Tangent Space and Dimension Estimation with the Wasserstein Distance

Uzu Lim, Harald Oberhauser, and Vidit Nanda

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

1. Introduction

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

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

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

$$
\Pi_k[\mathbf{x}] := \text{span}(v_1, \ldots, v_k)
$$

$$
\bar{\Lambda}[\mathbf{x}] := (\lambda_1, \ldots, \lambda_D)
$$

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

$$
\hat{\Pi}(\mathbf{x}, r, i, k) := \Pi_k[\mathbf{x}_i]
$$

$$
\hat{d}(\mathbf{x}, r, i, \eta) := \text{Thr}\left(\bar{\Lambda}[\mathbf{x}_i], \eta\right)
$$

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

When we calculate $\hat{\Pi}$ and $\hat{d}$ for a sample drawn near a $d$-dimensional manifold, we will get accurate estimations of tangent spaces and the intrinsic dimension $d$. Intuitively, this is because when a manifold is zoomed in closely enough at each point, its curvature flattens out and we essentially get a $d$-dimensional disk. Let’s translate this intuition to precise mathematics. To do this, we precisely describe how we draw a random sample near a manifold.
**Setup.** Let $M \subset \mathbb{R}^D$ be a smoothly embedded $d$-dimensional compact manifold. Let $\mu_0$ be a Borel probability measure on $\mathbb{R}^D$ with a probability density function $\varphi : M \to \mathbb{R}_{\geq 0}$: for each open $U \subseteq \mathbb{R}^D$, define

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

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

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

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

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

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

**Main Results.**

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

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

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

$$S_1(\tau, d, \varphi, \theta) = \frac{\sin \theta \varphi_{\text{min}}}{(d + 2)^{3/2} c_1 d \varphi_{\text{max}} + c_2 \alpha \tau}$$

$$S_2(\varrho, D, d, \varphi, \theta) = \frac{c_3 (d + 2)^3}{\omega_d \varphi_{\text{min}} \sin^2 \theta} \log \left(\frac{c_4 D \varrho}{\delta}\right)$$

where $(c_1, c_2, c_3, c_4) = (928, 192, 18574, 14).$
Theorem B (Intrinsic Dimension Estimation). Let $X = \{X_1, \ldots, X_m\}$ be a random sample as above. Given $\eta, \delta, q > 0$ with $\eta < (2D)^{-1}$, the following holds:

$$\sqrt{2\tau s} \leq r \leq S_1 \text{ and } \frac{mr^d}{\log m} \geq S_2 \implies P(\hat{d}_i = d \text{ for } i \leq qm) \geq 1 - \delta$$

where $\hat{d}_i = \hat{d}(X, r, i, \eta)$ is the dimension estimator defined in (1.1).

Remarks. If $\varphi$ vanishes in a small region, we may avoid division by zero by replacing $\varphi_{\min}$ by $\Phi(r_-)$. Here $\Phi$ quantifies local concentration of the measure $\mu_0$. It is defined as $\Phi(r) = \inf_{x \in M} \mu_0(U_{x,r})/\varphi_{\min}(r^d)$ and $U_{x,r} = \{y \in M \mid d(x, \Pi_x(y)) \leq r\}$, where $\Pi_x$ is the projection map to $T_x M$. Also $r_-$ is defined as $r_- = r(1 - r^2/4\tau^2) - 2s$. This stronger result is stated in Theorem 5.3. Also, conditions for $r$ given by two inequalities can be collectively replaced by one upper bound on a function $Q$, defined in Proposition 4.4. Lastly, a special case of our result is given by setting $r = \left(\frac{S_2 \log m}{m}\right)^{1/d}$, which makes our results directly comparable to Theorem 2 of [2]. The constant $S_2$ is fully calculated in our main theorems, improving Theorem 2 of [2].

1.1. Structure of the paper. Theorems A and B follow easily from Theorem 5.3 in Section 5, which is about estimating covariance matrices locally. Ingredients for its proof span Sections 2, 3, 4. In Section 2, we modify the matrix Hoeffding’s inequality to show that Local PCA correctly estimates covariance (Proposition 2.6). In Section 3, we show that given two compactly supported probability measures $\mu, \nu$ valued in $\mathbb{R}^D$, there is a Lipschitz relation of the form $\|\Sigma[\mu] - \Sigma[\nu]\| \leq C \cdot W_1(\mu, \nu)$ where $\Sigma[\mu]$ is the covariance matrix of $\mu$ (Proposition 3.3). In Section 4, we show that if a well-behaved measure on a manifold is restricted to a tiny ball, then its Wasserstein distance to the uniform measure over the unit tangential disk is small (Proposition 4.4). The Lipschitz relation in Section 3 then translates the Wasserstein bound to the bound on matrix norms.

We summarize the notations and conventions of this article in the Appendix (page 28).

1.2. Related works. The task of estimating geometric and topological quantities of manifolds from finitely many sample points lies at the crux of statistical inference, and as such the literature surrounding these topics is vast. Below we have described some of the techniques of which we are aware, and direct the reader to [35, 20, 6] for a more comprehensive survey.
Tangent space estimation. Probabilistic bounds on tangent space estimation using Local PCA have been studied in considerable detail, for example in [2, 31, 16, 28]. To the best of our knowledge, our work is the first in which the tangent space estimation applies to:

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

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

In [16] and [31], the underlying probability measure is assumed to be uniform, and only estimation at a single point is considered. In [28], various constants have not been explicitly computed, and there is no consideration of noise in data distribution. In [2], various constants have not been computed explicitly, thus not specifying the minimum sample size requirement and scaling factor $c$ for their prescription $r = (c \log m/m)^{1/d}$. Furthermore, their noise model is assumed to be orthogonal to the manifold.

Dimension estimation. The idea to use local principal component analysis for estimating intrinsic dimension is ancient, dating back at least to [11]. As such, there is a plethora of literature on the problem of estimating intrinsic dimensions. The work of [22] provides a practical and widely-used maximum likelihood estimator, but there are no known theoretical guarantees of its correctness even for synthetic data. The minimax-based estimator of [17] does come with such guarantees, but in order to compute it one is compelled to solve minimisation problems over the symmetric group on $m$ elements (with $m$ being the total size of the input dataset); thus, this estimator becomes intractable in practice. The recent work
of [5] introduces a far more efficient Wasserstein-based estimator with guarantees\(^1\), but does not adapt to noise. Our efforts in this paper were motivated by the desire to find a suitable balance between practical efficiency, theoretical soundness and compatibility with noise.

**Concentration inequality.** Our concentration inequality for covariance matrices, Proposition 2.4, is directly derived from the matrix Hoeffding inequality in [30]. A more sophisticated approach, such as the one from [18], may be used to improve our concentration inequality. For instance, the constants appearing in Proposition 2.4 may be improved. Similar methods for analyzing (non-local, non-manifold) PCA are also studied in [19, 26].

**Other Techniques.** We also list related techniques that appear in other papers. A cubic bound of the form \(\|\Sigma[\mu] - \Sigma[\nu]\| \leq Cr^3\), where \(\mu, \nu\) 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 (Proposition 3.3 and Corollary 4.5). The key difference in the two derivations is that our approach uses the Wasserstein distance rather than the total variation distance from [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 a similar idea as the proof of Proposition 3.1 in [29]. However, their transportation plan does not involve noise and applies to different types of local covariance matrices.

In [3], local polynomial regression were used to estimate manifolds and their tangent spaces from uniform point samples lying on tubular neighbourhoods. Compared to this work, our results have the advantage of not requiring the noise to be uniformly distributed. Our result only estimates tangent spaces and not higher-order information like curvature. However, the Wasserstein bound could potentially be leveraged to produce bounds on polynomial approximations.

Local PCA has been extensively used in contexts independent of the manifold hypothesis [11, 15, 32, 24], although the theoretical analysis is either heuristic or makes strong assumptions on the underlying distribution (e.g. Gaussian). Theoretical analysis in manifold learning is a flourishing field, with many significant examples including [13, 12, 1, 2, 10, 9, 17, 3, 29] and many others.

**Acknowledgements.**

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

UL is supported by the Korea Foundation for Advanced Studies.

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

\(^1\)We note in passing that the number of points we require to ensure a \(1 - \delta\) probability of correct dimension estimation in our result is \(m \sim \log(1/\delta)\), which improves on the rate \(m \sim \log(1/\delta)^3\) of [5].
HO is supported by the EPSRC grant “Datasig” [EP/S026347/1], The Alan Turing Institute, and the Oxford-Man Institute.

2. Local estimation of covariance matrices

The main result of this section is Proposition 2.6, where we establish bounds for local covariance estimation. Our main tool is the matrix Hoeffding inequality \[30, \text{Theorem 1.3}\]^2. Here onwards, we will use \(\|A\|\) to denote the operator norm of a given matrix \(A\): \(\|A\| := \sup_{\|x\|=1} \|Ax\|\).

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

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

This inequality can be used to establish concentration of vectors.\(^3\)

**Corollary 2.2.** Let \(X_1, \ldots X_m\) be independent random vectors in \(\mathbb{R}^D\) satisfying \(\mathbb{E} X_i = 0\), and \(\|X_i\| \leq \alpha_i\) for some real number \(\alpha_i\). Write \(\sigma^2 = \sum_{i=1}^m \alpha_i^2\). Then for every \(\epsilon \geq 0\),

\[
\mathbb{P}\left( \|X_1 + \cdots + X_m\| \geq \epsilon \right) \leq 2(D + 1) \cdot \exp\left( -\frac{\epsilon^2}{8\sigma^2} \right)
\]

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

**Definition 2.3.** Given \(X \sim \mu\), the covariance matrix of \(\mu\) is the following \(D \times D\) matrix:

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

Let \(\delta_x\) be the Dirac delta measure at a point \(x\). Given \(x = \{x_1, \ldots x_m\} \subset \mathbb{R}^D\), define the empirical measure \(\delta_x\):

\[
\delta_x := \frac{1}{m}(\delta_{x_1} + \cdots + \delta_{x_m})
\]

Given a Borel set \(U \subseteq \mathbb{R}^D\), the normalised restriction of \(\mu\) 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 \(\mu|_U\) constitutes a Borel probability measure on \(\mathbb{R}^D\) whenever \(\mu(U) > 0\).

\(^2\)Our version of the matrix Hoeffding inequality follows from the one in [30] 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 \(\alpha^2 \cdot \text{Id} - A^2\) is positive definite.

\(^3\)Apply Hermitian dilation, which takes a rectangular matrix \(A\) and produces a Hermitian matrix \(A_H = \begin{bmatrix} 0 & A^\top \\ A & 0 \end{bmatrix}\). Then \(\|A_H\|^2 = \|A_H^2\| = \|A\|^2\) and the result applies.
If $X = (X_1, \ldots, X_m)$ is $\mu$-i.i.d. sample, then $\Sigma[\delta X] = \frac{1}{m} \sum_{i=1}^{m} (X_i - \overline{X})(X_i - \overline{X})^\top$, where $\overline{X} = \frac{1}{m} \sum X_i$ is the sample mean. The expected value of $\Sigma[\delta X]$ is in fact $\frac{m-1}{m} \Sigma[\mu]$, but the following computation tells us that we may use it to estimate $\Sigma[\mu]$.

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

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

where, denoting $\overline{X} = \frac{1}{m} \sum X_i$,

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

**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 $\|X - \overline{E}X\| \leq 2$ by the triangle inequality and the constraint on the support of $\mu$. 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 $Y_i = \frac{1}{m}((X_i - \overline{E}X)(X_i - \overline{E}X)^\top - \Sigma)$. Then $\|Y_i\| \leq (4 + 4)/m$ and $\sigma^2 = m \cdot (8/m)^2 = 64/m$. Since $\hat{\Sigma}_0 = \hat{\Sigma} + (\overline{X} - \overline{E}X)(\overline{X} - \overline{E}X)^\top$, we have

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

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

$$
\mathbb{P}(\|\hat{\Sigma} - \Sigma\| \geq t) \leq \mathbb{P}(\|\hat{\Sigma}_0 - \Sigma\| \geq \alpha t) + \mathbb{P}(\|\overline{X} - \overline{E}X\|^2 \geq (1 - \alpha)t)
$$

$$
\leq \mathbb{P}(\|\hat{\Sigma}_0 - \Sigma\| \geq \alpha t) + \mathbb{P}\left(\|\overline{X} - \overline{E}X\| \geq \frac{1}{2}(1 - \alpha)t\right)
$$

$$
\leq 2D \cdot \exp\left(-\frac{\alpha^2 m t^2}{512}\right) + 2(D + 1) \cdot \exp\left(-\frac{(1 - \alpha)^2 m t^2}{128}\right).
$$

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. $\square$

We will estimate $\Sigma[\mu|U]$ with $\Sigma[\delta X|U]$ assuming that $U$ is bounded.

**Proposition 2.5.** Let $X = (X_1, \ldots, X_m)$ be an i.i.d. sample drawn from $\mu$ 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 $\Sigma[\delta X|U]$, and similarly write $\Sigma_U = \Sigma[\mu|U]$. Then for any error level $\epsilon > 0$, we have that $\hat{\Sigma}_U$ estimates $\Sigma_U$:

$$
\mathbb{P}(\|\hat{\Sigma}_U - \Sigma_U\| \leq \epsilon) \geq 1 - \delta,
$$
where \( \delta \) is an expression such that \( \lim_{m \to \infty} \delta = 0 \), defined as:

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

**Proof.** The proof follows from conditioning the membership of elements of \( X \) to \( U \). Denoting by \( S_I \) the event \( (X_i \in U \iff i \in I) \) and writing \( u := \mu(U) \), we have

\[
P(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon) = \sum_{I \subseteq \{1, \ldots, m\}} P(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon | S_I) \cdot P(S_I).
\]

Writing \( |I| \) for the cardinality of each \( I \), we have

\[
P(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon) = \sum_{I \subseteq \{1, \ldots, m\}} u^{|I|}(1 - u)^{m-|I|}P(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon | S_I)
\]

\[
= \sum_{k=0}^{m} \binom{m}{k} u^k (1 - u)^{m-k} P(\|\hat{\Sigma}_U - \Sigma_U\| \geq \epsilon | S_{\{1, \ldots, k\}})
\]

\[
\leq \sum_{k=0}^{m} \binom{m}{k} u^k (1 - u)^{m-k} \cdot (4D + 2) \xi^k
\]

\[
= (4D + 2) \cdot (1 - u(1 - \xi))^m.
\]

Here Proposition 2.4 was applied in the only inequality above. Note that the possibility \( S_\emptyset \) is correctly accounted for since we included \( k = 0 \) when indexing the sum in the second line above. \( \square \)

Now we prove the main result of this section, about estimating \( \Sigma[\mu|U_i] \) for open balls \( U_i \).

**Proposition 2.6.** Let \( \mu \) be a Borel measure supported on a compact subset \( K \subseteq \mathbb{R}^D \), and let \( X = (X_1, \ldots, X_m) \) be a \( \mu \)-i.i.d. sample. Given a radius \( r > 0 \), consider for \( 1 \leq i \leq m \) the covariances \( \hat{\Sigma}_i := \Sigma[\delta_{X_i}|U_i] \) and \( \Sigma_i = \Sigma[\mu|U_i] \), where \( X_i = \{X_j|j \neq i\} \) and \( U_i = B_r(X_i) \). Let \( \epsilon, \delta, \varrho > 0 \) where we assume\(^4\) that \( \epsilon \leq 2r^2 \). Then the following holds:

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

where \( u_0 = \inf_{x \in K} \mu(B_r(x)) > 0 \).

**Proof.** Let \( k = [\epsilon m] \). Define the set \( E_i \subseteq (\mathbb{R}^D)^m \) as:

\[
E_i :=\left\{ x = (x_1, \ldots, x_m) \mid \|\hat{\Sigma}[\delta_{X_i}|U_i] - \Sigma[\mu|U_i]\| > \epsilon\right\}.
\]

\(^4\)We lose nothing from this assumption; suppose \( \mu, \nu \) are two measures supported on a single ball of radius \( r \). Then \( \|\Sigma[\mu] - \Sigma[\nu]\| \leq 2r^2 \) since \( \|\Sigma[\mu] - \Sigma[\nu]\| = \sup_{\|x\|=1} x^T(\mathbb{E}_{X \sim \mu, Y \sim \nu} XX^T - YY^T)x = \sup_{\|x\|=1} (x, x)^2 - (x, x)^2 \leq 2r^2 \leq 2r^2 \).
where \( x_i = \{ x_j | j \neq i \} \). By the union bound, symmetry, and Proposition 2.5, we then have:

\[
\mu(E_1 \cup \cdots \cup E_k) \leq \mu(E_1) + \cdots + \mu(E_k)
\]

\[
=k \cdot \int \mu^{k-1} \left( \left\{ (x_2, \cdots, x_m) | (x_1, x_2, \cdots, x_m) \in E_1 \right\} \right) d\mu(x_1)
\]

\[
\leq k \cdot \int (4D + 2)(1 - u_x(1 - \xi))^m^{-1} d\mu(x)
\]

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

\[
\int (4D + 2)k(1 - u_x(1 - \xi))^{m-1} d\mu(x) \leq (4D + 2)k(1 - u_0(1 - \xi))^{m-1}
\]

(2.1)

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

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

\[
\iff \frac{-1}{\log(1 - u_0(1 - \xi))} \cdot \log \left( \frac{(4D + 2)k}{\delta} \right) \leq m - 1
\]

(2.2)

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

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

where the first inequality is due to Lemma 6.8, and the second inequality follows from the assumption that \( \epsilon^2 \leq 4r^4 \).

Using the fact that \( \log((4D + 2)/\delta) \geq 2 \) and Lemma 6.6, we obtain the claimed sufficient condition for (2.2):

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

To establish that \( u_0 > 0 \), consider the covering of \( K \) by balls of radius \( r/2 \). Since \( K \) is compact, it admits a subcover \( \{ B_{r/2}(x) \mid x \in J \} \), with \( J \) a finite set. Thus, every \( x \in K \) admits a \( y \in J \) satisfying \( x \in B_{r/2}(y) \). Triangle inequality guarantees that \( B_{r/2}(y) \subseteq B_r(x) \), so that \( \mu(B_{r/2}(y)) \leq \mu(B_r(x)) \) and hence \( \inf_{y \in J} \mu(B_{r/2}(y)) \leq \inf_{x \in K} \mu(B_r(x)) \). 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. \( \square \)

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 \([33]\) on its

\(^5\text{By similar reasoning, the left hand side of (2.2) is at least } \frac{1}{u_0} (1150r^4/\epsilon^2), \text{ so that this sufficient condition doesn’t weaken the bound much.}\)
domain, defined as follows. Let \((M, d_M)\) be a Polish metric space equipped with probability measures \(\mu\) and \(\nu\). For each \(p \geq 1\), the \(p\)-Wasserstein distance between \(\mu\) and \(\nu\) equals

\[
W_p(\mu, \nu) := \left( \inf_{\gamma \in \Pi(\mu, \nu)} \int_{M \times M} d_M(x, y)^p \, d\gamma(x, y) \right)^{1/p}
\]

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

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

1. \(\|\mathbb{E}X - \mathbb{E}Y\| \leq W_p(\mu, \nu)\)
2. \(W_p(\tilde{\mu}, \tilde{\nu}) \leq 2 \cdot W_p(\mu, \nu)\)

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

\[
\|x_0 - y_0\| = \left\| \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} (x - y) \, d\mu(x) \, d\nu(y) \right\|
\]

\[
= \left\| \int_{\mathbb{R}^D \times \mathbb{R}^D} (x - y) \, d\gamma(x, y) \right\|, \text{ for any } \gamma \in \Pi(\mu, \nu)
\]

\[
= \inf_{\gamma \in \Pi(\mu, \nu)} \left\| \int_{\mathbb{R}^D \times \mathbb{R}^D} (x - y) \, d\gamma(x, y) \right\|
\]

\[
\leq \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \|x - y\| \, d\gamma(x, y)
\]

\[
= W_1(\mu, \nu)
\]

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

\[
W_p(\tilde{\mu}, \tilde{\nu})^p = \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \|(x - x_0) - (y - y_0)\|^p \, d\gamma(x, y)
\]

\[
= 2^p \cdot \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \left( \frac{\|x - y\| + \|x_0 - y_0\|}{2} \right)^p \, d\gamma(x, y)
\]

\[
\leq 2^p \cdot \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \frac{\|x - y\|^p + \|x_0 - y_0\|^p}{2} \, d\gamma(x, y)
\]

\[
= 2^{p-1} (W_p(\mu, \nu)^p + \|x_0 - y_0\|^p)
\]

\[
\leq 2^p \cdot W_p(\mu, \nu)^p
\]

where the first inequality is the power mean inequality, and the second inequality follows from the first claim. \(\square\)

**Lemma 3.2.** For probability measures \(\mu, \nu\) defined on \(\mathbb{R}\) and supports contained the interval \([-R, +R]\), we have the \(2R\)-Lipschitz relation for all \(p \geq 1\):

\[
\mathbb{E}[X^2] - \mathbb{E}[Y^2] \leq 2R \cdot W_p(\mu, \nu)
\]
Proof. Since $W_p$ is increasing in $p$, it suffices to prove the assertion for $p = 1$.

$$\mathbb{E}[X^2] - \mathbb{E}[Y^2] = \int_{\mathbb{R}} \int_{\mathbb{R}} (x^2 - y^2) \, d\mu(x) \, d\nu(y)$$
$$= \int_{\mathbb{R} \times \mathbb{R}} (x^2 - y^2) \, d\gamma(x, y), \text{ for any } \gamma \in \Pi(\mu, \nu)$$
$$\leq 2R \cdot \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R} \times \mathbb{R}} |x - y| \, d\gamma(x, y)$$
$$= 2R \cdot W_1(\mu, \nu)$$

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

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

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

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 $\mu, \nu$ have zero means. We may also assume that both $\text{supp } \mu$ and $\text{supp } \nu$ are contained within $B_2(0)$ by the triangle inequality; there is a ball $B_1(x)$ of radius 1 containing $\text{supp } \mu$, so that by triangle inequality, $\text{supp } \mu \subseteq B_1(x) \subseteq B_2(0)$.

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

$$\|S\| = \max_i |\lambda_i| = \max_i \left| \mathbb{E}[U^T X X^T U]_{i,i} - \mathbb{E}[U^T Y Y^T U]_{i,i} \right|$$
$$\leq \max_i \left| \mathbb{E}(U^T X)^2 - \mathbb{E}(U^T Y)^2 \right|$$

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

$$\mathbb{E}(U^T X)^2 - \mathbb{E}(U^T Y)^2 \leq 4 W_1((U^T \mu)_i, (U^T \nu)_i)$$
$$\leq 4 W_1(U^T \mu, U^T \nu)$$
$$= 4 W_1(\mu, \nu)$$
where $U^\top \mu = \text{Law}(U^\top X)$ and $(U^\top \mu)_i$ denotes the marginal of $U^\top \mu$ at its $i$th coordinate. The first inequality is Lemma 3.2 with $2R = 4$. The second inequality is a general fact that applies to the Wasserstein distances between marginals. The last equality follows from the fact that the Wasserstein distance is invariant with respect to isometry applied simultaneously to the two measures. Finally, multiplying by the Lipschitz constant $2$ for the non-centered measures, we get the Lipschitz constant $8$. The inequality for other $p$ follows since $W_p$ is increasing in $p$. □

4. Wasserstein bound for Flattening a Measure on Manifold

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

**Definition 4.1.** Given a metric space and a positive integer $d$, denote by $H^d$ the $d$-dimensional Hausdorff measure $[27]$ on the metric space:

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

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

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

**Definition 4.2.** Suppose $M$ is a $d$-dimensional smooth compact manifold with a smooth embedding into $\mathbb{R}^D$ and $\varphi : M \to \mathbb{R}^+$ is a continuous function satisfying $\int_M \varphi \, dH^d = 1$. Let $\mu_0$ be the Borel probability measure given by defining for each open $U \subseteq \mathbb{R}^D$ the following:

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

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

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

Then $\mathcal{P}(M, s)$ is defined as the set of all such pairs $(\mu_0, \mu)$, given $M$ and $s$.

The following are notions from differential geometry relevant to us.

**Definition 4.3.** For each compact Riemannian manifold $M \subset \mathbb{R}^D$, ...
(1) For each \(x, y \in M\), let \(d_M(x, y)\) be the length of the shortest geodesic connecting \(x\) and \(y\). \(^6\)

(2) The reach \(\tau\) of \(M\) is the supremum of \(t \geq 0\) satisfying the following: If \(x \in \mathbb{R}^D\) satisfies \(d_{\mathbb{R}^D}(x, M) \leq t\), then there is a unique point \(x_1 \in M\) such that \(d_{\mathbb{R}^D}(x, x_1) = d_{\mathbb{R}^D}(x, M)\). Here, \(d_{\mathbb{R}^D}(x, y) = \|x - y\|\) is the Euclidean distance on \(\mathbb{R}^D\), and \(d_{\mathbb{R}^D}(x, M) = \inf_{y \in M} d_{\mathbb{R}^D}(x, y)\).

(3) For each point \(x \in M\), we denote by \(\mathcal{B}_r \subseteq T_x M\) the open ball of radius \(r\) around \(0 \in T_x M\), while the notation \(\mathcal{B}_r(x) \subseteq \mathbb{R}^D\) is reserved for the (usual) open ball of radius \(r\) around \(x \in \mathbb{R}^D\).

(4) Given \(x \in M\), the exponential map \(\exp_x\) sends each \(v \in T_x M\) to the endpoint of the unique geodesic on \(M\) starting at \(x\) with the initial velocity of \(v\).

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

**Proposition 4.4.** Let \((\mu_0, \mu) \in \mathcal{P}(M, s)\) where \(M \subseteq \mathbb{R}^D\) is a compact smoothly embedded \(d\)-dimensional manifold with reach \(\tau\) and \(s \geq 0\). Let \(x \in \text{supp} \mu\), let \(x_1 \in M\) be any point in \(\mathcal{B}_s(x) \cap M\), and let \(r\) be a number satisfying the conditions \(2s \leq r \leq (\sqrt{2} - 1)\tau - 2s\) and \(r \leq \tau/(2\sqrt{2}d)\). Then the following holds for any \(p \geq 1\):

\[
W_p(\nu, \tilde{\nu}) \leq \tau \cdot Q\left(\frac{\tau}{\tau^s}, \frac{s}{\tau}\right)
\]

where \(\nu := \mu|_{\mathcal{B}_r(x)}\), and \(\tilde{\nu} := \text{Unif}_d(\mathcal{B}_r(x_1) \cap T_{x_1} M)\)

where \(Q\) is given by:

\[
Q(\rho, \sigma) = 3\sigma + (\rho + 2\sigma)^2 + 2\rho(1 - \Omega^d) \frac{1}{\Phi} \varphi_{\max}\left(1 + 4\sqrt{2}d\rho\right) + \left(\frac{1}{\Phi} (\varphi_{\max} - \varphi_{\min}) (1 + 4\sqrt{2}d\rho) + 4\sqrt{2}d\rho\right) \cdot 2\rho + \frac{1}{4} \rho^3
\]

where \(\varphi_{\max}, \varphi_{\min}\) are extrema of \(\varphi\) taken over \(\mathcal{B}_{r+2s}(x_1)\) and

\[
\Phi = \frac{\mu_0(\Pi^{-1} \mathcal{B}_r)}{\omega_d r^d}, \quad \Omega = \frac{r}{r_+}, \quad r_- = r - \left(1 - \frac{r^2}{4\tau^2}\right) - 2s, \quad r_+ = r + 2s
\]

and \(\Pi\) is the projection map to \(T_{x_1} M\).

**Proof.** We use 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 \(\nu_0\) to get \(\nu_1\)
2. Apply projection to get \(\nu_2\)
3. Fold in the portion of \(\nu_2\) on the outer rim to the inside to get \(\nu_3\)

---

\(^6\)Equivalently, \(d_M(x, y)\) be the infimum of lengths of all piecewise regular curves that connect \(x\) and \(y\).

This follows from the Hopf-Rinow Theorem; see Corollary 6.21 and 6.22 in [21].
Figure 3. An overview of the transportation plan in the proof of Proposition 4.4. The last four sub-diagrams take place on the tangent space. Nonuniform shadings in the 3rd, 4th sub-diagrams indicate nonuniform probability distribution.

(4) Flatten out the nonuniformity and get $\nu_4$.
(5) Rescale radius uniformly to get $\nu_5$.

**Step 1.** Suppose that $X \sim \mu_0$ and $(X+Y) \sim \mu$. We define $\nu_1 := \text{Law}(X \mid X+Y \in B_r(x))$ and define the transportation plan $\nu_{01}$ by $\nu_{01} := \text{Law}((X+Y, X) \mid X+Y \in B_r(x))$, whose marginals are $\nu_0$ and $\nu_1$. Thus for each open $U \subseteq \mathbb{R}^D$, we have

$$\nu_1(U) = \mathbb{P}(X \in U \mid X+Y \in B_r(x)) = \frac{1}{u} \mathbb{P}(X \in U \text{ and } X+Y \in B_r(x))$$

where $u = \mu(B_r(x))$ (4.1)

where $u = \mu(B_r(x)) = \mathbb{P}(X + Y \in B_r(x))$, which follows by the definition of $\mu$. The transportation cost is bounded as

$$W_p(\nu_0, \nu_1) \leq \mathbb{E}_{(X+Y, X) \sim \nu_{01}} \| (X+Y) - X \| \leq s$$

Note that by the assumption $x \in \text{supp} \mu$, we have $u > 0$ and thus we are not conditioning on the null event.

By Equation (4.1), $\nu_1$ is well understood in regions where the condition $X + Y \in B_r(x)$ either always or never holds. If $X \in B_{r-s}(x)$, then since $\|Y\| \leq s$, the triangle inequality implies $X + Y \in B_r(x)$. Similarly if $X \notin B_{r+s}(x)$, then $X + Y \notin B_r(x)$. By also noting that $\|x - x_\bot\| \leq s$, the triangle inequality once again implies $B_{r-2s}(x_\bot) \subseteq B_{r-s}(x)$ and...
Figure 4. Measure $\mu$ and its restriction $\mu|_{\mathcal{B}_r(x)}$, where $x \in \mathbb{R}^D$ and $x_\perp \in M$.

$\mathcal{B}_{r+s}(x) \subseteq \mathcal{B}_{r+2s}(x_\perp)$. Applying Equation (4.1), we get the following:

\begin{align*}
\nu_1(U) &\leq \frac{\mu_0(U)}{u} \quad \text{for any } U \\
\nu_1(U) &= \frac{\mu_0(U)}{u} \quad \text{for } U \subseteq \mathcal{B}_{r-2s}(x_\perp) \\
\nu_1(U) &= 0 \quad \text{for } U \subseteq \mathcal{B}_{r+2s}(x_\perp)^c
\end{align*}

(4.2)

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

**Step 2.** We define $\nu_2$ by pushing forward $\nu_1$ along the projection map to the tangent space, and we must do it where the map is invertible. By Lemma 6.13, we know that the projection map is a diffeomorphism. Now, denote $\Pi$ by the projection map to $T_{x_\perp}M$. Then,

\begin{align*}
\mathcal{B}_-^o &\subseteq \Pi(\mathcal{B}_- \cap M), \quad \Pi(\mathcal{B}_+ \cap M) \subseteq \mathcal{B}_+^o
\end{align*}

(4.3)

where, denoting $\mathcal{B}_r$ by the open ball of radius $r$ in $T_{x_\perp}M$ centered at $0$,

\begin{align*}
\mathcal{B}_- &= \mathcal{B}_{r-2s}(x_\perp), \quad \mathcal{B}_+ = \mathcal{B}_{r+2s}(x_\perp) \\
\mathcal{B}_-^o &= \mathcal{B}_-^o, \quad \mathcal{B}_+^o = \mathcal{B}_+^o \\
r_- &= r \left(1 - \frac{r^2}{4\tau^2}\right) - 2s, \quad r_+ = r + 2s
\end{align*}

The transportation plan is the application of Lemma 6.3 to the pushforward along $\Pi$. 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 6.15, the transportation cost then is bounded as:

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

\footnote{Note that $(r - 2s)(1 - (r - 2s)^2/4\tau^2) \leq (r - 2s)(1 - r^2/4\tau^2) = r(1 - r^2/4\tau^2) - 2s(1 - r^2/4\tau^2) \leq r(1 - r^2/4\tau^2) - 2s$}
Thus by Equations (4.2) and (4.3),
\[ \nu_2(U) \leq \frac{\mu_0(\Pi^{-1}U)}{u} \quad \text{for } U \subseteq \hat{\mathcal{B}}_+ \]
\[ \nu_2(U) = \frac{\mu_0(\Pi^{-1}U)}{u} \quad \text{for } U \subseteq \hat{\mathcal{B}}_- \]
\[ \nu_2(U) = 0 \quad \text{for } U \subseteq (\hat{\mathcal{B}}_+)^c \] (4.4)

Meanwhile, we can evaluate \( \mu_0(U) \) when \( U \subseteq \hat{\mathcal{B}}_+ \) explicitly using the area formula from geometric measure theory\(^8\), which is a generalization of chain rule:
\[ \mu_0(\Pi^{-1}U) = \int_{\Pi^{-1}U} \varphi \, d\mathcal{H}^d = \int_U \varphi(\Pi^{-1}y) J \Pi^{-1}(y) \, dy \] (4.5)

Here, \( Jf \) denotes the Jacobian of a function \( f \) and \( dy \) is the \( d \)-dimensional Lebesgue measure.

Thus,
\[ \nu_2(U) \leq \frac{1}{u} \int_U \varphi(\Pi^{-1}y) J \Pi^{-1}(y) \, dy \quad \text{for } U \subseteq \hat{\mathcal{B}}_+ \]
\[ \nu_2(U) = \frac{1}{u} \int_U \varphi(\Pi^{-1}y) J \Pi^{-1}(y) \, dy \quad \text{for } U \subseteq \hat{\mathcal{B}}_- \]
\[ \nu_2(U) = 0 \quad \text{for } U \subseteq (\hat{\mathcal{B}}_+)^c \] (4.6)

**Step 3.** We saw that \( \nu_2 \) can be written in terms of \( \mu_0 \) inside radius \( r_- \) and vanishes outside radius \( r_+ \). 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 \( \nu_2 \) restricted to \( \hat{\mathcal{B}}_- \). We decompose \( \nu_2 \) as \( \nu_2 = \nu_2^- + \nu_2^+ \), where we define for each Borel set \( U \subseteq T_{x\perp}M \) the following:
\[ \nu_2^-(U) := \nu_2(U \cap \hat{\mathcal{B}}_-) \]
\[ \nu_2^+(U) := \nu_2(U \cap (\hat{\mathcal{B}}_+ - \hat{\mathcal{B}}_-)) \]

Define
\[ \nu_3 := m^{-1} \nu_2^- \]
\[ m := \nu_2^-(T_{x\perp}M) \]

The transportation plan is to: (a) transport \( \nu_2^+ \) to the Dirac delta distribution centered at \( 0 \in T_xM \) and (b) transport this Dirac delta distribution back to \( \frac{1-m}{m} \nu_2^- \). By Lemma 6.4, we have the bound:
\[ W_p(\nu_2, \nu_3) \leq (r_+ + r_-)(1-m) \leq 2r(1-m) \]
since the first part of this transportation moves by distance at most \( r_+ \), the second part moves by at most \( r_- \), and the total mass to move is \( (1-m) \). Equation (4.6) carries over

\(^8\)See for example [8] for a standard reference in geometric measure theory
since $\nu_3$ and $\nu_2^-$ are proportional; for each open $U \subseteq T_{x \perp} M$,

$$\nu_3(U) = \frac{1}{um} \int_{U \cap B^-} \varphi(\Pi^{-1} y) J \Pi^{-1}(y) \, d y$$

(4.7)

**Step 4.** We flatten out the non-uniformity in $\nu_3$. As in Equation (4.7) above, $\nu_3$ is given by the probability density function $\psi(y) := \varphi(\Pi^{-1} y) J \Pi^{-1}(y)$ times a constant. Defining $\nu_4 = \text{Unif}_d(\bar{B}_-)$, we can directly apply Lemma 6.5:

$$W_p(\nu_3, \nu_4) \leq \frac{\omega d r^d}{um} \cdot (\psi_{\text{max}} - \psi_{\text{min}}) \cdot 2 r_-$

where the factor $\omega d r^d$ is needed to rescale the Lebesgue measure $d y$ in Equation (4.7) into $\bar{d}y = d y/(\omega d r^d)$ so that $\int_{\bar{B}_-} \bar{d}y = 1$, so that Lemma 6.5 can be applied. In the above, extrema of $\psi$ are taken over $\bar{B}_-$. Since $\psi$ is the product of $\varphi$ and the Jacobian, the variation $\psi_{\text{max}} - \psi_{\text{min}}$ can be controlled with the triangle inequality as follows:

$$|\psi_{\text{max}} - \psi_{\text{min}}| \leq (\varphi_{\text{max}} - \varphi_{\text{min}}) J_+ + \varphi_{\text{min}} (J_+ - J_-)$$

Here the extrema of $\varphi$ are taken over the geodesic ball $\Pi^{-1} \bar{B}_-$. By Proposition 6.13, we see that:

$$J_- \leq J \Pi^- \leq J_+$$

where $J_- = 1$, $J_+ = \left(1 - \frac{\sqrt{2r}}{\tau}\right)^{-d}$

(4.8)

We furthermore note that, by Equation 4.6,

$$um = \int_{\bar{B}_-} \varphi(\Pi^{-1} y) J \Pi^{-1}(y) \, d y \geq \omega d r^d J_- \varphi_{\text{min}}$$

$$\implies \varphi_{\text{min}} \leq \frac{um}{\omega d r^d} \cdot \frac{1}{J_-}$$

Thus the transportation cost is bounded as:

$$W_p(\nu_3, \nu_4) \leq \left(\frac{\omega d r^d}{um} (\varphi_{\text{max}} - \varphi_{\text{min}}) J_+ + \frac{J_+ - J_-}{J_-}\right) \cdot 2 r_-$$

**Step 5.** Here we simply rescale $\bar{B}_-$ from radius $r_-$ to $r$ radially, which multiplies the associated probability density function by a constant factor (Lemma 6.14), so that we get another uniform distribution. By Lemma 6.3, the transportation cost is bounded by

$$W_p(\nu_4, \nu_5) \leq r - r_- = \frac{r^3}{4 \tau^2} + 2s$$

We note at this point that the extrema of $\varphi$ may be taken over $B_{r+2s}(x_\perp)$ instead, since $B_{r+2s}(x_\perp) \supseteq \Pi^{-1}(\bar{B}_-)$. This relaxation is done for a compatibility with another extrema of $\varphi$ taken later.
The Total Bound. Collecting the bounds\(^\text{10}\), we get:

\[
W_p(\nu_0, \nu_5) \\
\leq W_p(\nu_0, \nu_1) + W_p(\nu_1, \nu_2) + W_p(\nu_2, \nu_3) + W_p(\nu_3, \nu_4) + W_p(\nu_4, \nu_5) \\
\leq s + \frac{(r + 2s)^2}{\tau} + 2r(1 - m) + \\
+ \left(\frac{\omega_d r^d}{um} (\varphi_{\max} - \varphi_{\min}) J_+ + (J_+ - 1)\right) \cdot 2r + \left(\frac{r^3}{4\tau^2} + 2s\right) \\
\tag{4.9}
\]

Using Equations (4.4), (4.6) and (4.8), we obtain the following bounds:

\[
\begin{align*}
\text{um} &= \mu_0(\Pi^{-1}B_-) \leq \varphi_{\max} J_+ \omega_d r^- \\
u(1 - m) &\leq \mu_0(\Pi^{-1}(\bar{B}_+ - B_-)) \leq \varphi_{\max} J_+ \omega_d (r_+^d - r_-^d)
\end{align*}
\]

where \(\varphi_{\max}\) is the maximum of \(\varphi\) taken over \(B_{r+2\delta}(x_\perp)\).\(^\text{11}\) Combining these, we get:

\[
\frac{1 - m}{m} = \frac{u(1 - m)}{ur} \leq \frac{\varphi_{\max} J_+ \omega_d (r_+^d - r_-^d)}{um} = \Phi'(\Omega^{-d} - 1)
\]

with \(\Omega = \frac{r_-}{r_+}, \Phi' = \frac{\varphi_{\max} J_+ \omega_d r^d}{um} \geq 1\)

We can bound \(\nu_2^{\text{out}}\) using the above, as follows:

\[
1 - m = \left(1 + \frac{m}{1 - m}\right)^{-1} \leq \left(1 + \frac{1}{\Phi'(\Omega^{-d} - 1)}\right)^{-1} \leq \Phi'(1 - \Omega^d)
\]

where the first inequality holds by plugging in the upper bound for \((1 - m)/m\), and the second inequality holds since \(\Phi' \geq 1\). Plugging these into Equation (4.9), we get that

\[
W_p(\nu_0, \nu_5) \leq s + \frac{(r + 2s)^2}{\tau} + 2r(1 - \Omega^d)\varphi_{\max} J_+ \omega_d r^- \\
+ \left(\frac{\omega_d r^d}{um} (\varphi_{\max} - \varphi_{\min}) J_+ + (J_+ - 1)\right) \cdot 2r + \left(\frac{r^3}{4\tau^2} + 2s\right)
\]

We bound \(J_+\) using Lemma 6.9. By applying the assumption \(r \leq \tau/(2\sqrt{2}d)\), we see that the lemma applies with \(c = 1/2\):

\[
\left(1 - \frac{\sqrt{2}r}{\tau}\right)^{-d} \leq 1 + \frac{4\sqrt{2}d \cdot r}{\tau}
\]

Applying this to the above bound on \(W_p(\nu_0, \nu_5)\) and also plugging in \(\rho = r/\tau, \sigma = s/\tau\), we obtain the \(Q(\sigma, \tau)\) expression that was claimed in the beginning.  

\(^{10}\)We plug in the definition \(J_{\perp} = 1\), and we also use a slight abuse of notation and identify \(\nu_k\) with \(\iota_\kappa\nu_k\) for \(k = 2, \ldots 5\), where \(\iota : T_{\perp} M \to \mathbb{R}^D\) is the inclusion of tangent space. This is not a problem, since generally \(W_p(\iota_\kappa \mu_1, \iota_\kappa \mu_2) \leq W_p(\mu_1, \mu_2)\) holds for any measures \(\mu_1, \mu_2\) on \(T_{\perp} M\).

\(^{11}\)See Equation 4.3.
Corollary 4.5. In Proposition 4.4, suppose that we additionally assume that there exists $\alpha$ such that the following Lipschitz continuity holds for every $x, y \in M$:
\[
\|\varphi(x) - \varphi(y)\| \leq \alpha \cdot d_M(x, y)
\]
Suppose we also assume $s \leq \frac{r^2}{2\tau}$. Then we have the following quadratic bound:
\[
W(\nu, \tilde{\nu}) \leq Q_2 \cdot \tau \rho^2
\]
where $Q_2$ is defined as:
\[
Q_2 = \left( \frac{7}{2} + 8\sqrt{2}d \right) + \frac{(27/2)d\varphi_{\max} + 6\alpha \tau}{\Phi}
\]

Proof. We use the notation $\rho = r/\tau, \sigma = s/\tau$. Firstly by the assumption $\sigma \leq \rho^2/2$, 
\[
\Omega = \frac{\rho - \rho^3/4 - 2\sigma}{\rho + 2\sigma} \geq \frac{1 - \rho^2/4 - \rho}{1 + \rho} \geq \frac{1 - c\rho}{1 + c\rho}, \text{ where } c = \frac{9}{8}
\]
Then, assuming $\rho \in [0, 8/9]$, Lemma 6.10 says:
\[
1 - \Omega^d \leq 1 - \frac{(1 - c\rho)^d}{(1 + c\rho)^d} \leq 2dc \cdot \rho
\]
By the Lipschitz condition and the noise bound,
\[
Q(\rho, \sigma) = 3\sigma + (\rho + 2\sigma)^2 + 2\rho(1 - \Omega^d) \frac{1}{\Phi} \varphi_{\max} \left(1 + 4\sqrt{2}d\rho\right)
\]
\[
+ \left( \frac{1}{\Phi} (\varphi_{\max} - \varphi_{\min}) (1 + 4\sqrt{2}d\rho) + 4\sqrt{2}d\rho \right) \cdot 2\rho + \frac{1}{4}\rho^3
\]
\[
\leq \frac{3}{2} \rho^2 + (1 + \rho)^2 \rho^2 + \frac{1}{\Phi} \varphi_{\max}^2 4dc \left(1 + 4\sqrt{2}d\rho\right) \rho^2
\]
\[
+ \left( \frac{1}{\Phi} (2(r + 2s)\alpha) (1 + 4\sqrt{2}d\rho) + 4\sqrt{2}d\rho \right) \cdot 2\rho^2 + \frac{1}{4}\rho^3
\]
where the Lipschitz relation is applied to bound $\varphi_{\max} - \varphi_{\min} \leq 2(r + 2s)\alpha$ by using two radial geodesics of length $\leq r_+ = r + 2s$ in the unit ball of radius $r_+$ in the tangent space $T_{x_\perp}M$.
Factoring out $\rho^2$ and plugging back in the definition $c = \frac{9}{8}$, we get:
\[
\frac{1}{\rho^2} Q(\rho, \sigma) \leq \left(8\sqrt{2}d + \frac{5}{2} + \frac{9}{4}\rho + \rho^2\right) + \frac{\varphi_{\max}}{\Phi} \frac{9}{2} d(1 + 4\sqrt{2}d\rho) + \frac{4(\rho + \rho^2)\alpha \tau}{\Phi}(1 + 4\sqrt{2}d\rho)
\]
Using the assumption $\rho \leq \frac{1}{2\sqrt{2}d}$, we get the bounds $1 + 4\sqrt{2}d\rho \leq 3$, and $9\rho/4 + \rho^2 \leq 1$, and $\rho + \rho^2 \leq \frac{1}{2}$. We obtain the claimed bound by plugging them in. \(\square\)
5. Tangent space and dimension estimation

In this section, we combine the Propositions 2.6, 3.3, and 4.4 to prove Theorem 5.3. This in turn implies both Theorem A and B.\(^\text{12}\)

**Definition 5.1.** Given a \(d\)-dimensional subspace \(\Pi \subseteq \mathbb{R}^D\), denote the \(D \times D\) orthogonal projection matrix to \(\Pi\) by \(P_\Pi\), which is a real symmetric matrix, given concretely as:
\[
P_\Pi = A_\Pi A_\Pi^T
\]
where \(A_\Pi \in \mathbb{R}^{D \times d}\) is any matrix whose columns form an orthonormal basis of \(\Pi\).

**Definition 5.2.** Let \(X = (X_1, \ldots X_m)\) be an i.i.d. sample drawn from \(\mu\), a Borel probability measure on \(\mathbb{R}^D\). Given \(x \in \mathbb{R}^D\) and \(r > 0\), define:
\[
\hat{P}_i := \frac{d + 2}{r^2} \Sigma[\delta_X, |U_i|], \text{ where } X_i = \{X_j\}_{j \neq i}, U_i = B_r(X_i)
\]
If \(\Pi \subseteq \mathbb{R}^D\) is a \(d\)-dimensional subspace, then Lemma 6.1 says that:
\[
(d + 2) \Sigma[\text{Unif}(\Pi \cap B_1(0))] = P_\Pi
\]
Thus an approximation to this covariance matrix in Proposition 4.4 amounts to the approximation of a projection matrix, and justifies the definition of \(\hat{P}_i\).

**Theorem 5.3.** Let \((\mu, \mu_0) \in \mathcal{P}(M, s)\(^\text{13}\)) where \(M\) is a smoothly embedded compact \(d\)-dimensional manifold \(M \subseteq \mathbb{R}^D\) with reach \(\tau\) and \(s \geq 0\) is a real number. Let \(\varphi\) be the probability density function of \(\mu_0\) which satisfies \(\|\varphi(x) - \varphi(y)\| \leq \alpha \cdot d_M(x, y)\). Let \(X_1, \ldots X_m\) be an i.i.d. sample drawn from \(\mu\) and let \(X_1, \ldots X_m\) be their orthogonal projections to \(M\). Given \(\delta, \epsilon, \alpha > 0\) and assuming \(^\text{14}\) \(\epsilon < 2\), suppose \(r, m\) satisfy the following:
\[
\sqrt{\frac{2s}{\tau}} \leq \frac{r}{\tau} \leq \frac{\epsilon}{16(d + 2)Q_2} \quad \text{and} \quad \frac{m}{\log m} \geq \frac{4642(d + 2)^2}{u_0 \epsilon^2} \log \left(\frac{14D\rho}{\delta}\right)
\]
where \(u_0 = \inf_{x \in \text{supp } \mu} \mu(B_r(x))\). Then with probability at least \(1 - \delta\), the following holds:
\[
\max_{i \leq \alpha m} \left\|\hat{P}_i - P_i\right\| \leq \epsilon
\]
where \(P_i\) is the projection matrix to the tangent space \(T_{X_i} M\), and \(Q_2\) is defined as:
\[
Q_2 = \left(\frac{7}{2} + 8\sqrt{2d}\right) + \frac{(27/2)d\varphi_{\text{max}} + 6\alpha \tau}{\Phi}, \text{ where } \Phi = \frac{\mu_0(\Pi^{-1}B_{\infty})}{\omega_d r^d}
\]
\(^\text{12}\)Minor technical note: In the special cases discussed in the Introduction, we set \(k = m\) in Theorems A and B, use Lemma 6.6, and use \(\log(14D) > 1 + \log(4D + 2)\) assuming \(D \geq 2\).
\(^\text{13}\)See Definition 4.2.
\(^\text{14}\)Nothing is lost from this assumption since operator norm of the difference of two projection operators is at most 2.
Proof. Out of total allowed error $\epsilon$, we will allocate one half $\epsilon/2$ to the concentration inequality (Proposition 2.6) and the other half $\epsilon/2$ to the curvature (Proposition 4.4). Throughout the proof, we use the shorthand $U_i = B_r(X_i^\perp)$.

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

$$\|\Sigma[\delta X_i|U_i] - \Sigma[\mu|U_i]\| \leq \frac{r^2}{d + 2} \cdot \frac{\epsilon}{2},$$

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

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

Note that $\frac{r^2}{2r^2} P_{X_i^\perp}$ 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

$$\left\| \frac{d + 2}{r^2} \Sigma[\delta X_i|U_i] - P_i \right\| \leq \frac{d + 2}{r^2} \left( \|\Sigma[\delta X_i|U_i] - \Sigma[\mu|U_i]\| + \|\Sigma[\mu|U_i] - \frac{r^2}{d + 2} P_i\| \right)$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

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

5.1. **Proof of Theorem A.** To use Theorem 5.3, we relate the projection matrices to angular deviation between subspaces using the Davis-Kahan theorem (see [34], [7], [36]).

To work with local behaviour of the space given by a union of two manifolds, we must understand the space of pair of subspaces. Indeed, every pair of linear subspaces of the same dimension can be characterised by *principal angles*, up to (simultaneous) rigid motion.

**Definition 5.4.** Given $\pi_1, \pi_2 \in Gr(d, D)$, let $A_i \in \mathbb{R}^{D \times d}$ be a matrix with orthonormal columns that span $\pi_i$. Suppose $\cos \theta_1 \geq \cdots \geq \cos \theta_d$ are singular values of the matrix $A_1^\top A_2$. The *principal angles* of $(\pi_1, \pi_2)$ are defined as the angles $\angle(\pi_1, \pi_2) := (\theta_1, \ldots, \theta_d)$, which satisfy $0 \leq \theta_1 \leq \cdots \leq \theta_d \leq \pi/2$.

By abuse of notation, we will also refer to the largest principal angle $\theta_d$ as the "principal angle". This quantity has a simple interpretation:

---

15Applying Corollary 4.5 requires assuming $\rho + \rho^2 \leq \sqrt{2} - 1$ and $\rho \leq 1/(\sqrt{8}d)$. But this assumption is automatically satisfied by the $r$ in the assumption of the theorem, where we already assume $\rho \leq \epsilon/(16(d + 2)Q_2) \leq 1/(192(d + 2)^2)$. Thus these assumptions become redundant.
Lemma 5.5. If $\langle \pi_1, \pi_2 \rangle = (\theta_1, \ldots, \theta_d)$ for $\pi_1, \pi_2 \in \text{Gr}(d, D)$, then:

$$\theta_d = \max_{x \in \pi_1} \min_{y \in \pi_2} \angle(x, y)$$

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

Proof. Let $A_i \in \mathbb{R}^{D \times d}$ be a matrix whose columns form an orthonormal basis of $\pi_i$. We have:

$$\cos \theta_D = \min_{\|z\| = 1} \|A_i^T A_2 z\| = \min_{\|y\| = 1, y \in \Pi_2} \|A_i^T y\| = \min_{\|y\| = 1, y \in \Pi_2} \langle y_1, y \rangle$$

where $y_1$ is the unit vector in the direction of $A_1 A_i^T y$. Noting that $\langle y_1, y \rangle = \max_{\|x\| = 1, x \in \Pi_1} \langle x, y \rangle$, we have $\cos \theta_D = \min_{\|y\| = 1, y \in \Pi_2} \max_{\|x\| = 1, x \in \Pi_1} \langle x, y \rangle$.

Theorem 5.6 (Davis-Kahan-Wang-Samworth). Let $A, B \in \mathbb{R}^{D \times D}$ be real symmetric matrices. Let $1 \leq d_1 \leq d_2 \leq D$ and assume that $\min(\lambda_{d_1-1, A}, \lambda_{d_2-1, A}) > 0$, where $\lambda_{d_i-1, A} = \lambda_d A - \lambda_{d+1} A$ is the $k$-th spectral gap of the matrix $A$. Let $\pi_A$ be the span of the eigenspaces of eigenvalues $\lambda_{d_1} A, \lambda_{d_1+1} A, \ldots, \lambda_{d_2} A$, and let $\theta_1 \leq \ldots \leq \theta_d$ be the principal angles between $(\pi_A, \pi_B)$. Then we have:

$$\sqrt{\sin^2 \theta_{d_1} + \cdots + \sin^2 \theta_{d_2}} \leq \frac{2}{\min(\lambda_{d_1-1, A}, \lambda_{d_2-1, A})} \cdot \min \left( \|A - B\|_F, \sqrt{d} \|A - B\| \right)$$

In particular, for $(d_1, d_2) = (1, d)$, we have:

$$\sqrt{\sin^2 \theta_1 + \cdots + \sin^2 \theta_d} \leq \frac{2}{\lambda_{d-1}} \cdot \min \left( \|A - B\|_F, \sqrt{d} \|A - B\| \right)$$

We will only be using the case of $(d_1, d_2) = (1, d)$ above.

Proof of Theorem A. This is a direct corollary of plugging in $\epsilon = (\sin \theta)/(2\sqrt{d+2})$ in Theorem 5.3. Assuming that, the following holds for each $i \leq \lfloor \varphi m \rfloor$:

$$\|P_i - \hat{P}_i\| \leq \frac{\sin \theta}{2\sqrt{d+2}}$$

Since both $P_i$ and $\hat{P}_i$ are real symmetric matrices and since eigenvalues of $P_i$ are $(1, \ldots, 1, 0, \ldots, 0)$, its $d$-th spectral gap is 1 and therefore letting $A = P_i, B = \hat{P}_i$ in the Davis-Kahan theorem gives the following:\footnote{In the equation, note that we could choose $\epsilon = \epsilon/(2\sqrt{d})$ for a slightly tighter bound. Our choice of $\epsilon$ is for cleanliness of the final expression produced.}

$$\sin \angle \left( \Pi(P_i, d), \Pi(\hat{P}_i, d) \right) \leq 2\sqrt{d} \|P_i - \hat{P}_i\| \leq \frac{2\sqrt{d}}{2\sqrt{d+2}} \sin \theta \leq \sin \theta$$

Since $P_i$ is the projection matrix to $T_{X_i^+} M$, a $d$-dimensional space, we have $\Pi(P_i, d) = T_{X_i^+} M$. Furthermore, $\Pi(\hat{P}_i, d) = \Pi(\Sigma[\delta X_i, U_i], d) = \hat{P}_i$, where $U_i = B_i(X_i)$.

In Theorem A, the conditions for $(r, m)$ used in Theorem 5.3 are made stricter for the sake of easy interpretability. We explain how this is done.
Condition on \( r \). The following is the required upper bound on \( \rho = r/\tau \):

\[
\rho \leq \frac{\epsilon}{16(d+2)Q_2}, \text{ where } \epsilon = \frac{\sin \theta}{2\sqrt{d+2}}
\]

Using \( \Phi \geq \varphi_{\min} \) (follows from Equation (4.5) and the Jacobian of inverse-projection being \( \geq 1 \)), we get the following upper bound on \( Q_2 \):

\[
Q_2 = \left( \frac{7}{2} + 8\sqrt{2}d \right) + \frac{1}{\Phi} \left( \frac{27d}{2} \varphi_{\max} + 6\alpha \tau \right)
\]

\[
\leq \left( \frac{7}{2} + 8\sqrt{2}d \right) + \frac{1}{\varphi_{\min}} \left( \frac{27d}{2} \varphi_{\max} + 6\alpha \tau \right)
\]

\[
\leq \left( \frac{7}{2} + 8\sqrt{2} + \frac{27}{2} \right) d \cdot \frac{\varphi_{\max}}{\varphi_{\min}} + \frac{6\alpha \tau}{\varphi_{\min}}
\]

\[
\leq \frac{20d\varphi_{\max} + 6\alpha \tau}{\varphi_{\min}}
\]

Thus we get the required upper bound for \( \rho = r/\tau \) used in Theorem A, as follows:

\[
\frac{\epsilon}{16(d+2)Q_2} = \frac{\sin \theta}{32(d+2)^{3/2}} \cdot \frac{\varphi_{\min}}{29d\varphi_{\max} + 6\alpha \tau} = \frac{\sin \theta}{(d+2)^{3/2} c_1 d \varphi_{\max} + c_2 \alpha \tau}
\]

where \((c_1, c_2) = (928, 192)\).

Condition on \( m \). The required lower bound for \( m/\log m \) is obtained by also plugging in \( \epsilon = \sin \theta/(2\sqrt{d+2}) \) in Theorem 5.3, and noting that \( u_0 \geq \omega d^2 \varphi_{\min} \), by Equations (4.5) and (4.3). Furthermore, we use the following:

\[
r_0 = r \left( 1 - \frac{r^2}{4\tau^2} \right) - 2s \geq r \cdot \left( 1 - \frac{r}{\tau} - \frac{r^2}{4\tau^2} \right)
\]

Assuming that \( \rho = r/\tau \) satisfies \( \rho + \rho^2/4 \leq c/d \) for some constant \( c > 0 \) and applying Lemma 6.9 with \( t = \rho + \rho^2/4 \leq c/d \), we get:

\[
\frac{1}{r^d} \leq \frac{1}{r^d} \left( 1 - t \right)^{-d} \leq \frac{1}{r^d} \left( 1 + \frac{d}{(1-c)^2} t \right) \leq \frac{1}{r^d} \left( 1 + \frac{c}{(1-c)^2} \right)
\]

By assuming the condition on \( r \) derived above, we have that \( \rho \leq 1/(3^{3/2}\cdot 928) \leq 4820 \), so that we can take \( c = 0.00025 \), which implies \( c/(1-c)^2 \leq 0.0003 \). This yields \( 1.0003 \times (4642 \times 4) \leq 18574 = c_3 \).

5.2. Proof of Theorem B. To relate a perturbation of eigenvalues to a perturbation of covariance matrices, we use the Hoffman-Wielandt theorem [14].

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

\[
\sum_{i=1}^D |\lambda_i - \lambda'_i|^2 \leq ||A - A'||_F^2
\]

24
where \( \|A\|_F := \sqrt{\text{Tr}(A^T A)} \) denotes the Frobenius norm, with \( \text{Tr}(\bullet) \) denoting the trace. In particular, if \( A, A' \) are real symmetric matrices, then \(^{17}\)

\[
\|\vec{\lambda}[A] - \vec{\lambda}[A']\| \leq \|A - A'\|_F
\]

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

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

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

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

where \( \text{Thr} \) is defined in the Introduction.

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

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

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

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

Since \( \|\vec{\lambda}(d, D)\|_1 = d/(d + 2) \), triangle inequality implies that \( \frac{d}{d+2} - q \leq \|\vec{\lambda}\|_1 \leq \frac{d}{d+2} + q \).

Thus we can formulate the following sufficient conditions:

\[
q < \eta \left( \frac{d}{d+2} - q \right), \text{ and } \frac{1}{d+2} - q > \eta \left( \frac{d}{d+2} + q \right)
\]

\[
\iff (1 + \eta)q < \frac{\eta d}{d+2}, \text{ and } (1 + \eta)q < \frac{1 - \eta d}{d+2}
\]

\[
\iff q < \frac{\min(\eta d, 1 - \eta d)}{(1 + \eta)(d + 2)}
\]

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

**Proof of Theorem B.**

The proof goes verbatim except we use the Hoffman-Wielandt theorem instead of the Davis-Kahan theorem, and that we use the estimation error for the covariances \( \|\hat{\Sigma} - \Sigma\|_2 \),

\(^{17}\)This special case for real symmetric matrices follows from Lemma 6.18.
given by $\epsilon^{-1} = 3D(1 + \eta^{-1})$. Then the following chain of inequalities hold with probability $\geq 1 - \delta$:

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

The proof is then completed by applying Proposition 5.8. We note how the expression $Q_2$ is weakened by using Equation (5.1), which is also used in deriving Theorem A:

$$
\frac{\epsilon}{16(d + 2)Q_2} \geq \frac{1}{48(d + 2)D(1 + \eta^{-1})} \frac{\varphi_{\min}}{29d\varphi_{\max} + 6\alpha\tau} = \frac{1}{(d + 2)D(1 + \eta^{-1})} \frac{\varphi_{\min}}{c_1d\varphi_{\max} + c_2\alpha\tau}
$$

where $(c_1, c_2) = (1392, 288)$. The condition on $m$ is derived in a similar manner described in the proof of Theorem A. This time, we get $1.0003 \times (4642 \times 9) \leq 41791 = c_3$.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[16] D. N. Kaslovsky and F. G. Meyer. Non-asymptotic analysis of tangent space perturbation. *Inf. Inference*, 3(2):134–187, 2014.
[17] J. Kim, A. Rinaldo, and L. Wasserman. Minimax rates for estimating the dimension of a manifold. *arXiv preprint arXiv:1605.01011*, 2016.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[36] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.
6. Appendix

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

- The word ‘dimension’ and ‘intrinsic dimension’ are used interchangeably, where ‘intrinsic’ simply distinguishes it from the ‘ambient’ dimension.
- All manifolds are connected.
- All vectors are by default column vectors.
- \(|v| = \sqrt{v^\top v}\) denotes the Euclidean norm of a vector \(v \in \mathbb{R}^D\).
- \(|A|\) denotes the operator norm of a matrix \(A \in \mathbb{R}^{m \times n}\), seen as a map \(\mathbb{R}^n \to \mathbb{R}^m\).
  \(|A|_F = \sqrt{\text{Tr}(A^\top A)}\) 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 \(\mu\) over \(\mathbb{R}^D\).
- \(B_r(x) \subseteq \mathbb{R}^D\) 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 \(x \in M\), \(\tilde{B}_r \subseteq T_xM\) denotes the open ball of radius \(r\) centered at \(0 \in T_xM\), assuming that the choice of \(x\) is clear from the context.
- \(\lambda[A] \in \mathbb{R}^D\) denotes the eigenvalues of a real symmetric matrix \(A\) of size \(D \times D\), arranged in the decreasing order.
- \(\omega_d = \pi^{d/2}/\Gamma(d/2 + 1)\) denotes the volume of the \(d\)-dimensional unit ball.

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

| Notation | Meaning |
|----------|---------|
| \(M\)    | A compact manifold smoothly embedded in \(\mathbb{R}^D\) |
| \(d\)    | (Intrinsic) dimension of \(M\) |
| \(D\)    | Ambient dimension |
| \(\tau\) | Reach of \(M\) |
| \(\mu\)  | Main distribution of interest with noise |
| \(\mu_0\) | \(\mu\) before adding noise |
| \(\varphi\) | Probability density function on \(M\) used to define \(\mu_0\) |
| \(\alpha\) | Lipschitz constant for \(\varphi\) |
| \(m\)    | Sample size |
| \(r\)    | Local detection radius |
| \(s\)    | Noise radius |
| \(g\)    | Probabilistic guarantees hold for \(\lfloor gm \rfloor\) out of \(m\) points |
| \(\delta\) | Probabilistic guarantees hold with probability \(\geq 1 - \delta\) |
| \(\rho\) | Normalized local detection radius \(\rho = r/\tau\) |
| \(\sigma\) | Normalized noise radius \(\sigma = s/\tau\) |
6.2. Technical lemmas.

**Lemma 6.1.** (Lemma 13 from [4]) Given a d-dimensional subspace \( \Pi \) of \( \mathbb{R}^D \), the covariance matrix of the uniform distribution over the disk \( \Pi \cap B_1(0) \) is given by:

\[
\Sigma[\text{Unif}_d(\Pi \cap B_1(0))] = \frac{1}{d + 2} \Pi
\]

where \( \Pi \) is the \( D \times D \) projection matrix to \( \Pi \). Eigenvalues of this matrix are:

\[
\frac{1}{d + 2} \left( \frac{1}{d}, \ldots, \frac{1}{d}, 0, \ldots, 0 \right)
\]

**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 \( \text{Unif}_d(\Pi_{d,D} \cap B_1(0)) \) is:

\[
\frac{1}{\omega_d} \int_{\|x\| \leq 1} x^2 \, dx = \frac{1}{\omega_d} \int_{\|x\| \leq 1} \|x\|^2 \, dx = \frac{1}{d} \int_0^1 r^2 \cdot dr^{d-1} \, dr = \int_0^1 r^{d+1} \, dr = \frac{1}{d + 2}
\]

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,

\[
\Sigma[\text{Unif}_d(\Pi_{d,D} \cap B_1(0))] = \frac{1}{d + 2} \left[ I_d \ 0 \right] = \frac{1}{d + 2} \left[ I_d \ 0 \right] = \frac{1}{d + 2} \Pi
\]

Given any \( d \)-dimensional subspace \( \Pi \subseteq \mathbb{R}^D \), consider an orthonormal basis \( A = [v_1, \ldots, v_D] \) such that the first \( d \) vectors \( [v_1, \ldots, v_d] \) span \( \Pi \). If \( X \sim \text{Unif}(\Pi \cap B_1(0)) \), then \( A^{-1}X \sim \text{Unif}(\Pi_{d,D} \cap B_1(0)) \). Thus the covariance matrix of \( X \) is

\[
\frac{1}{d + 2} A \left[ I_d \ 0 \right] A^\top = \frac{1}{d + 2} [v_1, \ldots, v_d][v_1, \ldots, v_d]^\top = \frac{1}{d + 2} \Pi
\]

\( \square \)

**Lemma 6.2.** Suppose

\[
\lambda(d, D) := \frac{1}{d + 2} \left( \frac{1}{d}, \ldots, \frac{1}{d}, 0, \ldots, 0 \right)
\]

If \( d \leq d' \), then

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

Also for any \( k \neq d \), we have:

\[
\|\lambda(k, D) - \lambda(d, D)\| \geq \|\lambda(d, D) - \lambda(d + 1, D)\| = \frac{\sqrt{(d + 1)(d + 4)}}{(d + 2)(d + 3)}
\]
Proof. The norm of difference is given by direct computation:
\[
\| \tilde{\lambda}(d, D) - \tilde{\lambda}(d', D) \|^2 = d \cdot \left( \frac{1}{d + 2} - \frac{1}{d' + 2} \right)^2 + \frac{d' - d}{(d' + 2)^2} = \frac{(d' - d)(dd' + 4d + 4)}{(d + 2)(d' + 2)^2}
\]

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 \(d \geq 2\),
\[
\min_{d \neq d'} \| \tilde{\lambda}(d, D) - \tilde{\lambda}(d', D) \| = \min(\| \tilde{\lambda}(d, D) - \tilde{\lambda}(d + 1, D) \|, \| \tilde{\lambda}(d, D) - \tilde{\lambda}(d - 1, D) \|)
\]
\[
= \min \left( \frac{\sqrt{(d + 1)(d + 4)}}{(d + 2)(d + 3)}, \frac{\sqrt{d(d + 3)}}{(d + 1)(d + 2)} \right)
\]
\[
= \frac{\sqrt{(d + 1)(d + 4)}}{(d + 2)(d + 3)}
\]
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). \(\square\)

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 \(\iota^A : A \to M, \iota^B : B \to M\). 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 \(\mu\) be a Borel probability measure on \(A\). Then for any \(p \geq 1\), the Wasserstein distance between pushforwards of \(\mu\) and \(f_*\mu\) along inclusions are bounded by \(L\):
\[
W_p(\iota^A_*\mu, \iota^B_*f_*\mu) \leq L
\]

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:
\[
W_p(\iota^A_*\mu, \iota^B_*f_*\mu) \leq (\mathbb{E}_X d_M(X, f(X))^p)^{1/p} \leq L
\]
\(\square\)

Lemma 6.4. Let \(M\) be a Polish metric space with metric \(d_M\) and a finite diameter \(L := \sup_{x,y \in M} d_M(x, y)\). For a Borel probability measure \(\mu\) on \(M\) and a Dirac delta measure \(\delta_x\) centered at \(x \in M\), we have:
\[
W_p(\mu, \delta_x) \leq L
\]

Proof. Define the transportation plan \(\nu\) 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 \(\mu\) and \(\delta_x\). 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 $L := \sup_{x,y \in M} d_M(x, y)$. Fix a Borel probability measure $\mu$ 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 $\mu_f$ be the Borel probability measure on $M$ given by taking $f$ as the probability density function. Then for any $p \geq 1$,

$$W_p(\mu_f, \mu) \leq CL$$

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:

$$\mu_f = \mu + \mu_f^+ - \mu_f^-$$

where $\mu_f^+(U) = \int_U \max(0, f(x) - 1) \, d\mu(x)$

$$\mu_f^-(U) = \int_U \max(0, 1 - f(x)) \, d\mu(x)$$

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

$$W_p(\mu_f, \mu) = W_p(\mu + \mu_f^+ - \mu_f^-, \mu) \leq W_p(\mu_f^+ \mu_f^-)$$

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

$$W_p(\mu_f^+, \mu_f^-) \leq W_p(\mu_f^+, A \cdot \delta_x) + W_p(A \cdot \delta_x, \mu_f^-) \leq 2AL$$

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

$$A = \mu_f^+(M) \leq \sup_{x \in M} f(x) - 1$$

$$A = \mu_f^-(M) \leq 1 - \inf_{x \in M} f(x)$$

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

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

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

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

$$\frac{x}{y} > a(1 + c) \implies x > a(y + c) \implies \frac{x}{y} > a$$

\hfill \square

Lemma 6.7. For the following function

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

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

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

31
Proof. Let’s always assume $x \in [0, 1/a]$ here. By direct evaluation, $f'(0) = -(1 + 2a)$ and thus the claim is equivalent to $f(x)^k \geq 1 + kf'(0)x$. Since $f(0) = 1$, it’s sufficient to show that $(f^k)'(x) \geq kf'(0)$ for any $x$. We have $f' < 0$ since $f$ is decreasing, and we can also directly check that $0 \leq f \leq 1$. Thus $(f^k)' = kf^{k-1}f' \geq kf'$. Thus it suffices to show that $f' \geq f'(0)$. By direct computation, we have:

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

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

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

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

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

$$\frac{1}{1 - e^{-1/t}} - t \in \left[\frac{1}{2}, 1\right]$$

$$\frac{1}{\log(1 - s^{-1})} + s \in \left[\frac{1}{2}, 1\right]$$

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:

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

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:

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

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 6.9.** Suppose $0 < c \leq 1, d \geq 1$ and $t \leq c/d$. Then we have the following linear bound:

$$(1 - t)^{-d} \leq 1 + \frac{d}{(1-c)^2} \cdot t$$

**Proof.** Let $f_d(t)$. The first and second derivatives are:

$$f'_d(t) = d(1-t)^{-d-1}, \quad f''_d(t) = d(d+1)(1-t)^{-d-2}$$

and thus $f'_d(t)$ is an increasing function at $t \in [0,1]$. This implies that, for each $0 \leq t \leq t_0 \leq 1$, we have:

$$f_d(t) \leq 1 + f'_d(t_0)t$$

Take $t_0 = c/d$. Then:

$$f'_d(c/d) = \frac{d}{1-c/d} \cdot \frac{1}{(1-c/d)^d} \leq \frac{d}{(1-c)^2}$$

where we used the fact that $s \mapsto (1 - 1/s)^d$ is an increasing function for $s \geq 0$ to see that $(1 - c/d)^d \geq (1 - c)$. \hfill \Box

**Lemma 6.10.** Suppose $d \geq 1, t \in [0,1]$. Then

$$\left(\frac{1-t}{1+t}\right)^d \geq 1 - 2d \cdot t$$

**Proof.** The first and second derivative of the function $f_d(t) = ((1-t)/(1+t))^d$ are:

$$f'_d(t) = \frac{2d}{t^2-1} \left(\frac{1-t}{1+t}\right)^d, \quad f''_d(t) = \frac{4d(d-t)}{(t^2-1)^2} \left(\frac{1-t}{1+t}\right)^d$$

For $t \in [0,1]$, the second derivative is $\geq 0$. Therefore we have $f_d(t) \geq 1 + f'_d(0)t$. Since $f'_d(0) = -2d$, we get the claim. \hfill \Box

**Lemma 6.11.** Let $M \subset \mathbb{R}^D$ be a compact set and let $\tau$ be its reach. Let $\pi_M$ be the projection map to $M$, such that for any $x \in \mathbb{R}^D$, $\pi_M(x)$ is the set of points on $M$ that minimises the distance to $M$. The following hold:

1. The distance function $x \mapsto d(x, M) = \inf\{\|y - x\| \mid y \in M\}$ is continuous.
2. For $0 < r < \tau$, $\pi_M|_{B(M,r)}$ is a single-valued continuous function.

**Proof.** (1) From the definition it easily follows that $d(-, M)$ is a Lipschitz function; we have that: $\|d(x, M) - d(x', M)\| \leq \|x - x'\|$.

(2) Let’s write $\pi = \pi_M|_{B(M,r)}$ for the moment. Let $x \in B(M, r)$. Suppose that $x_n \to x$ but $\pi(x_n)$ doesn’t converge to $\pi(x)$. Then there exists $s > 0$ such that $\pi(x_n) \notin B(\pi(x), s)$.

Since $d(y, M) = \|y - \pi(y)\|$ for each $y \in B(M, r)$, the continuity of $d(-, M)$ implies that there is a convergence $\|x_n - \pi(x_n)\| \to \|x - \pi(x)\|$. Since we also have $x_n \to x$, we have $\|x - \pi(x_n)\| \to \|x - \pi(x)\|$. Thus inf $\{\|x - y\| \mid y \in M \setminus B(\pi(x), s)\} = \|x - \pi(x)\| = d(x, M)$.

This is a contradiction. Since $M \setminus B(\pi(x), s)$ is a compact set, the distance function $y \mapsto \|y - x\|$ attains a minimum on some $z \in M \setminus B(\pi(x), s)$. This violates the definition of
reach, which requires a unique nearest point of \( x \) on \( M \), which can’t be simultaneously \( \pi(x) \) and \( z \).

**Lemma 6.12.** Let \( M \subset \mathbb{R}^D \) be a compact path-connected set and let \( \tau \) be its reach. If \( x, y \in M \) satisfies \( \| x - y \| < \tau \), then there exists a continuous path on \( M \) that connects \( (x, y) \) such that every point on the path has distance at most \( \| x - y \| \) from both \( x \) and \( y \).

**Proof.** Define a path \( \bar{\gamma} : [0, 1] \to M \) by \( \bar{\gamma}(t) = (1 - t)x + ty \), the line segment connecting \( (x, y) \). Since \( \| x - y \| < \tau \), every point on \( \bar{\gamma} \) is within distance \( \tau \) from \( x \), and thus \( \pi_M \circ \bar{\gamma} : [0, 1] \to M \) is a (single-valued) continuous function. Let’s write \( \gamma = \pi_M \circ \bar{\gamma} \).

Let \( t_0 \in [0, 1] \) and write \( z = \gamma(t_0) \) and \( \bar{z} = \bar{\gamma}(t_0) \). Then we have:

\[
\| z - x \| \leq \| z - \bar{z} \| + \| \bar{z} - x \| \leq \| y - \bar{z} \| + \| \bar{z} - x \| = \| y - x \|
\]

where the first inequality is the triangle inequality, the second inequality is due to the definition of \( \gamma \), and the last equality is due to \( (x, \bar{z}, y) \) lying on one line. Therefore \( \| z - x \| \leq \| y - x \| \), and by symmetry of the argument in \( (x, y) \), we also get \( \| z - y \| \leq \| y - x \| \). \( \square \)

**Proposition 6.13.** Let \( M \) be a \( d \)-dimensional submanifold. Let \( \pi_x : \mathbb{R}^D \to T_x M \) be the projection map to \( T_x M \), and let \( \tilde{\pi}_x := \pi_x|_M : M \to T_x M \) and \( \tilde{\pi}_{x,r} := \pi_x|_{M \cap B(x,r)} \). The following hold:

1. When \( r < \tau/2 \), \( \tilde{\pi}_{x,r} \) has nonsingular derivatives and is a diffeomorphism.
2. For any \( y \in M \), we have \( J_y \tilde{\pi}_x = \text{det}(A_x^T A_y) \), where \( A_x \in \mathbb{R}^{D \times d} \) is any orthonormal frame of \( T_x M \).
3. For any \( y \in M \), the following bound holds:

\[
\cos \theta_{x,y} \geq 1 - \frac{d_M(x, y)}{\tau}
\]

where \( \theta_{x,y} = \angle_{\max}(T_x M, T_y M) \).
4. For any \( y \in M \), the following bound holds:

\[
J_y \tilde{\pi}_x \in \left[ (\cos \theta_{x,y})^d, 1 \right]
\]

If \( r < (\sqrt{2} - 1)\tau \), then we furthermore get:

\[
J_y \tilde{\pi}_x \in \left[ (1 - \sqrt{2}r/\tau)^d, 1 \right]
\]

**Proof.** (1) The nonsingularity is Lemma 5.4 from [?]. By applying the inverse function theorem locally at each point where the derivative is non-singular, we see that \( \tilde{\pi}_{x,r} \) is a diffeomorphism.

(2) This is because \( d \tilde{\pi}_x(v) = A_x^T v \) for each (embedded) tangent vector \( v \in T_y M \).

(3) This is Proposition 6.2 from [?].

(4) The first bound follows from (2) and the definition of principal angles. The second bound follows from (3) and Lemma 6.15, which implies \( d_M(x, y)/\tau \leq (r/\tau) + (r/\tau)^2 \leq \sqrt{2}r/\tau \). \( \square \)
**Lemma 6.14.** 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 \( f(x) = f_0(x) \) for \( x \neq 0 \). Then the Jacobian determinant of \( f \) is given by:

\[
J_f(x) = f_0(x)
\]

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

\[
J_f = \text{det}(f_0 I + (\nabla g)x^\top) = f_0 + (\nabla f_0)^\top x = f_0
\]

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

The following lemma, which is a simple extension of Proposition 6.3 of [25], controls the deviation of geodesic from the first order approximation:

**Lemma 6.15.** Let \( M \) be a smooth compact \( n \)-manifold embedded in \( \mathbb{R}^D \) with reach \( \tau \). 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,

\[
r = \|\gamma(\tilde{r}) - \gamma(0)\| \geq \|\tilde{r}\dot{\gamma}(0)\| - \left\| \int_0^{\tilde{r}} \int_0^{t_1} \ddot{\gamma}(t_2) \, dt_2 \, dt_1 \right\| \geq \tilde{r} - \frac{\tilde{r}^2}{2\tau}
\]

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:

\[
\gamma(\tilde{r}) - \gamma(0) = \int_0^{\tilde{r}} \dot{\gamma}(t_1) \, dt_1 = \int_0^{\tilde{r}} \left( \dot{\gamma}(0) + \int_0^{t_1} \ddot{\gamma}(t_2) \, dt_2 \right) \, dt_1 = \tilde{r}\dot{\gamma}(0) + \int_0^{\tilde{r}} \int_0^{t_1} \ddot{\gamma}(t_2) \, dt_2 \, dt_1
\]

and thus

\[
\|\gamma(\tilde{r}) - (\gamma(0) + \tilde{r}\dot{\gamma}(0))\| = \left\| \int_0^{\tilde{r}} \int_0^{t_1} \ddot{\gamma}(t_2) \, dt_2 \, dt_1 \right\| \leq \int_0^{\tilde{r}} \int_0^{t_1} \frac{1}{\tau} \, dt_2 \, dt_1 = \frac{\tilde{r}^2}{2\tau}
\]

35
where the last inequality holds because for any \( t \), \( \| \dot{\gamma}(t) \| \leq \tau^{-1} \) (the norm of the second fundamental form is bounded above by \( \tau^{-1} \). See Proposition 6.1 of \([25]\)).

To get simpler bounds, now suppose that \( r \leq (\sqrt{2} - 1)\tau \). We note that \( x \in [0, \sqrt{2} - 1] \) implies\(^\text{18} \) \( 1 - \sqrt{1 - 2x} \leq x + x^2 \). Thus
\[
\tilde{r} \leq \tau - \tau \sqrt{1 - 2\tau^{-1}r} \leq r + \frac{r^2}{\tau}
\]
\[
\| \gamma(\tilde{r}) - (\gamma(0) + \tilde{r}\dot{\gamma}(0)) \| \leq \tilde{r}^2 \leq \frac{r^2}{2\tau^3}(r + \tau)^2 \leq \frac{r^2}{\tau}.
\]

Sectional curvature may be used to bound the Jacobian of the exponential map, as follows\([23]\):

**Theorem 6.16.** Let \( M \) be a Riemannian manifold with sectional curvature bounded below and above by \( \kappa_- \) and \( \kappa_+ \). Then for \( x \in M \) and \( v \in T_x M \), the following holds:
\[
\min \left( 1, \frac{\sin \sqrt{\kappa_+} \| v \|}{\sqrt{\kappa_+} \| v \|} \right) \leq \| (d \exp_x)_v \| \leq \max \left( 1, \frac{\sin \sqrt{\kappa_-} \| v \|}{\sqrt{\kappa_-} \| v \|} \right)
\]
for all \( \| v \| \) if \( \kappa_+ \leq 0 \), and for \( \| v \| \leq \pi / \sqrt{\kappa_+} \) otherwise. The quantity \( \frac{\sin \sqrt{x}}{x} \) is taken to be 1 when \( x = 0 \).

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

**Corollary 6.17.** Let \( M \subseteq \mathbb{R}^D \) be a smoothly embedded compact Riemannian manifold with reach \( \tau \). Then for \( x \in M \) and \( v \in T_x M \) satisfying \( r := \| v \| \leq \pi \tau \), we have:
\[
\frac{\sinh \sqrt{2\tau^{-1}r}}{\sqrt{2\tau^{-1}r}} \leq \| (d \exp_x)_v \| \leq \frac{\sin \sqrt{\tau^{-1}r}}{\tau^{-1}r}
\]
In particular, if \( r \leq 2\tau \), then
\[
1 - \frac{r^2}{6\tau^2} \leq \| (d \exp_x)_v \| \leq 1 + \frac{r^2}{2\tau^2}.
\]

**Proof.** Norm of the second fundamental form is bounded above by \( \tau^{-1} \) \([25]\), and thus by the Gauss equation applied to sectional curvature (i.e. \( K(u, v) = \langle R(u, v)u, v \rangle = \langle [u, u], [v, v] \rangle - \| [u, v] \|^2 \) 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 \( x \in [0, 2\sqrt{2}] \). \(\square\)

\(^{18}\)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.

\(^{19}\)This can be manually checked by computing the first and the second derivative of \( x + x^3/4 - \sinh x \).
Lemma 6.18. For a metric space $M$ and its $n$-fold product space $M^n$, the following function is a metric on $M^n$:

$$d_\circ(x, y) := \min_{\sigma, \tau \in S_n} d_M(\sigma \cdot x, \tau \cdot y) = \min_{\sigma \in S_n} d_M(x, \sigma \cdot y)$$

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

$$d_\circ(x, y) = \|x - y\|$$

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

$$d_\circ(x, y) + d_\circ(y, z) = d_M(x, \sigma_{xy} \cdot y) + d_M(y, \sigma_{yz} \cdot z)$$

$$= d_M(x, \sigma_{xy} \cdot y) + d_M(\sigma_{xy} \cdot y, \sigma_{xy} \cdot \sigma_{yz} \cdot z)$$

$$\geq d_M(x, \sigma_{xy} \cdot \sigma_{yz} \cdot z)$$

$$\geq d_\circ(x, z)$$

This shows that $d_\circ$ 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: $\|x - \tau_{ij} \cdot z\| < \|x - z\|$, 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. □