, D) = \frac{1}{k+2}(\underbrace{1, \dots 1}_{k}, \underbrace{0, \dots 0}_{D-k}) \tag{1.1}$$

Here, $\vec{\lambda}(k, D)$ are eigenvalues of (the covariance matrix for) the uniform distribution over a k-dimensional unit disk embedded in $\mathbb{R}^D$ (see Lemma 6.1). Thus, the estimator $\hat{d}^{(i)}$ determines for which k the sample is the closest to a k-dimensional unit disk.

[1] The empirical covariance matrix we consider is a biased estimator in which we take $\frac{1}{m}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$ instead of the unbiased estimator $\frac{1}{m-1}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$. This is not an issue when m goes to infinity; the relevant notions are proved precisely in Section 2.

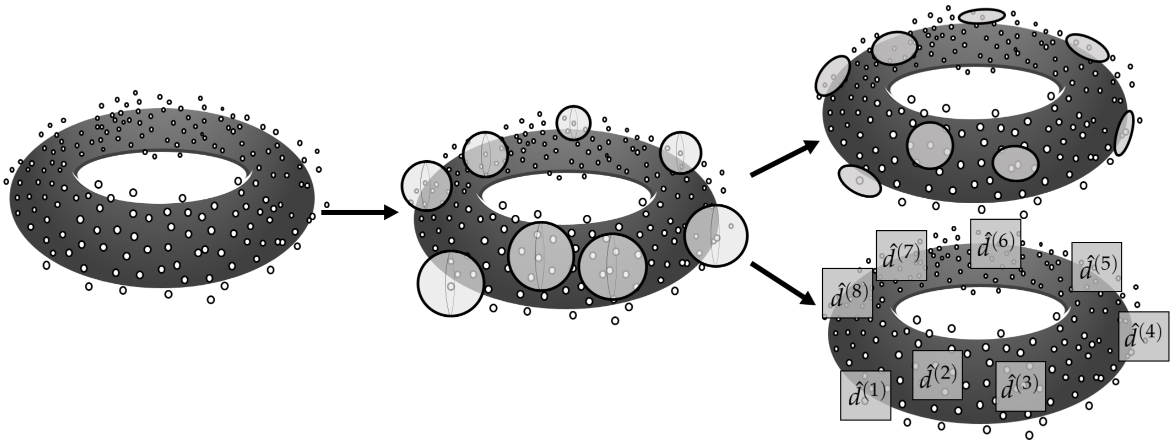

Figure 1. Local PCA on a dataset concentrated near a torus. The two diagrams on the right respectively indicate tangent space estimation and intrinsic dimension estimation.

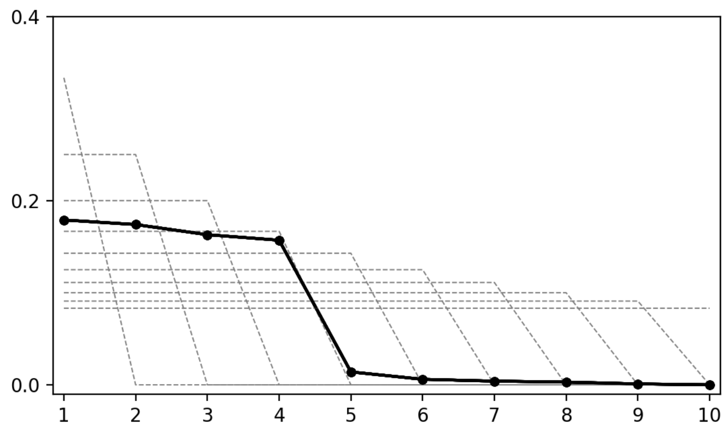

Figure 2. An illustration of the dimension estimation process. The dotted lines are plots of $\vec{\lambda}(d, D)$ for $D = 10$ and $d = 1, \ldots 10$. The solid line is a plot of empirically obtained eigenvalues, which is close to $\vec{\lambda}(4, 10)$, indicating that the estimated intrinsic dimension is 4.

Probabilistic guarantees. When the estimators $\hat{\Pi}_d^{(i)}$ and $\hat{d}^{(i)}$ are calculated for a sample drawn from a probabiltiy distribution on a d-dimensional manifold, we expect that they will respectively estimate the tangent spaces and the dimension d of the manifold. 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 set the stage for our main theorems, which show that $\hat{\Pi}_d^{(i)}$ and $\hat{d}^{(i)}$ work as expected. Let $M \subset \mathbb{R}^D$ be a smoothly embedded d-dimensional compact manifold. Denote by τ the reach of M, which is the maximum length to which M can be thickened normally without self-intersection. Let μ_0 be a Borel probability measure on $\mathbb{R}^D$ defined using a probability density function $\varphi : M \to \mathbb{R}^+$: 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. We assume that $\int_M \varphi \, \mathrm{d}\,\mathcal{H}^d = 1$ so that μ_0 is a probability measure. We also assume that φ satisfies the Lipschitz condition $\|\varphi(x) - \varphi(y)\| \le \alpha \cdot \mathrm{d}_M(x, y)$, where d_M is the geodesic distance on M. Let $X \sim \mu_0$, and let Y be a random variable valued in $\mathbb{R}^D$ with bounded norm $\|Y\| \le s$. Here Y represents noise and s represents the noise radius. Finally, define the measure of interest as the law

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

Here onwards, ω_d denotes the volume of the unit d-dimensional ball:

$$\omega_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}$$

Let $\mathbf{X} = (X_1, \ldots X_m)$ be an i.i.d.[2] sample drawn from μ and let $X_1^\perp, \ldots X_m^\perp$ be their orthogonal projections to M. Then we have the following guarantee for tangent space estimation:

Theorem A (Tangent Space Estimation). Suppose θ, δ are positive real numbers and k is a positive integer. Let $\epsilon = \sin\theta$. If r and m satisfy:

$$\sqrt{\frac{2s}{\tau}} < \frac{r}{\tau} < \frac{\epsilon}{c_1} \quad \text{and} \quad m \ge c_2 \cdot \log\left(\frac{(4D+2)k}{\delta}\right) + 1$$

Then with probability at least $1 - \delta$, the following holds for every $i \le k$:

$$\measuredangle\left(\hat{\Pi}_d^{(i)}, T_{X_i^\perp} M\right) \le \theta$$

where $\hat{\Pi}_d^{(i)}$ is defined in Equation (1.1), and $\measuredangle(\Pi_1, \Pi_2)$ is the principal angle between subspaces Π_1, Π_2 (see Definition 5.6). The constants c_1, c_2 are given by:

$$c_1 = c_1(d, \varphi, \tau) = 16(d+2)\left(3 + \frac{8\varphi_{\max} d + 5\alpha\tau}{\varphi_{\min}}\right)$$

$$c_2 = c_2(d, \varphi, s, r, \epsilon) = \frac{1}{\omega_d (r - 2s)^d \varphi_{\min}} \cdot \frac{4794(d+2)^2}{\epsilon^2} \tag{1.2}$$

where $\varphi_{\max} = \max_{x \in M} \varphi(x)$ and $\varphi_{\min} = \min_{x \in M} \varphi(x)$.

The constants above can be improved in several ways:

[2] i.i.d. stands for 'independent and identically distributed'.

- The constants c_1, c_2 above can be improved by replacing each occurrence of $\varphi_{\min}$ by 1.04Φ, where Φ is defined as:

$$\Phi = \inf_{x \in M} \frac{\mu_0(\exp_x \mathring{\mathcal{B}}_{r-2s})}{\omega_d (r-2s)^d}$$

Here $\exp_x \mathring{\mathcal{B}}_{r-2s}$ is the geodesic ball of radius $r-2s$ centered at x. The intuitive meaning of Φ is a lower bound of local concentration of the measure μ_0. Replacing $\varphi_{\min}$ by 1.04Φ avoids division by zero in case $\varphi_{\min}$ vanishes at a small region. This replacement is possible since $\varphi_{\min} < 1.04\Phi$. Note that the constant 4794 in c_2 appears by rounding up 1.04×4609, with 4609 coming from Lemma 5.4. See the proof of Theorem A in Section 5 for details.
- The condition given by two inequalities between s, r, ϵ can be collectively replaced by the following (single) weaker condition:

$$Q\left(\frac{r}{\tau}, \frac{s}{\tau}\right) \le \frac{r\epsilon}{16(d+2)\tau}$$

where Q is a function defined in Proposition 4.4. See the curvature control part of the proof of Theorem 5.3 to see how this modification may be done.

We now state the guarantee for intrinsic dimension estimation:

Theorem B (Intrinsic Dimension Estimation). Suppose δ is a positive real number and k is a positive integer. Let

$$\epsilon = \frac{\sqrt{(d+1)(d+4)}}{2\sqrt{D}(d+3)}$$

If r and m satisfy:

$$\sqrt{\frac{2s}{\tau}} < \frac{r}{\tau} < \frac{\epsilon}{c_1} \quad \text{and} \quad m \ge c_2 \cdot \log\left(\frac{(4D+2)k}{\delta}\right) + 1$$

Then with probability at least $1-\delta$, the following holds for every $i \le k$:

$$\hat{d}^{(i)} = d$$

where $\hat{d}^{(i)}$ is defined in Equation (1.1). Here, c_1, c_2 are defined as in Theorem A, only with a different prescription of ϵ in c_2.

We make a few remarks regarding the main results.

- We emphasize that the noise term Y is *not* assumed to be independent of $X \sim \mu_0$, when defining the probability measure of interest $\mu = \mathrm{Law}(X+Y)$.

- In the case of uniform distribution, we note that the constants c_1, c_2 in Theorem A become easily interpretable:

$$c_1 = 16(d+2)(8d+2) \approx 128d^2$$
$$c_2 = \left(\frac{\omega_d (r-2s)^d}{\mathrm{Vol}_M}\right)^{-1} \cdot \frac{4794(d+2)^2}{\epsilon^2}$$

  where Vol_M is the d-dimensional Hausdorff measure of M. In particular, the fraction $\omega_d (r-2s)^d / \mathrm{Vol}_M$ indicates how large a d-dimensional disk of radius $r-2s$ is, compared to the manifold M.
- Both Theorems require both a lower and upper bound for the radius r. This reflects the observation that r must be large enough to overcome the effects of noise, but small enough to ignore the effects of curvature.
- Theorem A relies on the knowledge of the true intrinsic dimension d. This quantity can be either obtained using Theorem B, or other methods for intrinsic dimension estimation, such as in [5].
- Constants in Theorem B depend on the intrinsic dimension d, which is the quantity to be estimated. This doesn't make the theorem circular, since d is a well-defined quantity. Nevertheless, a practitioner will not know d a priori. Therefore for a practical application, d in the constants can be replaced by D, or any d' that is a priori known to be greater: $d \leq d'$.
- Theorem A and B are corollaries of Theorem 5.3, and this can be seen as the main theorem of our work. Theorem 5.3 gives probabilistic guarantees for estimating covariance matrices locally, for a probability distribution concentrated near a manifold.
- A conventional method of estimating intrinsic dimension from eigenvalues is by testing how many principal components account for (say) 95% of the total variance. This is stated and proven in Theorem B′. This method introduces additional complexity because the threshold parameter must also fall within a certain range.

Structure of the paper. We prove the building blocks of the paper in Sections 2 to 4, and then derive the main Theorem 5.3 and its corollaries Theorem A and B in Section 5. More specifically,

- **Section 2:** We modify the matrix Hoeffding's inequality to show that Local PCA correctly estimates covariance matrix of the underlying distribution (Proposition 2.8).

- **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 $\|\Sigma[\mu]-\Sigma[\nu]\| \le C \cdot W_p(\mu,\nu)$ where $\Sigma[\mu]$ is the covariance matrix of μ (Proposition 3.3).
- **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).

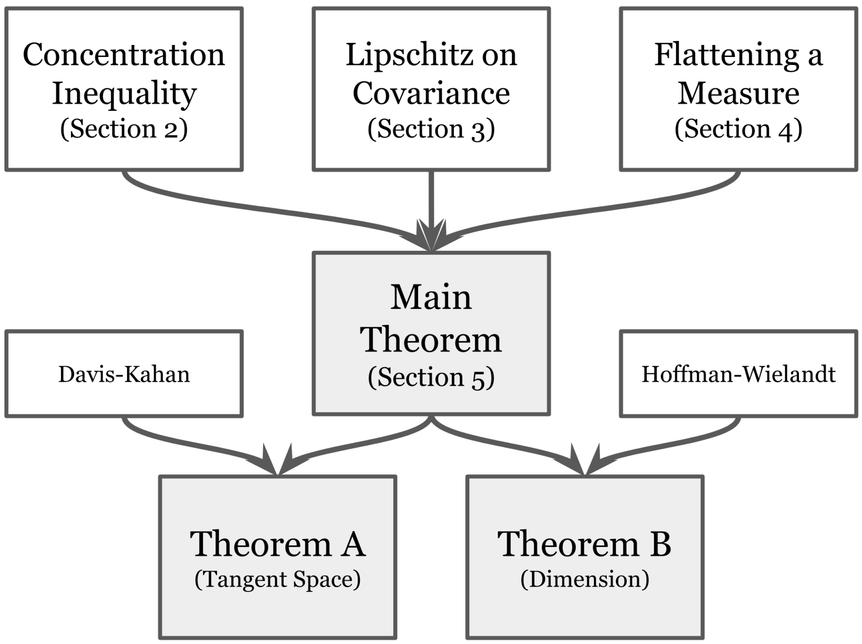

Figure 3. Summary of the relations between the main results.

We summarize the notations and conventions of this article in the Appendix (page 33).

Related work. Probabilistic bounds on tangent space estimation using Local PCA have been studied in considerable detail, for example in [2, 26, 13, 23]. To the best of our knowledge, our work is the first in which probabilistic guarantees for tangent spaces and dimension estimation were produced with *all relevant constants* in the probabilistic bounds computed explicitly, while *allowing for noise and non-uniformity*. A major advantage offered by the constants we computed is its interpretability; see Equation (1.2). In [13] and [26], the underlying probability measure is assumed to be uniform, and only estimation at a single point is considered (instead of simultaneous estimation at multiple points). In [2] and [23], various constants have not been computed explicitly.

Our concentration inequality for covariance matrices (Proposition 2.6 in this paper) is directly derived from the matrix Hoeffding inequality, which appears in [25]. A more sophisticated approach, such as the one from [14], may be used to generalize our results. Similar methods for analyzing (non-local, non-manifold) PCA are also studied in [15, 21].

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 [4]. We also obtain a similar inequality (in Proposition 3.3 and Corollary 4.5 below). The key difference in the two derivations is that our approach uses the Wasserstein distance rather than the total variation distance from [4], 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 the same idea as the proof of Proposition 3.1 in [24]. However, their transportation plan doesn't involve noise and applies to different types of local covariance matrices.

More recently, the authors of [3] have used local polynomial regression 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. On the other hand, our result only estimates tangent spaces, and does not produce any polynomial approximations which may be used to estimate higher-order information (such as curvature).

In [5], an intrinsic dimension estimator that doesn't use Local PCA is introduced. We note that the number of points we require to ensure a $1 - \delta$ probability of dimension estimation has the rate of $m \sim \log(1/\delta)$, which improves the rate $m \sim \log(1/\delta)^3$ in [5].

Finally, we note that local PCA has been extensively used in contexts independent of the manifold hypothesis [10, 12, 27, 19]. As far as we are aware, in all cases the theoretical analysis is either heuristic or makes strong assumptions on the underlying distribution (e.g., by requiring that the data be Gaussian).

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. 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 2.8, where we establish bounds for local covariance estimation. Our main tool is the *matrix Hoeffding inequality* [25, Theorem 1.3][3]. Here onwards, we will use $\|A\|$ to denote the operator norm of a given matrix A, i.e.,

$$\|A\| := \sup_{\|x\|=1} \|Ax\|.$$

[3]Our version of the matrix Hoeffding inequality follows from the one in [25] 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.

Theorem 2.1 (Matrix Hoeffding). *Let $Y_1, \dots Y_m$ be independent Hermitian random $D \times D$ matrices so that for each i in $\{1, \dots, m\}$ we have both $\mathbb{E}Y_i = 0$ and $\|Y_i\| \leq \alpha_i$ almost surely for some real number $\alpha_i \geq 0$. Then writing $Y = \sum_1^m Y_k$, for every $\epsilon \geq 0$ we have*

$$\Pr\Big(\|Y\| \geq \epsilon \Big) \leq 2D \cdot \exp\left(\frac{-\epsilon^2}{8\sigma^2}\right),$$

where $\sigma^2 = \sum_1^m \alpha_k$.

This inequality can be used to establish concentration of vectors under 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 one obtains the following consequence.

Corollary 2.2. *Let $X_1, \dots X_m$ be independent random vectors in $\mathbb{R}^D$ satisfying $\mathbb{E}X_i = 0$, and $\|X_i\| \leq \alpha_i$ almost surely for some real number α_i. Writing $Y = \sum_1^m X_k$, for every $\epsilon \geq 0$ we have*

$$\Pr\Big(\|Y\| \geq \epsilon \Big) \leq 2(D+1) \cdot \exp\left(\frac{-\epsilon^2}{8\sigma^2}\right)$$

where $\sigma^2 = \sum_1^m \alpha_i^2$.

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

Definition 2.3. 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 $\mu(U) > 0$.

Definition 2.4. Let $\mathbf{x} = (x_1, \dots x_m) \subset \mathbb{R}^D$ and let $U \subseteq \mathbb{R}^D$ be a Borel set. Then we define:

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

$$N_{\mathbf{x},U} := \sum_{i=1}^m \mathbf{1}_{x_i \in U}$$

$$\delta_{\mathbf{x}|U} := \begin{cases} \frac{1}{N_{\mathbf{x},U}} \sum_{i=1}^m \mathbf{1}_{x_i \in U} \cdot \delta_{x_i} & \text{if } N_{U,\mathbf{x}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

where δ_x is the Dirac delta measure at x and $\mathbf{1}_*$ is the indicator function.

In elementary terms, $\delta_{\mathbf{x}|U}$ averages the Dirac delta measures at all those points x_i in $\mathbf{x}$ which lie within U.

Definition 2.5. Given $X \sim \mu$, the *covariance matrix* (or simply the *covariance*) of μ is the $D \times D$ matrix defined by:

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

If $\mathbf{X} = (X_1, \ldots X_m)$ is μ-i.i.d. sample, the covariance of the measure $\delta_{\mathbf{X}}$ evaluates as follows:

$$\Sigma[\delta_{\mathbf{X}}] = \frac{1}{m} \sum_{i=1}^{m} (X_i - \bar{X})(X_i - \bar{X})^\top$$

where $\bar{X} = \frac{1}{m} \sum_i X_i$ is the sample mean. We recall that the expected values of

$$\frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})(X_i - \bar{X})^\top \text{ and } \frac{1}{m} \sum_{i=1}^{m} (X_i - \mathbb{E}X)(X_i - \mathbb{E}X)^\top$$

are both equal to $\Sigma[\mu]$ whereas the expected value of $\Sigma[\delta_{\mathbf{X}}]$ is $\frac{m-1}{m}\Sigma[\mu]$. Nevertheless, the following result shows that we can use $\Sigma[\delta_{\mathbf{X}}]$ to estimate $\Sigma[\mu]$.

Proposition 2.6 (Concentration inequalities for covariance). *Let μ be a Borel probability measure on $\mathbb{R}^D$ and let $\mathbf{X} = (X_1, \ldots X_m)$ 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 $\epsilon \geq 0$,*

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

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

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

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

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

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

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

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

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

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. □

Let $\mathbf{X} = (X_1, \ldots X_m)$ be a μ-i.i.d. sample. We will consider estimating $\Sigma[\mu|_U]$ with $\Sigma[\delta_{\mathbf{X}|U}]$ in the special case where the Borel set U is entirely contained within an open ball of some fixed radius $r > 0$.

Proposition 2.7. *Let $\mathbf{X} = (X_1, \ldots X_m)$ 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:*

$$\Pr\left(\|\hat{\Sigma}_U-\Sigma_U\|\leq\epsilon\right)\geq 1-\delta,$$

where

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

(Note that $\delta \to 0$ as $m \to \infty$.)

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

$$\Pr\left(\|\hat{\Sigma}_U-\Sigma_U\|\geq\epsilon\right)=\sum_{I\subseteq\{1,\ldots m\}}\Pr\left(\|\hat{\Sigma}_U-\Sigma_U\|\geq\epsilon|\mathcal{S}_I\right)\cdot\Pr(\mathcal{S}_I).$$

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

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

Here Proposition 2.6 was applied in the only inequality above. Note that the possibility $\mathcal{S}_\emptyset$ is correctly accounted for since we have included $k = 0$ when indexing the sum in the second line above. □

If we apply the previous proposition to the special case where U is the open ball $\mathcal{B}_r(X)$ of radius r around $X \sim \mu$, then we arrive at the main result of this section.

Proposition 2.8. *Assume that μ is 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 ϵ and δ be positive real numbers and let k be a positive integer. If the sample size m satisfies:*

$$m \geq \gamma \cdot \log\left(\frac{(4D+2)k}{\delta}\right) + 1$$

then we have:

$$\Pr\left(\|\hat{\Sigma}_i - \Sigma_i\| \leq \epsilon \textit{ for all } i \leq k\right) \geq 1 - \delta.$$

Here $\gamma = -1/\log(1 - \gamma_1\gamma_2)$, where

$$\gamma_1 := \inf_{x \in K} \mu(\mathcal{B}_r(x)) \quad \textit{and} \quad \gamma_2 := 1 - \exp\left(\frac{-\epsilon^2}{1152 r^4}\right).$$

Furthermore, $\inf_{x \in K} \mu(\mathcal{B}_r(x)) > 0$ holds and thus $\gamma > 0$.

Proof. For each i, let U_i be the ball $\mathcal{B}_r(x_i)$ and define the set $E_i \subseteq (\mathbb{R}^D)^m$ as:

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

where $\mathbf{x}_i = \{x_j | j \neq i\}$. By the union bound, symmetry, and Proposition 2.7, we then have:

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

where $u_x = \mu(\mathcal{B}_r(x))$, $\xi = \exp(-\epsilon^2/1152 r^4)$, 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:

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

The condition on m now follows by forcing the right hand side of (2.1) to be $\leq \delta$:

$$(4D+2)k(1 - u_0(1-\xi))^{m-1} \leq \delta$$

By taking logarithm and rearranging terms, we get:

$$m - 1 \geq \frac{\log\left(\frac{(4D+2)k}{\delta}\right)}{-\log(1 - u_0(1-\xi))} = \gamma \cdot \log\left(\frac{(4D+2)k}{\delta}\right)$$

as desired.

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 $x \in \mathcal{B}_{r/2}(y)$. The triangle inequality guarantees a containment $\mathcal{B}_{r/2}(y) \subseteq \mathcal{B}_r(x)$, from which we obtain $\mu(\mathcal{B}_{r/2}(y)) \le \mu(\mathcal{B}_r(x))$ and hence

$$\inf_{y\in J} \mu(\mathcal{B}_{r/2}(y)) \le \inf_{x\in K} \mu(\mathcal{B}_r(x)).$$

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 [28] on its domain, defined as follows. Let (M, d_M) be a Polish metric space equipped with probability measures μ and ν. For each $p \ge 1$, 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 $1 \le p \le q$, we have $\mathrm{W}_p(\mu,\nu) \le \mathrm{W}_q(\mu,\nu)$ 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 $p \ge 1$,*

(1) $\|\mathbb{E}X - \mathbb{E}Y\| \le \mathrm{W}_p(\mu,\nu)$ where $X \sim \mu$ and $Y \sim \nu$.
(2) $\mathrm{W}_p(\tilde{\mu},\tilde{\nu}) \le 2 \cdot \mathrm{W}_p(\mu,\nu)$

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

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

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

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

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 $p \ge 1$:*

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

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

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

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 $p \ge 1$, we have the following Lipschitz property:*

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

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 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, $\operatorname{supp}\mu \subseteq \mathcal{B}_1(x) \subseteq \mathcal{B}_2(0)$.

Perform diagonalization as follows:

$$S := \Sigma[\mu] - \Sigma[\nu] = U\Lambda U^\top$$

where $U = [u_1, \dots u_D]$ is orthogonal and Λ is a diagonal matrix with entries $\lambda_1 \geq \cdots \geq \lambda_D \geq 0$. Then the operator norm of S is λ_1, which can be written as:

$$\begin{aligned}
\|S\| = \lambda_1 &= (U^\top S U)_{1,1} \\
&= \mathbb{E}[U^\top X X^\top U]_{1,1} - \mathbb{E}[U^\top Y Y^\top U]_{1,1} \\
&= \mathbb{E}(U^\top X)_1^2 - \mathbb{E}(U^\top Y)_1^2
\end{aligned}$$

where $A_{1,1}$ refers to the $(1,1)$th entry of a matrix A and w_1 refers to the 1st entry of a vector w. We furthermore get:

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

where $U^\top \mu = \mathrm{Law}(U^\top X)$ and $(U^\top \mu)_1$ denotes the marginal of $U^\top \mu$ at its 1st 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.

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 [22] on the metric space:

$$\mathcal{H}^d(U) = \lim_{\delta \downarrow 0} \mathcal{H}^d_\delta(U),$$

$$\text{where } \mathcal{H}^d_\delta(U) = \frac{\omega_d}{2^d} \inf_{\substack{\mathrm{diam}(C_j)<\delta \\ U \subseteq \cup C_j}} \left(\sum_{j=1}^{\infty} \mathrm{diam}(C_j)^d \right) \text{ and } \omega_d := \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$$

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 $\mathrm{Unif}(U)$ the uniform probability measure over U with respect to $\mathcal{H}^d$; for each V,

$$\mathrm{Unif}(U)(V) = \frac{\mathcal{H}^d(U \cap V)}{\mathcal{H}^d(U)}$$

Recall that in $\mathbb{R}^D$, the Hausdorff measure $\mathcal{H}^D$ agrees with the Lebesgue measure.

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) the *metric* d_M is defined as follows: and points $x, y \in M$, define the metric d_M by letting $\mathrm{d}_M(x, y)$ be the infimum of lengths of all piecewise regular[4] curves that connect x and y. Equivalently, $\mathrm{d}_M(x, y)$ is the length of the shortest geodesic connecting x and y.[5]

(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 $\mathrm{d}_{\mathbb{R}^D}(x, M) = \inf_{y \in M} \mathrm{d}_{\mathbb{R}^D}(x, y)$.

[4] A regular curve is a continuously differentiable curve with nonvanishing velocity.

[5] This follows from the Hopf-Rinow Theorem; see Corollary 6.21 and 6.22 in [16].

(3) For each point $x \in M$, we denote by $\mathring{\mathcal{B}}_r \subseteq T_xM$ the open ball of radius r around $0 \in T_xM$, 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 $v \in T_xM$ to the endpoint of the unique geodesic on M starting at x with the initial velocity of v.

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 $p \geq 1$:*

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

Furthermore, we may take:

$$Q(\rho, \sigma) = 3\sigma + (\rho + 2\sigma)^2 + \frac{1.18\varphi_{\max}}{\Phi}(2\rho + (\rho + 2\sigma)^2)(1 - \Omega^d) + \frac{2.18\rho}{\Phi}(\varphi_{\max} - \varphi_{\min}) + 1.38\rho^3$$

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.

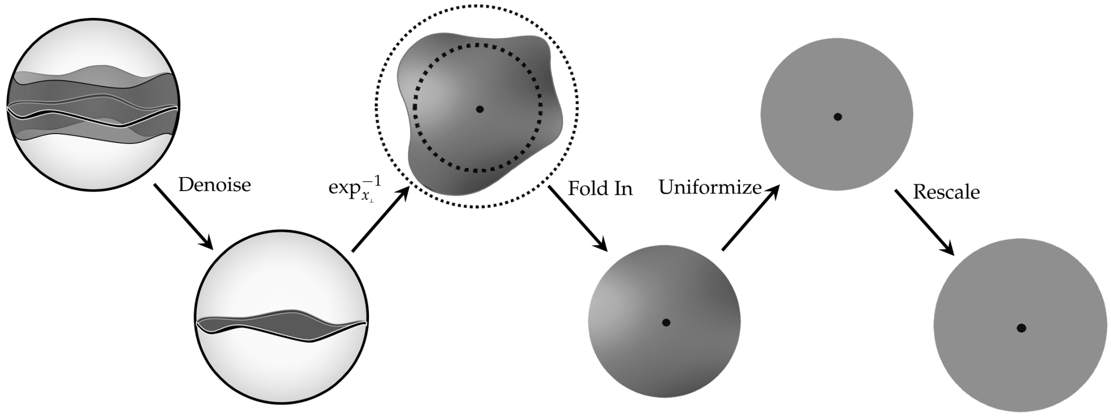

FIGURE 4. An overview of the transportation plan in the proof of Proposition 4.4. The last four sub-diagrams take place on the tangent space. Nonuniform shadings in the 3rd, 4th sub-diagrams indicate nonuniform probability distribution.

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

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

where $\mu(\mathcal{B}_r(x)) = \Pr(X+Y \in \mathcal{B}_r(x))$, which follows by the definition of μ. The transportation cost is bounded as $\mathrm{W}_p(\nu_0, \nu_1) \leq \mathbb{E}_{(X+Y,X)\sim\nu_{01}} \|(X+Y) - X\| \leq s$. 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 $\|Y\| \leq s$, 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 $\|x - x_\perp\| \leq s$, 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:

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

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

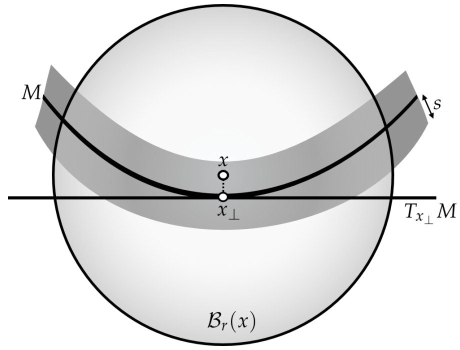

FIGURE 5. Measure μ and its restriction $\mu|_{\mathcal{B}_r(x)}$, where $x \in \mathbb{R}^D$ and $x_\perp \in M$.

Step 2. We define ν_2 by pushing ν_1 forward 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 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 $0 \in T_zM$. It is known that the injectivity radius is at least $\pi \cdot \tau$ (Proposition A1 of [1]). Meanwhile, Lemma 6.10 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 6.9, the transportation cost then is bounded as:

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

Thus by Equations (4.2) and (4.3),

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

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[6], 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,

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

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

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

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 $0 \in T_xM$ 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

[6] See for example [7] for a standard reference in geometric measure theory

cost $W_p(\nu_2, \nu_3)$ 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 $\int \nu_2^{\text{out}}$.

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 $\nu_4 = \mathrm{Unif}(\mathring{\mathcal{B}}_{r_{\text{in}}})$, we can directly apply Lemma 6.5:

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

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 $\mathring{\mathcal{B}}_{r_{\text{in}}}$. Writing $\psi^{(1)} := \varphi \circ \exp_{x_\perp}$ and $\psi^{(2)} := \mathrm{J} \exp_{x_\perp}$ so that $\psi = \psi^{(1)} \psi^{(2)}$, the variation $\psi_{\max} - \psi_{\min}$ can be controlled with the triangle inequality as follows:

$$\begin{aligned}
|\psi_{\max} - \psi_{\min}| &\leq |\psi_{\max}^{(1)} \psi_{\max}^{(2)} - \psi_{\min}^{(1)} \psi_{\min}^{(2)}| \\
&\leq |\psi_{\max}^{(1)} \psi_{\max}^{(2)} - \psi_{\min}^{(1)} \psi_{\max}^{(2)}| + |\psi_{\min}^{(1)} \psi_{\max}^{(2)} - \psi_{\min}^{(1)} \psi_{\min}^{(2)}| \\
&= |\psi_{\max}^{(1)} - \psi_{\min}^{(1)}| \cdot |\psi_{\max}^{(2)}| + |\psi_{\min}^{(1)}| \cdot |\psi_{\max}^{(2)} - \psi_{\min}^{(2)}| \\
&\leq (\varphi_{\max} - \varphi_{\min}) \left(1 + \frac{r_{\text{in}}^2}{2\tau^2}\right) + \varphi_{\min} \frac{2 r_{\text{in}}^2}{3\tau^2}.
\end{aligned}$$

Here the extrema of φ are taken over the geodesic ball $\exp_{x_\perp}(\mathring{\mathcal{B}}_{r_{\text{in}}})$. In the last inequality above we used Corollary 6.12, 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:

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

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 6.8), so that we get another uniform distribution. By Lemma 6.3, the transportation cost is bounded by $r - r_{\text{in}} = 2s$.

The Total Bound. Collecting the bounds[7], we get:

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

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

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

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

where $\varphi_{\max}$ is the maximum of φ taken over $\mathcal{B}_{r+2s}(x_\perp)$.[8] Combining these, we get:

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

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

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

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

where the first inequality holds by plugging in the upper bound for $\int \nu_2^{\text{out}} / \int \nu_2^{\text{in}}$, and the second inequality holds since $\Phi' \geq 1$. Plugging these into Equation (4.9), we get that

[7] We use a slight abuse of notation and identify ν_k with $\iota_* \nu_k$ for $k = 2, \ldots 5$, 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 $T_{x_\perp} M$.

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

$W_p(\nu_0, \nu_5)$ is no larger than

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

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

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

Plugging these numbers into our bound above for $W_p(\nu_0, \nu_5)$ 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 $p \geq 1$:

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

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

$$Q_1(\rho, \beta) = 3\beta + \beta_1^2 + \frac{1.18\varphi_{\max} d}{\Phi}(2 + \beta_1^2\rho)(1 + 4\beta) + \frac{4.36\alpha\tau}{\Phi}(1 + \beta_1\rho)\beta_1 + 1.38\rho$$
$$\text{and } \beta_1 = 1 + 2\beta\rho$$

In particular, for $\beta = 1/2$, we have:

$$Q_2(\rho) := Q_1(\rho, \frac{1}{2}) = (2.5 + 3.38\rho + \rho^2) + \frac{3.54\varphi_{\max} d}{\Phi}(2 + \rho + 2\rho^2 + \rho^3) + \frac{4.36\alpha\tau}{\Phi}(1 + 2\rho + 2\rho^2 + \rho^3)$$

Proof. We first have:

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

where the first line is by the definition of β_1 and the second line is by Lemma 6.6. This almost derives Q_1, except the bound on $\varphi_{\max} - \varphi_{\min}$. Since geodesic distance is used, the Lipschitz assumption on φ implies:

$$\varphi_{\max} - \varphi_{\min} \leq 2\alpha r_{\mathrm{out}}$$

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 6.6. 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 2.8, 3.3, and 4.4 to prove Theorem 5.3. This in turn implies both Theorem A and B from the Introduction.

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 $\mathbf{X} = (X_1, \ldots X_m)$ 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 $(\mu, \mu_0) \in \mathcal{P}(M, s)$ 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 $X_1, \ldots X_m$ 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 positive real numbers δ, ϵ, k where $k \in \mathbb{Z}$, suppose r, m satisfy the following:

$$r + 2s \leq (\sqrt{2} - 1)\tau$$
$$\sqrt{\frac{2s}{\tau}} < \frac{r}{\tau} < \frac{\epsilon}{16(d+2)Q_2(r/\tau)}$$
$$m \geq \gamma \cdot \log\left(\frac{(4D+2)k}{\delta}\right) + 1$$

Then with probability at least $1 - \delta$, the following holds for every $i \leq k$:

$$\left\|\hat{\mathrm{P}}_i - \mathrm{P}_i\right\| \leq \epsilon$$

where $\mathrm{P}_i = \mathrm{P}_{T_{X_i^\perp}M}$ is the projection to the tangent space $T_{X_i^\perp}M$,

$$\gamma = \frac{-1}{\log(1 - \gamma_1\gamma_2)} > 0, \text{ with } \gamma_1 = \inf_{x \in \operatorname{supp}\mu} \mu(\mathcal{B}_r(x)),\ \gamma_2 = 1 - \exp\left(\frac{-\epsilon^2}{4608(d+2)^2}\right)$$

and $Q_2(r/\tau)$ is defined as:

$$Q_2(\rho) = (2.5 + 3.38\rho + \rho^2) + \frac{3.54\varphi_{\max} d}{\Phi}(2 + \rho + 2\rho^2 + \rho^3) + \frac{4.36\alpha\tau}{\Phi}(1 + 2\rho + 2\rho^2 + \rho^3)$$
$$\text{where } \Phi = \frac{\mu_0(\exp_{x_\perp} \mathring{\mathcal{B}}_{r-2s})}{\omega_d (r-2s)^d}$$

Proof. Out of total allowed error ϵ, we will allocate one half $\epsilon/2$ to the concentration inequality (Proposition 2.8) 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 2.8, we may use k points for local covariance estimation by error level $r^2\epsilon/2(d+2)$:

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

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 $x \in \operatorname{supp}\mu$:

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

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 $i \leq k$ we have

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

as desired. We note lastly that the assumption $2s \leq r$ was dropped because it follows from the other assumptions $r + 2s \leq (\sqrt{2} - 1)\tau$ and $\sqrt{2s/\tau} < r/\tau$. □

We now note that the constant γ can be simplified with a slight relaxation:

Lemma 5.4. *In the setup of Theorem 5.3, suppose that $\epsilon \leq 1$. Then γ satisfies the following:*

$$\frac{1}{\gamma_1} \frac{4608(d+2)^2}{\epsilon^2} - 1 \leq \gamma \leq \frac{1}{\gamma_1} \frac{4609(d+2)^2}{\epsilon^2}$$

Proof. Applying Lemma 6.7, we have:

$$\gamma = \frac{-1}{\log(1 - \gamma_1 \gamma_2)} \in \left[\frac{1}{\gamma_1 \gamma_2} - 1, \frac{1}{\gamma_1 \gamma_2} - \frac{1}{2} \right]$$

Again applying Lemma 6.7, we have:

$$\frac{1}{\gamma_2} = \frac{1}{1 - e^{-\xi}} \in \left[\frac{1}{\xi} + \frac{1}{2}, \frac{1}{\xi} + 1 \right], \text{ where } \xi = \frac{\epsilon^2}{4608(d+2)^2}$$

Therefore,

$$\frac{1}{\gamma_1} \left(\frac{1}{\xi} + 1 \right) - 1 \leq \gamma \leq \frac{1}{\gamma_1} \left(\frac{1}{\xi} + 1 \right) - \frac{1}{2}$$

Since $\epsilon \leq 1$, we have $\xi^{-1} + 1 \leq 4609(d+2)^2/\epsilon^2$. Therefore, we get the claim. □

5.1. Proof of Theorem A. Theorem 5.3 was stated using a projection matrix, although an empirical estimate for the tangent space is given using a set of basis vectors. To directly estimate error of tangent space estimation, we introduce some additional notions.

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 $k \leq D$, define the k-dimensional subspace $\Pi(A, k) \subseteq \mathbb{R}^D$ as the span of the first k columns of U.

Definition 5.6. Suppose $\Pi_1, \Pi_2 \subseteq \mathbb{R}^D$ are two subspaces of $\mathbb{R}^D$ and let $A_1, A_2 \in \mathbb{R}^{D \times d}$ be matrices such that columns of A_i form an orthonormal basis of Π_i. Let $\sigma_1 \geq \cdots \geq \sigma_D$ be the singular values of $A_1^\top A_2$. The *principal angle* between Π_1 and Π_2 is defined as:

$$\measuredangle(\Pi_1, \Pi_2) := \cos^{-1} \sigma_D$$

Or equivalently,[9]

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

with $\measuredangle(x, y) = \cos^{-1}(\langle x, y\rangle/(\|x\| \cdot \|y\|))$.

The following special case of the Davis-Kahan theorem (see [29] or [6]) then allows us to bound the principal angle:

Theorem 5.7 (Davis-Kahan). *Suppose that A is a real symmetric matrix with eigenvalues $\lambda_1^A \geq \lambda_2^A \geq \cdots$. Then for any other real symmetric matrix B and a positive integer k such that $\lambda_k^A \neq \lambda_{k+1}^A$,*

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

We now give the proof of Theorem A, stated in the Introduction.

Proof of Theorem A. The setup is that we take $\epsilon = \sin\theta$ in Theorem 5.3. Then we have the following for each $i \leq k$:

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

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

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

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, $\Pi(\hat{\mathrm{P}}_i, d) = \Pi(\Sigma[\delta_{\mathbf{X}_i|U_i}], d) = \hat{\Pi}_d^{(i)}$, where $U_i = \mathcal{B}_r(X_i)$.

We explain how the Q_2 in Theorem 5.3 is replaced by $c_1' = 3 + (8\varphi_{\max} d + 5\alpha\tau)/\varphi_{\min}$ in Theorem A. (Note that $c_1 = 16(d + 2)c_1'$) This is because

$$\frac{r}{\tau} < \frac{\epsilon}{16(d + 2)c_1'} \text{ implies } \frac{r}{\tau} < \frac{\epsilon}{16(d + 2)Q_2}$$

Firstly, we see that $Q_2'(\rho) \geq Q_2(\rho)$, where $Q_2'(\rho)$ is obtained by replacing Φ by $0.97\varphi_{\min}$ in the definition of $Q_2(\rho)$. This holds since $0.97\varphi_{\min} \leq \Phi$:

$$\Phi = \frac{1}{\omega_d(r - 2s)^d} \int_{\mathring{\mathcal{B}}_{r-2s}} \varphi(\exp_{x_\perp} y) \,\mathrm{J}\exp_{x_\perp} y \,\mathrm{d}\, y \geq \varphi_{\min} \cdot (1 - \frac{r^2}{6\tau^2}) \geq 0.97\varphi_{\min}$$

[9]Note that $\sigma_D = \min_{\|z\|=1} \|A_1^\top A_2 z\| = \min_{\|y\|=1, y\in\Pi_2} \|A_1^\top y\| = \min_{\|y\|=1, y\in\Pi_2} \langle y_1, y\rangle$ where y_1 is the unit vector in the direction of $A_1 A_1^\top y$. Noting that $\langle y_1, y\rangle = \max_{\|x\|=1, x\in\Pi_1} \langle x, y\rangle$, we have $\sigma_D = \min_{\|y\|=1, y\in\Pi_2} \max_{\|x\|=1, x\in\Pi_1} \langle x, y\rangle$.

where we used $r \leq (\sqrt{2}-1)\tau$. Now $Q_2'(\rho)$ is an increasing function in ρ, and by numerical computation we see that $Q_2'(1/48) \leq c_1'$. Thus whenever $\rho < 1/48$,

$$\frac{1}{c_1'} \leq \frac{1}{Q_2'(1/48)} \leq \frac{1}{Q_2'(\rho)} \leq \frac{1}{Q_2(\rho)}$$

Thus if we assume

$$\frac{r}{\tau} < \frac{\epsilon}{16(d+2)c_1'}$$

then this implies $\rho < 1/48$ since $\epsilon = \sin\theta \leq 1$, $c_1' \geq 1$, and $16(d+2) \geq 48$. Therefore, we get the assumption:

$$\frac{r}{\tau} < \frac{\epsilon}{16(d+2)Q_2}$$

We finally apply Lemma 5.4 to simplify γ from Theorem 5.3 to get Theorem A.

The condition $r + 2s \leq (\sqrt{2}-1)\tau$ is dropped from Theorem A because $r/\tau < 1/48$ and $\sqrt{2s/\tau} < r/\tau$ implies $r + 2s \leq (\sqrt{2}-1)\tau$.

5.2. Proof of Theorem B. For dimension estimation, we claim that we must use the following estimation error level to ensure that the dimension estimation works correctly:

$$\epsilon = \frac{\sqrt{(d+1)(d+4)}}{2\sqrt{D}(d+3)}$$

This is basically a consequence of Lemma 6.2. To relate a perturbation of eigenvalues to a perturbation of covariance matrices, we use the Hoffman-Wielandt theorem [11].

Theorem 5.8 (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*

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

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

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

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

The special case for real symmetric matrices follows from Lemma 6.13. We now give the proof of Theorem B, stated in the Introduction.

Proof of Theorem B. Let $\epsilon = \sqrt{(d+1)(d+4)}/2\sqrt{D}(d+3)$. Then $\|\mathrm{P}_i - \hat{\mathrm{P}}_i\| \leq \epsilon$ implies

$$\|\vec{\lambda}[\mathrm{P}_i] - \vec{\lambda}[\hat{\mathrm{P}}_i]\| \leq \|\mathrm{P}_i - \hat{\mathrm{P}}_i\|_{\mathrm{F}} \leq \sqrt{D}\|\mathrm{P}_i - \hat{\mathrm{P}}_i\| \leq \frac{\sqrt{(d+1)(d+4)}}{2(d+3)}$$

where the first inequality is the Hoffman-Wielandt Theorem and the second inequality is a general inequality between the Frobenius and the operator norm of real symmetric matrices. Here, the constant $\sqrt{D}$ is optimal.[10] Recall the definitions $\hat{\mathrm{P}}_i = \frac{d+2}{r^2}\Sigma[\delta_{\mathbf{X}_i|\mathcal{B}_r(X_i)}]$ and $\mathrm{P}_i = \mathrm{P}_{T_{X_i^\perp}M}$. The eigenvalues of the latter are $(1,\dots 1, 0, \dots 0)$. Therefore, dividing by $(d+2)$, the above implies:

$$\left\| \frac{1}{r^2}\vec{\lambda}\Sigma[\delta_{\mathbf{X}|\mathcal{B}_r(X_i)}] - \frac{1}{d+2}(1,\dots 1, 0, \dots 0) \right\| \leq \frac{\sqrt{(d+1)(d+4)}}{2(d+2)(d+3)}$$

where there are d of ones in $1, \dots 1$ above. Then by Lemma 6.2, we have

$$\operatorname{argmin}_k \left\| \frac{1}{r^2}\vec{\lambda}\Sigma[\delta_{\mathbf{X}|\mathcal{B}_r(X_i)}] - \frac{1}{d+2}(1,\dots 1, 0, \dots 0) \right\| = d$$

5.3. Dimension estimation using tail sum: Theorem B′. We prove one more result on dimension estimation, using the tail sum of eigenvalues. Given a tolerance parameter $0 \leq \eta' \leq 1$ and a real symmetric matrix A with eigenvalues $\lambda_1 \geq \dots \geq \lambda_D \geq 0$, we consider the estimator measuring the longest possible tail sum:

$$\hat{d}_{\text{tail}}(A, \eta') := \min \left\{ k \,\middle|\, \sum_{i>k} \lambda_i^2 \leq \eta' \cdot \sum_{i=1}^{D} \lambda_i^2 \right\}$$

For instance, for $(\eta')^2 = 0.05$, we are looking for how many principal components explain 95% of the variance. Observe that the tolerance parameter η' shouldn't be zero in order to ignore a certain degree of noise. This is explained in Proposition 6.14 in Section 6.3. We combine this with Theorem 5.3 as follows:

Theorem B′ (Dimension estimation using tail sum). Let $(\mu, \mu_0) \in \mathcal{P}(M, s)$, where M is a smoothly embedded compact d-dimensional manifold $M \subseteq \mathbb{R}^D$ with reach τ and $s \geq 0$. 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 $\mathbf{X} = (X_1, \dots X_m)$ be an i.i.d. sample drawn from μ. Let η, η' be numbers that satisfy:

$$\eta < \frac{d}{(d+2)(2d+1)}$$
$$\eta' \in \left(\frac{(d+2)^2(2d+1)^2}{4d^3} \cdot \eta, \frac{(d+2)^2(2d+1)^2}{4d(d+1)^2}\left(\frac{1}{d+2} - \eta\right) \right)$$

Given a real $\delta > 0$ and an integer $k > 0$, suppose that a real r, m satisfy the following:

$$\sqrt{\frac{2s}{\tau}} < \frac{r}{\tau} < \frac{(d+2)\eta}{\sqrt{D}c_1}$$
$$m \geq c_2 \cdot \log\left(\frac{(4D+2)k}{\delta}\right) + 1$$

[10] For example, $\|\vec{\lambda}[A] - \vec{\lambda}[A']\| = \sqrt{D}\|A - A'\|$ whenever $A = a \cdot I_D$ and $A' = a \cdot I_D$ for some $a, a' \in \mathbb{R}$.

Then with probability at least $1-\delta$, the following holds for every $i \le k$:

$$\hat{d}_{\text{tail}}\left(\frac{1}{d+2}\hat{\mathrm{P}}_i, \eta'\right) = d$$

Also, c_1, c_2 are defined in Theorem A, with the only difference being that we take

$$\epsilon = \frac{(d+2)\eta}{\sqrt{D}}$$

in the definition of c_2.

Proof. Let $\epsilon = (d+2)\eta/\sqrt{D}$. Then $\|\mathrm{P}_i - \hat{\mathrm{P}}_i\| \le \epsilon$ implies

$$\|\vec{\lambda}[\mathrm{P}_i] - \vec{\lambda}[\hat{\mathrm{P}}_i]\| \le \|\mathrm{P}_i - \hat{\mathrm{P}}_i\|_{\mathrm{F}} \le \sqrt{D}\|\mathrm{P}_i - \hat{\mathrm{P}}_i\| \le (d+2)\eta$$

As before, dividing by $(d+2)$, the above implies:

$$\left\|\frac{1}{r^2}\vec{\lambda}\Sigma[\delta_{\mathbf{X}_i|\mathcal{B}_r(X_i)}] - \frac{1}{d+2}(1,\dots 1,0,\dots 0)\right\| \le \eta$$

where there are d of ones in $1,\dots 1$ above. Then by Proposition 6.14, we have

$$\min\left\{k \,\middle|\, \sum_{i>k}(\lambda_j^{(i)})^2 \le \eta' \cdot \sum_{i=1}^{D}(\lambda_j^{(i)})^2\right\} = d$$

Note that the appearance of c_1, c_2 and the omission of the condition $r+2s \le (\sqrt{2}-1)\tau$ are by the same reason in the proof of Theorem A. □

References

[1] Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic journal of statistics*, 13(1):1359–1399, 2019.

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

[3] Yariv Aizenbud and Barak Sober. Non-parametric estimation of manifolds from noisy data. *arXiv:2105.04754 [math.ST]*, 2021.

[4] Ery Arias-Castro, Gilad Lerman, and Teng Zhang. Spectral clustering based on local PCA. *J. Mach. Learn. Res.*, 18:Paper No. 9, 57, 2017.

[5] Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Intrinsic dimension estimation. *arXiv preprint arXiv:2106.04018*, 2021.

[6] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[7] Herbert Federer. *Geometric measure theory*. Springer, 2014.

[8] Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Fitting a putative manifold to noisy data. In *Conference On Learning Theory*, pages 688–720. PMLR, 2018.

[9] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

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

[11] Alan J Hoffman and Helmut 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.

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

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

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

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

[16] John M Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

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

[18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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

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

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

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

[23] Amit Singer and Hau-Tieng Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.*, 65(8):1067–1144, 2012.

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

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

[26] Hemant Tyagi, Elif Vural, and Pascal Frossard. Tangent space estimation for smooth embeddings of Riemannian manifolds. *Inf. Inference*, 2(1):69–114, 2013.

[27] Sergio Valle, Weihua Li, and S Joe 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.

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

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

[30] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.*, 26(1):313–338, 2004.

6. Appendix

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

- 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.

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
m	Sample size
r	Local detection radius
s	Noise radius
ρ	Normalized local detection radius $\rho = r/\tau$
σ	Normalized noise radius $\sigma = s/\tau$
Ω	$(\rho - 2\sigma)/((\rho + 2\sigma) + (\rho + 2\sigma)^2)$
ϵ	Estimation error level
δ	Estimation within error ϵ happens with probability $\geq 1 - \delta$

6.2. Technical lemmas. We prove technical lemmas not proven in the main body of the paper.

The following is Lemma 13 from [4].

Lemma 6.1. *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:*

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

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 $\mathrm{Unif}(\Pi_{d,D} \cap \mathcal{B}_1(0))$ is:

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

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,

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

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 $d \leq d'$, 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 $k \neq d$, we have:

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

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. $\square$

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 $2AL \leq CL$. $\square$

LEMMA 6.6. *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 6.7. *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:

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

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 6.8. *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 [20], controls the deviation of geodesic from the first order approximation:

Lemma 6.9. *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:

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

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

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

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

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

where the last inequality holds because for any t, $\|\ddot{\gamma}(t)\| \le \tau^{-1}$ (the norm of the second fundamental form is bounded above by τ^{-1}. See Proposition 6.1 of [20]).

To get simpler bounds, now suppose that $r \le (\sqrt{2}-1)\tau$. We note that $x \in [0, \sqrt{2}-1]$ implies[11] $1 - \sqrt{1-2x} \le x + x^2$. Thus

$$\tilde{r} \le \tau - \tau\sqrt{1 - 2\tau^{-1}r} \le r + \frac{r^2}{\tau}$$

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

□

Lemma 6.10. *Let $M \subseteq \mathbb{R}^D$ be a compact smoothly embedded d-dimensional manifold with reach τ. Let $x \in M$ and let $0 \le r \le (\sqrt{2}-1)\tau$ be a radius parameter. Then*

$$\exp_x(\mathring{\mathcal{B}}_r(0)) \subseteq \mathcal{B}_r(x) \cap M \subseteq \exp_x(\mathring{\mathcal{B}}_{r+r^2/\tau}(0))$$

Proof. The first inclusion $\exp_x(\mathring{\mathcal{B}}_r(0)) \subseteq \mathcal{B}_r(x) \cap M$ holds because a straight line is a geodesic in the ambient space $\mathbb{R}^D$. To see the second inclusion, suppose that $\|x - y\| = s \le (\sqrt{2}-1)\tau$. Then Lemma 6.9 tells us that any geodesic connecting (x, y) has length at most $s + s^2/\tau$. Applyig this to every $s \le r$, we get the inclusion. □

Sectional curvature may be used to bound the Jacobian of the exponential map, as follows[17]:

Theorem 6.11. *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:*

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

for all $\|v\|$ if $\kappa_+ \le 0$, and for $\|v\| \le \pi/\sqrt{\kappa_+}$ 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 6.12. *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 $r := \|v\| \le \pi\tau$, we have:*

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

[11] 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.

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

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

Proof. Norm of the second fundamental form is bounded above by τ^{-1} [20], 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 $r \leq \pi\tau$. Then we have:

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

where in the end we used $\sinh x \leq x + \frac{x^3}{4}$ for $x \in [0, 2\sqrt{2}]$[12]. □

Lemma 6.13. *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

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

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

Consider $M = \mathbb{R}$. Suppose that $x_1 \leq \cdots \leq x_n$, $y_1 \leq \cdots \leq y_n$. Then we claim that for any $\sigma \in S_n$,

$$\|x - y\| \leq \|x - \sigma\cdot y\|$$

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:

$$\|x - \tau_{ij}\cdot z\| < \|x - z\|$$

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

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. $\square$

6.3. Dimension estimation with tail sum. We introduce another intrinsic dimension estimator based on tail sum. Given a tolerance parameter $\epsilon \in (0, 1)$ and a real symmetric matrix A with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_D \geq 0$, we consider the estimator measuring the longest possible tail sum:

$$\hat{d}_{\mathrm{tail}}(A, \epsilon) := \min \left\{ k \middle| \sum_{i>k} \lambda_i^2 \leq \epsilon \cdot \sum_{i=1}^{D} \lambda_i^2 \right\}$$

or equivalently, $d = \hat{d}_{\mathrm{tail}}(A, \epsilon)$ is the unique number such that:

$$\sum_{i>d} \lambda_i^2 \leq \epsilon \cdot \sum_{i=1}^{D} \lambda_i^2, \text{ and } \sum_{i>d-1} \lambda_i^2 > \epsilon \cdot \sum_{i=1}^{D} \lambda_i^2 \tag{6.1}$$

Let's derive a sufficient condition for $\hat{d}_{\mathrm{tail}}(A, \epsilon) = d$. Suppose that the following holds:

$$\|\vec{\lambda}[A] - \vec{\lambda}(d, D)\| < \eta$$

or equivalently,

$$\left(\lambda_1 - \frac{1}{d+2}\right)^2 + \cdots + \left(\lambda_d - \frac{1}{d+2}\right)^2 + \lambda_{d+1}^2 + \cdots + \lambda_D^2 < \eta^2 \tag{6.2}$$

By a Lagrange multipler argument, extrema of $\sum_{i=1}^{D} \lambda_i^2$ under the above constraint are found when the gradient of the left hand side of the above is proportional to $(2\lambda_1, \ldots 2\lambda_D)$. If we assume that $\eta < \frac{1}{2(d+2)}$, then the proportionality forces us to have $\lambda_{d+1} = \cdots = \lambda_D = 0$ and also $\lambda_1 = \cdots = \lambda_d$. Therefore,

$$d\left(\frac{1}{d+2} - \frac{\eta}{d}\right)^2 < \sum_{i=1}^{D} \lambda_i^2 < d\left(\frac{1}{d+2} + \frac{\eta}{d}\right)^2 \tag{6.3}$$

Observe that:

$$\sum_{i>d} \lambda_i^2 < \eta^2, \text{ and } \sum_{i>d-1} \lambda_i^2 > \left(\frac{1}{d+2} - \eta\right)^2$$

Now by applying the above and (6.3), a sufficient condition for (6.1) is given by:

$$\eta^2 < \epsilon \cdot d\left(\frac{1}{d+2} - \frac{\eta}{d}\right)^2, \text{ and } \left(\frac{1}{d+2} - \eta\right)^2 > \epsilon \cdot d\left(\frac{1}{d+2} + \frac{\eta}{d}\right)^2$$

or equivalently,

$$\frac{d(d+2)^2}{(d-(d+2)\eta)^2} \cdot \eta^2 < \epsilon < \frac{d(d+2)^2}{(d+(d+2)\eta)^2} \cdot \left(\frac{1}{d+2} - \eta\right)^2$$

If we assume that $\eta < \frac{d}{(d+2)(2d+1)}$, the lower bound for ϵ is indeed smaller than the upper bound (this assumption on η implies the previous assumption $\eta < \frac{1}{2(d+2)}$). Thus, by plugging in $\eta < \frac{d}{(d+2)(2d+1)}$, we get the following sufficient condition for $\hat{d}_{\text{tail}}(A, \epsilon) = d$:

$$\|\vec{\lambda}[A] - \vec{\lambda}(d, D)\| < \eta < \frac{d}{(d+2)(2d+1)}$$

$$\text{and } \frac{(d+2)^2(2d+1)^2}{4d^3} \cdot \eta^2 < \epsilon < \frac{(d+2)^2(2d+1)^2}{4(d+1)^2 d} \cdot \left(\frac{1}{d+2} - \eta\right)^2$$

In summary, we have:

Proposition 6.14. *Let $v = (\lambda_1, \ldots, \lambda_D)$ with $\lambda_1 \geq \cdots \geq \lambda_D \geq 0$. If a number η satisfies:*

$$\|v - \vec{\lambda}(d, D)\| < \eta < \frac{d}{(d+2)(2d+1)}$$

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

and if the tolerance parameter η' satisfies:

$$\eta' \in \left(\frac{(d+2)^2(2d+1)^2}{4d^3} \cdot \eta^2, \frac{(d+2)^2(2d+1)^2}{4d(d+1)^2}\left(\frac{1}{d+2} - \eta\right)^2\right)$$

Then

$$\min\left\{k \middle| \sum_{i>k} \lambda_i^2 \leq \eta' \cdot \sum_{i=1}^{D} \lambda_i^2\right\} = d$$