HADES: Fast Singularity Detection
with Local Measure Comparison

Uzu Lim, Harald Oberhauser, and Vidit Nanda

ABSTRACT. We introduce HADES, an unsupervised algorithm to detect singularities in data. This algorithm employs a kernel goodness-of-fit test, and as a consequence it is much faster and far more scaleable than the existing topology-based alternatives. Using tools from differential geometry and optimal transport theory, we prove that HADES correctly detects singularities with high probability when the data sample lives on a transverse intersection of equidimensional manifolds. In computational experiments, HADES recovers singularities in synthetically generated data, branching points in road network data, intersection rings in molecular conformation space, and anomalies in image data.

1. Introduction

The Manifold Hypothesis asserts that high-dimensional datasets encountered in practice tend to concentrate near smooth manifolds of low intrinsic dimension. It is often used to justify the effectiveness of machine learning algorithms in high-dimensional settings, since the curse of dimensionality can be circumvented if the data concentrates on a low-dimensional manifold. It is, however, evident that several low-dimensional (and hence, visualisable) datasets do not satisfy the Manifold Hypothesis. Instead, such data can have *singularities* — points at which the local geometry does not resemble n-dimensional Euclidean space for any n. Prime examples of singular loci of datasets include branching points in neurons and cosmic filaments. Furthermore, standard image datasets (such as MNIST and CIFAR-10) are known to have non-constant intrinsic dimension [17], whereas a connected manifold must possess the same intrinsic dimension throughout.

Whenever such non-manifold behaviour within datasets is of interest, it becomes natural to wonder whether it can be accurately and automatically identified. Particularly in large, high-dimensional datasets where visual inspection is impossible, we seek tools to identify and locate singularities within datasets. Our focus here is on unsupervised singularity detection, where one has recourse neither to a plethora of training data, nor the opportunity to regenerate samples along an unknown probability measure.

MATHEMATICAL INSTITUTE, UNIVERSITY OF OXFORD, RADCLIFFE OBSERVATORY, ANDREW WILES BUILDING, WOODSTOCK RD, OXFORD OX2 6GG

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

1

This Paper. Here we propose HADES, a *Hypothesis-testing Algorithm for the Detection and Exploration of Singularities*. The basic philosophy is rooted in two elementary observations. First, by definition, an n-manifold locally resembles a standard Euclidean n-dimensional disk; and second, this resemblance can be precisely quantified by measuring the distance between (the local restriction of) an empirical measure and the uniform measure on the n-disk. HADES employs a goodness-of-fit test to measure this distance; we are therefore able to obtain a p-value for rejecting the null hypothesis that a given data point lies in the nonsingular locus of the underlying space.

Before proceeding to the details, we highlight three important features of the proposed algorithm below —

(1) **Efficiency:** HADES uses an explicit formula for kernel MMD (maximum mean discrepancy) to perform its goodness-of-fit test; this has a linear time complexity in the dimension of data, which forms a substantial improvement on the exponential complexity of the existing topological methods.

(2) **Correctness:** We show in Theorem 7 that HADES correctly identifies the singular set arising from the union of two transversely intersecting equidimensional submanifolds of Euclidean space. The proof uses tools from differential geometry, optimal transport theory, and concentration inequalities.

(3) **Validation:** In Section 5, we run HADES on several synthetic and real datasets. On synthetic data, we observe that singularities are correctly detected (Figure 3, 4). And in the real datasets where we have no access to ground truth, the singularities identified by HADES exhibit interesting and observable anomalous behaviour when compared to their nonsingular counterparts.

Related Work. Identifying non-manifold points and studying their structure often goes under the name of *stratified learning*, which attempts to model data using stratified spaces, instead of manifolds. An early example of studying non-manifold behaviour in data is seen in [25], where a Poisson mixture model was used to measure locally evaluated intrinsic dimension that may vary across data. Follow-up works considered data sampled from a union of multiple manifolds. In multi-manifold clustering, one starts with a data sampled from a union of intersecting manifolds and clusters data by separating them into the individual manifolds [44, 52, 43, 5, 2, 3]. Evidence for real world data containing multiple manifolds of mixed dimension have been recently studied [17, 18, 34]. We remark that unions of manifolds only constitute a small subset of all stratified spaces. While our algorithm doesn't recover the structural information of manifolds, it detects more diverse types of singularities not present in a union of manifolds.

Stratification learning received considerable attention from the topological data analysis community. The flagship tool here is *persistent homology*, which extracts topological
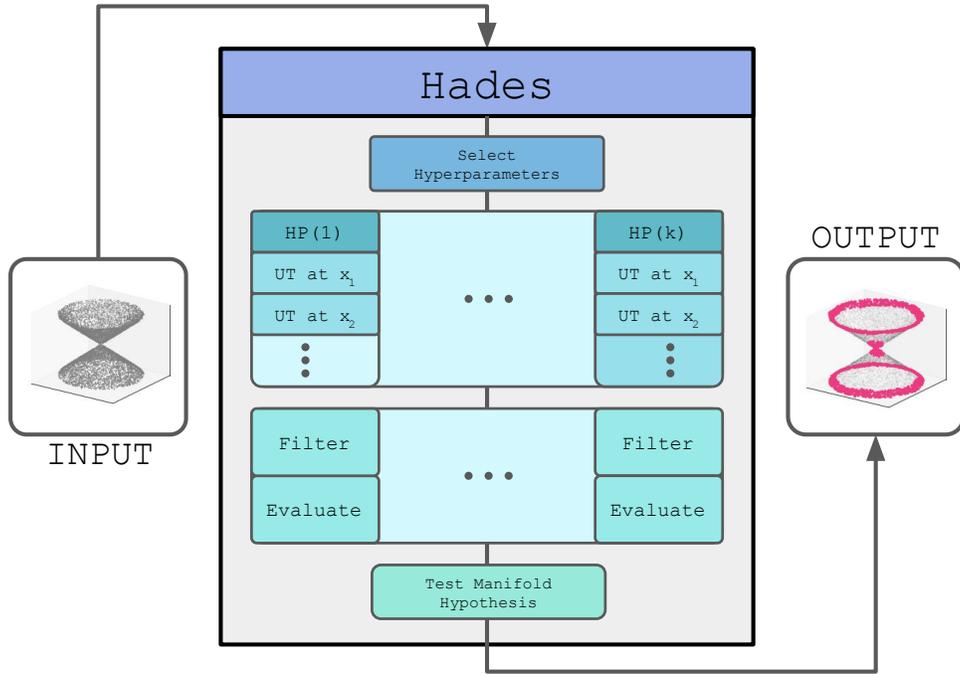
FIGURE 1. Schematic for a fully automated run of HADES. UT stands for Uniformity Test.

information at multiple scales of data. In [9, 10, 11], persistent intersection homology was used to discover stratified structure of data. In [51, 14, 15], algorithms for recovering low-dimensional stratification structure and homotopy type of a stratified space has been studied. Discovering a stratification structure of a given simplicial complex [37] and a complex projective variety [26] has also been studied. In [45, 50], persistent homology was used to detect singularities in data, and their algorithms have the same objective as our algorithm. Compared to their algorithms, our algorithm has a significantly improved time complexity and theoretical foundation.

Dimension estimation and reduction are key steps in our algorithm, for which we simply apply PCA locally. Nevertheless there are many more advanced dimension estimation methods available, such as [30, 21, 20, 24, 28, 12, 50]. Dimension reduction methods in the literature include [8, 54, 46, 35, 47]. For a survey of dimension estimation and dimension reduction algorithms, see [19, 48].

2. Algorithm

The main idea of HADES is to perform the Uniformity Test at the neighborhood of each data point. The Uniformity Test measures resemblance of each neighborhood to a flat disk, and thus determines whether a data point is smooth or singular. Hyperparameters are required for the Uniformity Test, which can be chosen either manually or automatically. The output is then filtered and evaluated, and the best set of hyperparameters is chosen based on the output evaluation. The fully automated version of HADES is illustrated on Figure 1.

2.1. Uniformity Test. The Uniformity Test works in two steps:

(1) Dimensionality reduction
(2) Goodness-of-fit test against the uniform distribution over a disk

We use PCA projection to reduce dimension of each *local* neighborhood, and this requires a threshold hyperparameter η. Suppose the points $\mathbf{z} \subset \mathbb{R}^D$ are plugged into the Uniformity Test. Firstly its estimated dimension $\hat{d}$ is the number of principal components required to explain η of the total variance. Then the local neighborhood of data is projected to the $\hat{d}$ principal components, producing $\tilde{\mathbf{z}}$.

The goodness-of-fit test uses a kernel method. Namely, we first compute the MMD (maximum mean discrepancy) and then compute the p-value associated to a null hypothesis. Given $\tilde{\mathbf{z}}$ obtained from projection, define the empirical measure $\hat{\mu}_z = n_{\hat{z}}^{-1} \sum_{x \in \tilde{\mathbf{z}}} \delta_x$, where δ_x is the Dirac delta measure centered at x. Let $\mathfrak{u}_{\hat{d}}$ be the uniform measure over the unit $\hat{d}$-dimensional disk. Let Δ be the kernel MMD (maximum mean discrepancy) associated to a kernel κ. Let $S = \Delta(\mathbf{X}_n, \mathfrak{u}_d)$ follow the null distribution, where $\mathbf{X}_n$ is an i.i.d. sample of size n drawn from $\mathfrak{u}_d$. We define the *singularity score $\sigma(z)$* and the *singularity p-value $\tilde{\sigma}(z)$* as follows:

$$\sigma(z) = \Delta(\hat{\mu}_z, \mathfrak{u}_{\hat{d}})$$

$$\tilde{\sigma}(z) = \mathbb{P}\left[S \geq \sigma(z)\right]$$

Let's understand the intuition behind the Uniformity Test. Suppose a data distribution μ is a uniform distribution on $M \subseteq \mathbb{R}^D$. Given x, consider $\mu_{x,r}$ = the restriction of μ to the ball of radius r centered at x. There are two possible outcomes of the Uniformity Test:

(1) **If M is smooth at x.** Then $\mu_{x,r}$ is supported on a slightly curved disk when r is sufficiently small. The dimensionality reduction flattens the slightly-curved disk into $\mu_{x,r}^\perp$, which is supported on a flat disk and has a small non-uniformity. Then $\mu_{x,r}^\perp \approx \mathfrak{u}_{\hat{d}}$ and thus the goodness-of-fit test fails to reject the null hypothesis. The point x is declared as highly *unlikely* to be singular.

4

(2) **If M is singular at x.** The support of $\mu_{x,r}$ does not resemble a flat disk no matter how small r is. The dimensionality reduction identifies a low-dimensional subspace containing the $\tilde{\mu}_{x,r}$, but $\tilde{\mu}_{x,r}$ only takes up a small portion of the Euclidean ball it spans. As such, the goodness-of-fit test will reject the null hypothesis. The point x is declared as highly *likely* to be singular.

Technical details. We use the following formula to compute the MMD of a power series kernel. Its proof is given in Appendix B.

THEOREM 1. *Let $\hat{\mu}_n = \frac{1}{n}(\delta_{x_1} + \cdots + \delta_{x_n})$ be a discrete (non-random) measure and let $\mathfrak{u}_d$ be the uniform distribution over the unit d-dimensional disk in $\mathbb{R}^d$. Let κ be a kernel given by $\kappa(x,y) = \sum_{k=0}^{\infty} a_k \langle x, y \rangle^k$, and let Δ be the MMD associated to κ. Then we have:*

$$\Delta^2(\hat{\mu}_n, \mathfrak{u}_d) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(x_i, x_j) + \sum_{k=0}^{\infty} a_{2k} \beta_{d,k} \left(\frac{d}{d+2k} - \frac{2}{n} \sum_{i=1}^{n} \|x_i\|^{2k} \right)$$

where the numbers $\beta_{d,k}$ are defined using the Gamma function Γ as:

$$\beta_{d,k} = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{d}{2}+1)\Gamma(k+\frac{1}{2})}{\Gamma(k+\frac{d}{2}+1)}$$

To evaluate the p-value arising from the MMD, we use its asymptotic distribution for large sample size n. The MMD is a V-statistic, for which asymptotic convergence under scaling by sample size holds true (Section 5, [42]):

THEOREM 2. *Let μ be a Borel measure on $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\hat{\mu}_n$ be the empirical measure of size n drawn from μ. Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function satisfying $\kappa(x,y) = \kappa(y,x)$, and let Δ be the MMD associated to κ. Then there is a converge in distribution as $n \to \infty$:*

$$n \cdot \Delta^2(\hat{\mu}_n, \mu) \longrightarrow c_\kappa + \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1)$$

Here Z_k are independent standard normals, $c_K := \mathbb{E}[\kappa(X,X)] - \mathbb{E}[\kappa(X,Y)]$, and λ_k are eigenvalues of the integral operator:

$$L[\phi] = \int \tilde{\kappa}(x,-)\phi(x)\,\mathrm{d}\,\mu(x)$$

where $\tilde{\kappa}(x,y) = \kappa(x,y) - \mathbb{E}[\kappa(X,y)] - \mathbb{E}[\kappa(x,Y)] + \mathbb{E}[\kappa(X,Y)]$.

We obtain the asymptotic distributions by Monte Carlo, i.e. by directly sampling the null statistics $n \cdot \Delta^2(\hat{\mu}_n, \mu)$ and using this to construct an empirical cumulative distribution function. To compute p-values for events that lie far outside the Monte Carlo simulation, we use exponential decay to estimate the p-values (see [39]).

Remark. There are many alternative choices for dimension estimation, dimensionality reduction, and the statistical distance used to perform the goodness-of-fit test. These parts

can be swapped out in a modular way, and the algorithm can be modified to match the user's needs. For example, a more sophisticated application might use a fractal dimension estimator, use UMAP [35] to perform local dimensionality reduction, and use statistical distances such as the Wasserstein distance. In practice, we found that the kernel MMD is more sensitive to detecting non-uniformity compared to the Wasserstein distance or its regularised Sinkhorn approximation [22]. Future improvements of the algorithm could be made by fine-tuning each of this step while keeping using the same conceptual framework. We also remark that Uniformity Test can be trivially parallelized since all computations are local. Thus running HADES parallel on multiple cores directly speeds it up.

2.2. Filtering and Evaluation. We now explain how to filter the singularity p-values into a binary label and evaluate the quality of the labeling. The labeling quality is ascribed to the hyperparameter set used to run the Uniformity Tests, and this gives us a way to choose the best set of hyperparameters.

The singularity p-values are filtered by applying a knee detection algorithm to the empirical probability density function of $\log(1/\tilde{\sigma}_i)$. Here, we apply the logarithm to separate very small p-values. We use Gaussian kernel density estimation to produce an empirical probability density function. We use the Kneed [40] algorithm to detect the knee of the probability density, and declare all points appearing after the knee to be singular points. The knee detection effectively identifies smooth points since their singularity p-values are relatively large, so that their distribution of $\log(1/\tilde{\sigma}_i)$ forms a concentrated mass near 0.

The quality of the binary label produced by the filtering step is evaluated using a metric we named *dispersion score*. The dispersion score is defined purely using data points and any binary label on them. The dispersion score is defined using *purity score* and *separation score*:

DEFINITION 3. Let $\mathbf{x} = (x_1, \ldots x_n)$, $\mathbf{y} = (y_1, \ldots y_n)$ be points and their binary labels, $x_i \in \mathbb{R}^D$, $y_i \in \{0, 1\}$. For each $i = 1, \ldots n$, let $\mathcal{N}(i) \subseteq \{1, \ldots n\}$ be a set satisfying $i \in \mathcal{N}(i)$. Define a partition $\mathcal{I}_0 \sqcup \mathcal{I}_1 = \{1, \ldots n\}$, where $\mathcal{I}_a = \{i \mid y_i = a\}$.

The *purity score* p_i is the proportion of indices $j \in \mathcal{N}(i)$ with $y_j = 1$:

$$p_i(\mathbf{y}, \mathcal{N}) = \frac{\#(\mathcal{N}(i) \cap \mathcal{I}_1)}{\#\mathcal{N}(i)}$$

The *separation score* is defined as an AUC (area-under-curve) score:

$$s_i(\mathbf{x}, \mathbf{y}, \mathcal{N}) = \text{AUC}\left\{ (t_{ij}, y_j) \,\middle|\, j \in \mathcal{N}(i) \right\}$$

where t_{ij} are real numbers defined as follows:

$$t_{ij} = \left\langle x_j - x_i, \frac{\tilde{x}_i}{\|\tilde{x}_i\|} \right\rangle, \quad \text{where} \quad \tilde{x}_i = \sum_{j \in \mathcal{N}(i) \cap \mathcal{I}_1} (x_j - x_i)$$

The *dispersion score* is defined as:

$$\mathfrak{D}(\mathbf{x}, \mathbf{y}, \mathcal{N}) = \alpha \cdot \mathcal{D}_1(P) + \sum_{i \in \mathcal{I}_1} \mathcal{D}_2(q_i), \quad \text{where} \quad q_i = 1 - \frac{1}{2}(s_i + p_i)$$

where $P = \#(\mathcal{I}_1)/n$ is the global purity score, α is a regularisation constant, and $\mathcal{D}_1, \mathcal{D}_2$ are *damping functions*, which are bijections $\mathcal{D}_i : [0,1] \to [0,1]$ satisfying $\mathcal{D}_i(x) \leq x$. [1]

Separation score quantifies how well the binary labels are cleanly separated along locally defined axes of direction, $\tilde{x}_i$. Indeed $\tilde{x}_i$ is the sum of displacements $x_j - x_i$ for which $y_j = 1$, and t_{ij} is the projected length of the displacement $x_j - x_i$ onto $\tilde{x}_i$. Thus, s_i measures how well the numbers t_{ij} can classify the binary labels y_j when $j \in \mathcal{N}(i)$.

Dispersion score detects points x_i for which both s_i and p_i are *simultaneously* small, whilst also penalising the degenerate case $P \approx 1$, when almost all points satisfy $y_i = 1$. The points x_i satisfying $i \in \mathcal{I}_1$ and $s_i + p_i \approx 0$ are far away from other indices $j \in \mathcal{I}_1$, and have poorly defined local boundary for separating the label 1 from the label 0. By using the damping functions $\mathcal{D}_1, \mathcal{D}_2$, we ensure that only the points x_i for which q_i is sufficiently large make a meaningful contribution to $\mathfrak{D}$, and also only the degenerate case for which $P \approx 1$ makes a meaningful contribution to $\mathfrak{D}$.

Remark. HADES is an *unsupervised learning* algorithm, for which there is no training dataset whose loss value can be minimised over many sets of hyperparameters. Instead, like clustering algorithms, the best set of hyperparameters is chosen by optimising a qualitatively defined criterion - the dispersion score. The dispersion score differs from the classical clustering quality measures that rewards concentration around centroids of clusters. The difference is that it aggregates *local* clustering information gathered from the data points, and thus the dispersion score can still be made small for complex shapes formed by the binary labels. This is adequate since the set of singularities of a stratified space have no reason to be concentrated around their centroid. (See Figure 3, the singular points marked in blue are not point-like clusters sought by the classical clustering quality measures.)

2.3. Hyperparameter selection. HADES uses the following three hyperparameters:

(1) **Local radius r.** Used to isolate neighborhoods.
(2) **PCA threshold η.** Used for dimension estimation.

---

[1]In the code, the default choice of the damping functions is given by $\mathcal{D}_1 = F_{0,2}$ and $\mathcal{D}_2 = F_{0.5,5}$, where $F_{a,b}(t) = (\frac{t-a}{1-a})^b$.

(3) **Kernel parameter α.** Used in MMD of the Uniformity Test.

The hyperparameters have the following effects. The radius r and threshold η both need to be at the right range to ignore noise and curvature. (For a thorough mathematical analysis, see [32]) The effect of the kernel parameter $\alpha \in (0,1)$ is less obvious. Choosing a different kernel parameter causes a different local singular geometry to be penalised. However, we found that the correctness of the output has a low sensitivity to the kernel parameter.

We explain how the sets of hyperparameters to run are automatically chosen by HADES. As explained before, the basic idea is to optimise the dispersion score over multiple sets of hyperparameters. These can be either supplied manually by the user or be chosen automatically by HADES. In the automatic hyperparameter selection, we use a grid $r \in [r_{min}, r_{max}]$, $\eta \in [\eta_{min}, \eta_{max}]$, $\alpha \in [\alpha_{min}, \alpha_{max}]$, where we use default values of $\eta \in [0.7, 0.9]$ and $\alpha \in [0.3, 0.7]$ for the PCA threshold and the kernel parameter.

Meanwhile, the range of radius hyperparameter r is chosen using a local scale detection algorithm. The idea here is to slowly enlarge a local neighborhood until the intrinsic dimension of the neighborhood stabilises. This process is done for multiple data points, and curves of intrinsic dimension estimates are averaged over them. The Kneed algorithm [40] is used to detect the threshold at which intrinsic dimension stabilises, by going backwards from the dimension estimate of the largest neighborhood and shrinking them, and detecting a knee of the curve. The standard intrinsic dimension estimator by Levina-Bickel [30] was used to calculate the intrinsic dimensions of the expanding neighborhoods. After obtaining the knee $\tilde{r}$, we use the range $r \in [1.5\tilde{r}, 5.0\tilde{r}]$.

When the optimal set of hyperparameters is found at the boundary of the grid search, HADES expands the search range towards that direction of the hyperparameter grid. For example, consider the grid search on $(r, \eta, \alpha) \in [0.1, 0.2] \times [0.7, 0.9] \times [0.3, 0.7]$, and suppose the dispersion score was minimised for $(r, \eta, \alpha) = (0.2, 0.8, 0.5)$. Since the optimal choice of r is found at the maximum of the range $[0.1, 0.2]$, HADES will do another grid search on the range $[0.2, 0.3] \times [0.7, 0.9] \times [0.3, 0.7]$ afterwards[2]. This process is repeated until a pre-specified end of search bounds are reached.

2.4. Testing the Manifold Hypothesis. We explain an algorithm used to test whether the geometric space underlying a dataset is a manifold. The main idea is the following: Given an iid sample $X_1, \ldots X_n$ drawn from a geometric space M and singularity p-values $\sigma_1, \ldots \sigma_n$ calculated from them, the following should hold:

- If M is a manifold, then $\sigma_1, \ldots \sigma_n$ should distribute uniformly over $[0, 1]$.

---

[2]In the implementation, we actually use a slightly more sophisticated method for expanding radius range. Observe that the volume of a d-dimensional ball with radius r is $\omega_d r^d$. We expand the radius parameter range such that this estimated local volume expands linearly.

- If M has a singularity, then $\sigma_1, \ldots \sigma_n$ should be concentrated near 0.

We give a heuristic argument for the above criterion. Firstly given any random variable Z with the probability density φ, the p-value variable $\tilde{Z} := \int_Z^\infty \varphi(t) \, d t$ is follows the uniform distribution over $[0, 1]$. This is because if we let $\tilde{a} = \int_a^\infty \varphi$, we have:

$$\Pr[\tilde{Z} \leq \tilde{a}] = \Pr\left[\int_Z^\infty \varphi \leq \int_a^\infty \varphi\right] = \Pr[Z \leq a] = \tilde{a}$$

Now for a fixed i, consider the singularity p-value σ_i calculated at the neighborhood of X_i by using the random sample $X_1, \ldots X_n$ drawn iid from a d-dimensional *manifold* M. Assuming that the local radius parameter is sufficiently small and n is sufficiently large, (1) the estimated dimension at X_i is d and (2) the empirical measure formed by the local neighborhood at X_i closely approximates the uniform distribution over a tangential disk at X_i.

Conditioning on k points among $X_1, \ldots X_n$ landing in the local neighborhood of X_i, we see that this marginal distribution of σ_i approximates $\tilde{Z}_k$. Here, $Z_k = \Delta(\hat{v}_k, \mathfrak{u}_d)$ where $\hat{v}_k$ is the empirical distribution constructed from an iid sample of size k drawn from the uniform distribution $\mathfrak{u}_d$, and $\tilde{Z}_k$ is the p-value of Z_k constructed in the way described above. Therefore, we expect σ_i to be approximately uniformly distributed over $[0, 1]$ when $r \to 0, n \to \infty$. Lastly, assuming sufficiently small r, most pairs of local neighborhoods at $X_1, \ldots X_n$ do not overlap, and we may expect the singularity scores $\sigma_1, \ldots \sigma_n$ to behave almost independently, so that their distribution over $[0, 1]$ is almost uniform. On the contrary, if M possessed a singularity, then near each singularity the singularity score (which is kernel MMD) becomes large and the singularity p-value will become small. Thus we expect a high concentration of singularity p-values near 0 if M has a singularity.

To differentiate between a uniform distribution of p-values over $[0, 1]$ and a distribution possessing a sharp spike of p-values near 0, we use the following three methods:

(1) SUPC (Small Uniformity p-value Concentration). Choose threshold values $\{q_1, \ldots q_k\} \subset [0, 1]$ and for each $q \in \{q_1, \ldots q_k\}$, calculate

$$\text{SUPC} := \max(q_1^\dagger, \ldots q_k^\dagger), \text{ where } q^\dagger = \frac{\#\{\sigma_i \leq q\}}{nq}$$

(2) UPUP (Uniformity p-value Uniformity p-value). Construct an empirical distribution $\hat{v}$ from $\sigma_1, \ldots \sigma_n$ and perform the uniformity test using the kernel MMD method (Theorem 1).

(3) KS (Kolmogorov-Smirnov). Again construct $\hat{v}$ from $\sigma_1, \ldots \sigma_n$ and perform the one-sample Kolmogorov-Smirnov test against the uniform distribution over $[0, 1]$.
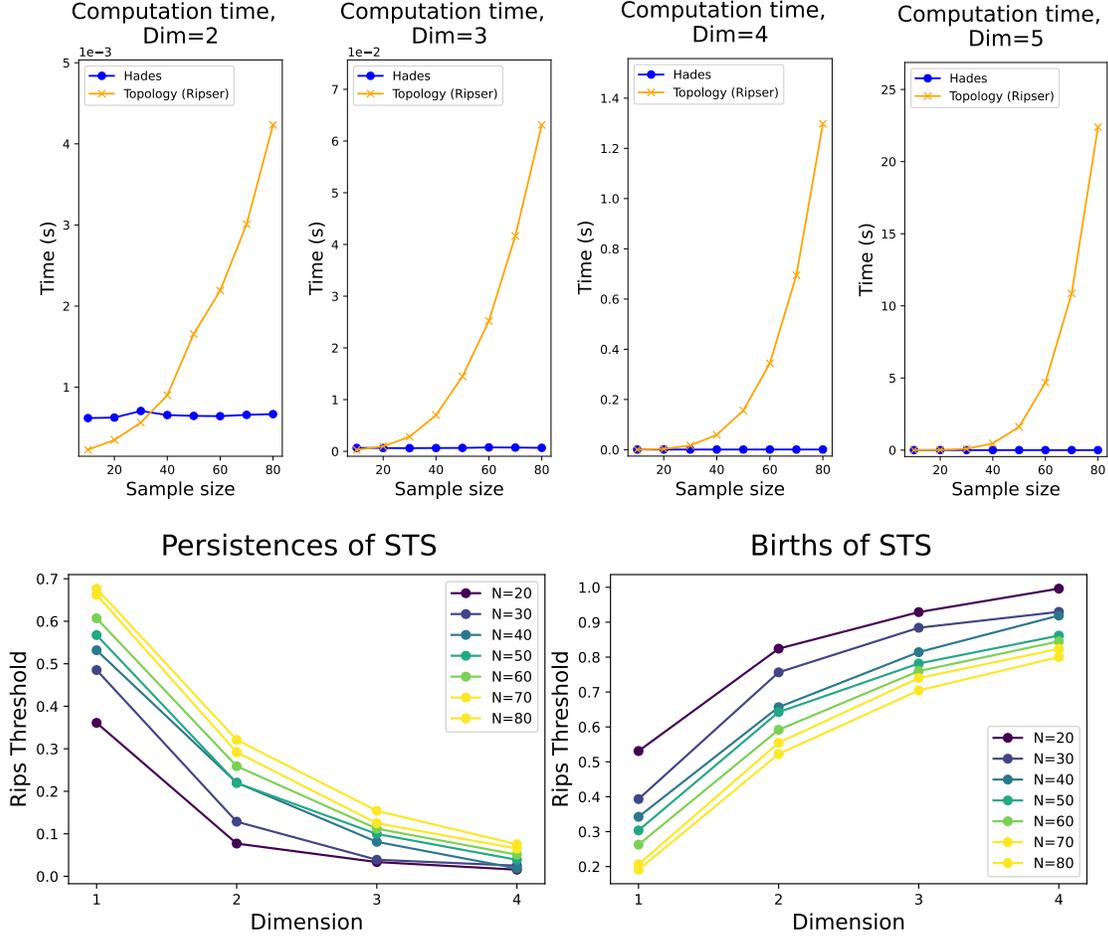
FIGURE 2. Top Row: Comparison of computation time of local shape analysis in HADES (blue) versus Ripser (orange), a highly optimised library for computing persistent homology. Bottom Row: Persistence of the significant topological signature (STS) in high-dimensional spheres decay significantly across dimensions.

3. Comparison with topological methods

We demonstrate significantly improved time complexity and statistical foundation of HADES in the singularity detection task, compared to the previous topological methods. Topological methods of singularity detection are based on persistent homology, a prominent tool from *topological data analysis* [**50, 45, 11, 9, 10, 51, 14, 15**]. Persistent homology computes topological features at varying scales of data, and the main idea behind topological methods for singularity detection is to compute persistent homology on local neighborhoods of data. In particular, the recent algorithms in [**50, 45**] uses the fact that a small annular neighborhood of a point on a manifold has the topology of a sphere, whose topology is well-understood.

Time complexity. A major advantage of HADES over singularity detection algorithms based on persistent homology is that HADES scales much better to high-dimensional data. We firstly have the following:

THEOREM 4. *The time complexity of the Uniformity Test on n points in $\mathbb{R}^D$ is $O(n^2D)$.*

PROOF. The dimensionality reduction step is an application of PCA. SVD is performed on a rectangular matrix of dimension $k \times D$, which has the time complexity of $O(k^2D)$ [49]. With the estimated dimension of $\hat{d}$, a matrix multiplication between a rectangular matrix of size $(k \times \hat{d})$ and a diagonal matrix of size $(\hat{d} \times \hat{d})$ is performed, for which the time complexity is $O(k\hat{d})$, which is less than $O(k^2D)$. This step thus amounts to the time complexity of:

$$T_1 = O(k^2D)$$

The goodness-of-fit step computes the MMD of a $\hat{d}$-dimensional point set of size k. Following the expression computed in Theorem 1, the time complexity for this step is:

$$T_2 = O(k^2d + k + kd) = O(k^2d)$$

$\square$

In comparison, the time complexity of persistent homology increases exponentially in the intrinsic dimension of data. The computational complexity of Ripser [7], a highly optimised Python package for computing persistent homology, is $O(s^3)$ where s is the number of simplices constructed. However a dataset of k points has a total of $s = \binom{k}{d+1} = O(k^{d+1})$ simplices of dimension d. A small annular local neighborhood of a d-dimensional manifold is topologically a $(d-1)$-sphere, and requires computationally constructing d-simplices. Therefore, the computational complexity of the $(d-1)$-th persistent homology group is $O(k^{3d+3})$. Persistent homology computation corresponds to the Uniformity Test, where in our algorithm we instead use PCA and kernel MMD. Using the computational complexity of the Uniformity Test given above, we have the following comparison of computational complexity incurred by local shape analysis:

$$\text{Uniformity test:} \quad O(k^2D)$$
$$\text{Persistent homology:} \quad O(k^{3d+3})$$

Thus we observe an exponential dependence of persistent homology computation on the intrinsic dimension d of data, whereas the Uniformity Test has a linear dependence on the ambient dimension D. When $D \gg d$, preprocessing data by dimensionality reduction allows us to circumvent the dependence of HADES on the ambient dimension D.

In Figure 2, we compare computation times of HADES (blue curve) and Ripser (orange curve). On each of the five plot shows computation times for a fixed dimension[3], but varying sample size. We observe that while HADES shows poorer performance than Ripser in low-dimensional data, the situation is quickly reversed in high-dimensional data.

Diminishing persistence. We observe from computational experiments that the topological signature of a high-dimensional sphere has a small persistence. This appears to present problem in applying the standard practice in topological data analysis, which declares a point on the persistent diagram as a genuine signal only if the point has a high persistence. In the case of the d-dimensional sphere, one seeks one highly persistent point on the d-dimensional persistence diagram, since the d-dimensional sphere has a 1-dimensional d-th homology group, and all other k-th homology groups of are zero for $k > 0$.

As such we define the *significant topological signature* (STS) at (d, n) to be the most persistent point of $\mathrm{PD}_d(\mathbf{X}_n)$, where $\mathrm{PD}_d(\mathbf{X}_n)$ is the d-th persistence diagram of the Rips filtration on $\mathbf{X}_n$, and $\mathbf{X}_n$ is an independently and identically distributed sample of size n from the d-dimensional sphere. Figure 2 tabulates birth times and persistences (y-axis) of the STS at (d, n) for varying sample size n (x-axis) the dimension d (different curves, colour-coded). The STS is significant because it is the main signal sought by the standard practice of topological data analysis.

Figure 2 indicates that the STS of a high-dimensional sphere has a small persistence and a large birth time. The small persistence tells us that STS becomes increasingly unreliable in high dimensions, due to it resembling "topological noise". This appears to defy the current paradigm of topological data analysis where highly persistent topological features are to be seen as genuine signal and other topological features are to be seen as noise. The large birth time tells us that one cannot use small connectivity threshold to detect STS, and therefore that it is difficult to reduce the number of high-dimensional simplices appearing in the full filtration of a point cloud.

This situation may be improved by using low-dimensional topological signal of high-dimensional spheres, which runs on smaller time complexity. In fact, even the 1-dimensional sphere (circle) exhibits systematic high-dimensional topological signals in large connectivity thresholds [31, 1], and high-dimensional spheres exhibit systematic low-dimensional topological signals.

---

[3]For d-dimensional data, we use samples of the unit d-dimensional ball for HADES and samples of the unit $(d-1)$-dimensional sphere for Ripser.

4. Theoretical Guarantee

In this section, we will define a mathematically precise version of the singularity detection algorithm, and state a theorem guaranteeing that the algorithm detects singularities correctly. In the following, we fix the ambient dimension D, threhsold parameter $\eta \in (0,1)$. We first define the PCA dimension estimator and projector precisely:

DEFINITION 5. Given μ, a Borel probability measure on $\mathbb{R}^D$, its *estimated dimension* $\hat{d}(\mu)$ and *linear regression* $\mathcal{L}(\mu)$ are defined as:

$$\hat{d}(\mu) = \min \left\{ k \,\middle|\, \frac{\lambda_{k+1} + \cdots + \lambda_D}{\lambda_1 + \cdots + \lambda_D} \leq \eta \right\}$$

$$\mathcal{L}(\mu) = \text{span}\left(\mathcal{E}(\mu, \lambda_1), \ldots, \mathcal{E}(\mu, \lambda_{\hat{d}(\mu)}) \right)$$

where $(\lambda_1, \ldots \lambda_D)$ are eigenvalues of $\Sigma[\mu]$, $\mathcal{E}(\mu, \lambda)$ is the λ-eigenspace of $\Sigma[\mu]$, and $\Sigma[\mu]$ is the covariance matrix of μ. The parameter η is implicit from the notations $\hat{d}(\mu), \mathcal{L}(\mu)$.

Using $\hat{d}$ and $\mathcal{L}$, we define the mathematically precise version of the *singularity score*. In the following let $\Delta(\mu, \nu)$ denote the kernel MMD associated to the Gaussian kernel $\kappa(x, y) = \exp(-\gamma \cdot \|x - y\|^2)$ for some fixed $\gamma > 0$. Also denote by $\mathfrak{u}_d$ the uniform measure over the unit d-dimensional disk centered at the origin. We first define the *abstract singularity score*, and use this for empirical measures to define the *empirical singularity score*.

DEFINITION 6. The *abstract singularity score* is defined as:

$$\sigma(\mu) = \Delta(\mu_\perp, \mathfrak{u}_{\hat{d}})$$

where $\hat{d} = \hat{d}(\mu)$ and $\mu_\perp = \Pi(\mu, \mathcal{L}\mu)$ is the pushforward of μ along the projection to $\mathcal{L}\mu$.

Let $\mathbf{x} = \{x_1, \ldots x_n\} \subset \mathbb{R}^D$ and let $r > 0$. Denote $\mathbf{x}(z) = \mathbf{x} \cap \mathcal{B}(z, r) \setminus \{z\}$, where $\mathcal{B}(z, r) \subseteq \mathbb{R}^D$ is the open ball of radius r, centred at z[4]. The *local empirical measure* of $\mathbf{x}$ at z is:

$$\hat{\mu}(z) = g_{z,r}\left(\frac{1}{\#\mathbf{x}(z)} \sum_{y \in \mathbf{x}(z)} \delta_y \right)$$

where $g_{z,r}(\nu)$ is the pushforward of a measure ν by the affine map $x \mapsto r^{-1}(x - z)$. The i-th *empirical singularity score* of $\mathbf{x}$ is defined as:

$$\hat{\sigma}_i(\mathbf{x}, r, \eta) = \sigma(\hat{\mu}(x_i))$$

Note that all of $\hat{d}, \mathcal{L}, \sigma$ depend on the choice of dimension estimation threshold η. We now state the setup and the main theorem.

---

[4]The point z is excluded for a technical reason concerning Wasserstein concentration inequalities, although the proofs can be modified to be included if necessary.

Setup. Let $M = M_1 \cup M_2$, where $M_1, M_2 \subseteq \mathbb{R}^D$ are smooth compact d-dimensional manifolds embedded in $\mathbb{R}^D$. Suppose there exist $d_0, \phi > 0$ such that the following holds for every $x \in M_1 \cap M_2$: the tangent spaces $T_x M_1$ and $T_x M_2$ intersect at a d_0-dimensional subspace, and all principal angles of the pair are $\geq \phi$. Let μ be the uniform measure over M, and let $\mathbf{X}_n = (X_1, \ldots X_n)$ be an iid[5] sample of size n drawn from μ.

THEOREM 7 (Theoretical guarantee). *There exist constants $\xi, \eta_-, \eta_+, c_A, c_B, r_0 > 0$ depending only on M such that the following holds. Given $\eta \in [\eta_-, \eta_+]$, $r \leq r_0$, and $q \in (0, 1)$, the following implications both hold for all i with probability at least q, when n is large enough:*

(1) When the distance of X_i from $M_1 \cap M_2$ is less than $c_A r$, then $\hat{\sigma}_i > 2\xi$.

(2) When the distance of X_i from $M_1 \cap M_2$ is greater than $c_B r$, then $\hat{\sigma}_i < \xi$.

where $\hat{\sigma}_i = \hat{\sigma}_i(\mathbf{X}, r, \eta)$.

The proof of the theorem requires much work, and it is presented in Appendix C. One main tool for the theorem is the Wasserstein distance, instead of the kernel MMD, which is possible since $\Delta(\mu, \nu) \leq \sqrt{2\gamma} \cdot W(\mu, \nu)$ for the Gaussian kernel $\kappa(x, y) = e^{-\gamma \|x - y\|^2}$ (Lemma 27). The advantage of the Wasserstein distance is that it is intuitively easy to prove geometric claims.

It has the following key ingredients:

(1) For a fixed $z \in M$ and as $r \to 0, n \to \infty$, the empirical measure $\hat{\mu}(z)$ converges to the uniform distribution over $T_z M° := T_z M \cap \mathcal{B}(0, 1)$, where $\mathcal{B}(0, 1) \subseteq \mathbb{R}^D$ is the unit ball of radius 1. Convergence is quantified using the Wasserstein distance. (Proposition 20)

(2) The singularity score function $\mu \mapsto \sigma(\mu)$ is a Lipschitz continuous function in μ, where Lipschitz continuity is quantified using the Wasserstein distance. (Proposition 46)

(3) The singularity score of the limiting measure at each point as $r \to 0, n \to \infty$ is zero at smooth points and positive at singular points (Propositions 48, 49).

(4) By moving sufficiently far away from the singularities, the local neighborhood of a point only isolates one manifold M_i at a time, instead of cutting through both M_1 and M_2 (Proposition 26).

To understand the proof, the reader is advised to start from the last part, Subsection C.5, and work backwards to identify the components used in the proof.

We remark that the constants c_A, c_B appearing in the theorem are unfortunately intrinsic features of the singularity detection. Suppose that $x \in M$, the ball of radius r is used to isolate local neighborhood of x, and that the distance of x to the singularities of M is $c \cdot r$

---

[5]independently and identically distributed

where $c \in \mathbb{R}^+$. Then there is an inherent ambiguity in choosing c_0 such that whenever $c > c_0$, x is declared non-singular, and whenever $c < c_0$, x is declared singular.

5. Experiments

We implemented HADES in Python and performed various computational experiments. Singularity detection lacks a ground truth for most real-world datasets, and is an *unsupervised learning* algorithm. We follow the standard 2-step approach to assess the performance of a singularity detection algorithm:

  (i) **Synthetic data.** We plot singularities detected from 2- and 3-dimensional datasets and visually inspect that the singularities are detected correctly. Then we detect singularities from families of high-dimensional synthetic datasets whose singularities are completely understood by construction, and use receiver-operating-characteristic (ROC) curve to quantitatively assess accuracy of the algorithm.

  (ii) **Real data.** We study datasets of road networks, cyclo-octane conformation, images of handwritten digits, and images of clothing items. For the road network and cyclo-octane conformation datasets, we recover the already-known locations of the singularities. For the image datasets whose geometry are not well-understood, we observe that images with high singularity score are anomalous from visual inspection.

For details of the experiments, see Appendix A.

5.1. Synthetic data: Visualisation and ROC Curves. We first apply HADES to the 2- and 3-dimensional point clouds in Figure 3, where singular points detected by the algorithm are marked blue. These synthetic datasets are generated from known data distributions of various geometric shapes, and uniform noise has been added to the datasets. They demonstrate that the algorithm is robust to noise and curvature. The algorithm simultaneously detects multiple types of singularities such as intersections, branching points, sharp corners, and cones. We also observe that no singularities are detected for the first row, which consist entirely of manifolds. This is enabled by the manifold hypothesis testing algorithm SUPC described in 2.4. The sizes of datasets range from 5,000 to 10,000. The time taken to extract singularities was about 3 minutes per dataset.

Going beyond visual inspection, we quantify accuracy of HADES on three families of geometric spaces:

  (1) One solid d-dimensional ball. (Singularity at boundary)

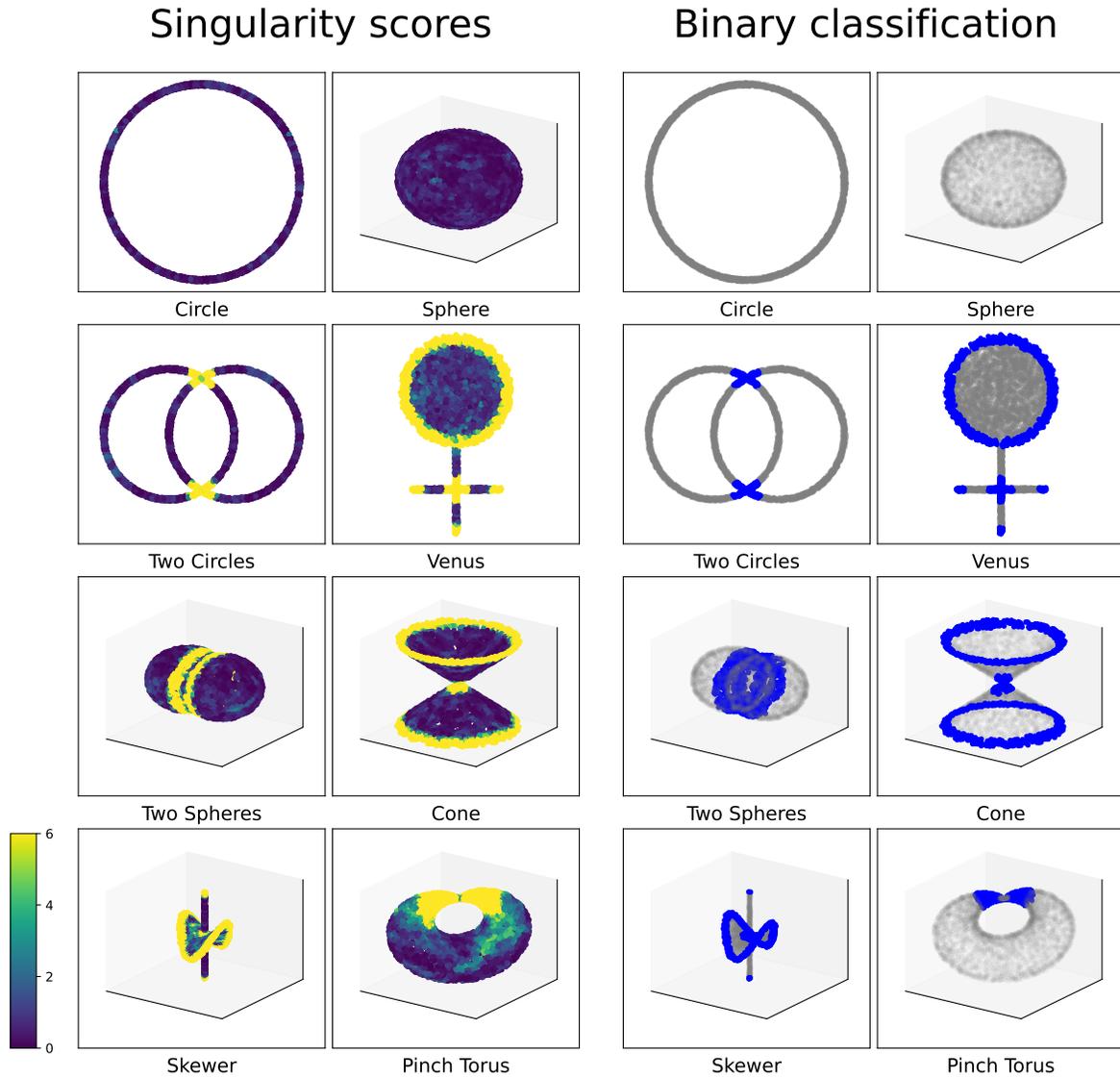 (2) Two d-dim. spheres intersecting at a $(d-1)$-dim. sphere. (Singularity at intersection)

FIGURE 3. Singularities discovered by HADES marked blue in synthetic datasets.

(3) Two 2d-dim. disks intersecting orthogonally at a d-dim. disk. (Singularity at intersection and boundary)

Visual inspection is inadequate for inspecting high-dimensional singularities, so we use receiver-operating characteristic (ROC) curve and its area-under-curve (AUC) to assess the performance. The AUC scores we obtain were all ≥ 0.89. The ROC curves and the AUC values are shown in Figures 4.
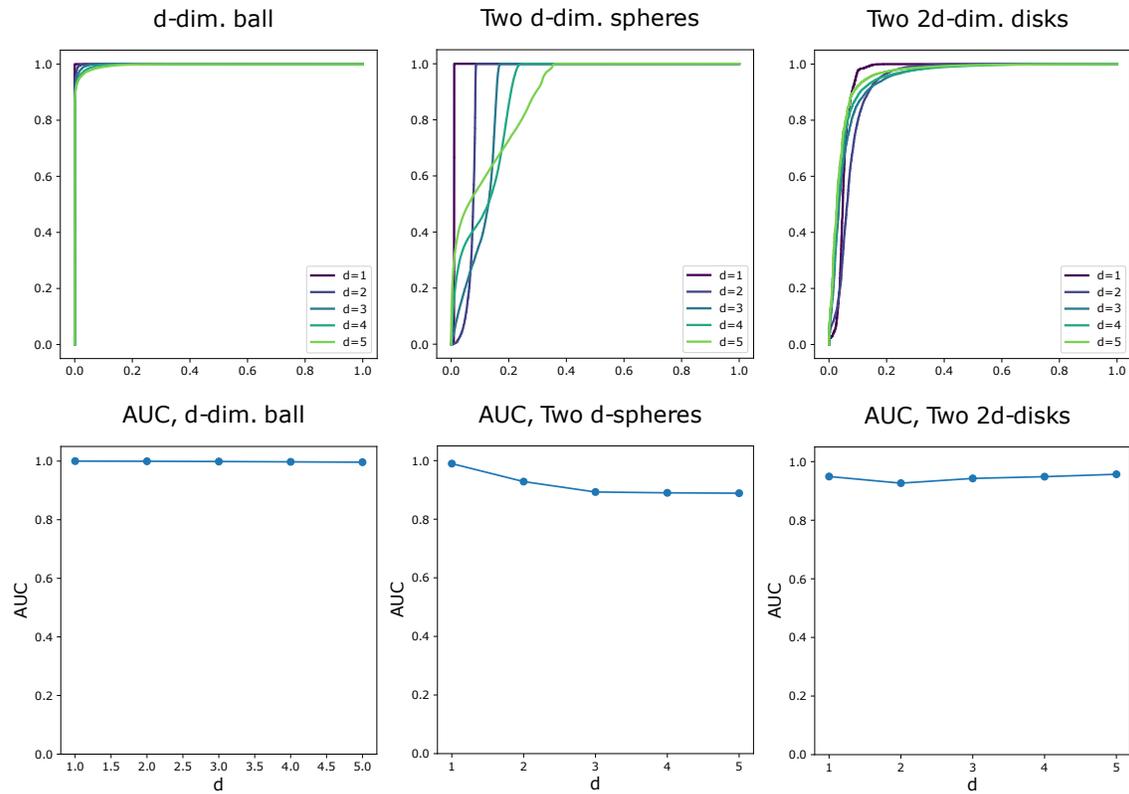
FIGURE 4. ROC curve and AUC scores of singularities discovered by HADES in synthetic datasets.
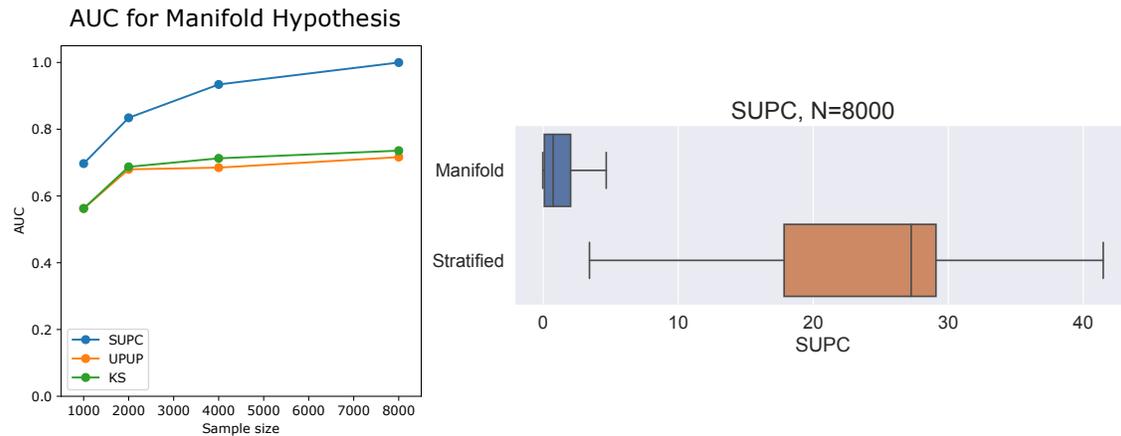

FIGURE 5. Testing the manifold hypothesis with HADES. Left: AUC of manifold hypothesis performed with three different scores: SUPC, UPUP, KS with synthetic datasets. Right: Boxplots of SUPC for manifolds vs. stratified spaces at sample size 8000.

5.2. Synthetic data: Manifold hypothesis. We test the manifold hypothesis with HADES using the methods described in Subsection 2.4. Unlike the *local* tests of detecting singularities, we perform a *global* test, on one dataset at a time. For this, datasets consisting of synthetically generated point clouds were created, with a *binary* label on whether each point cloud was a stratified space (with singularities) or a manifold (without singularities). Figure 5 shows the results. As sample sizes increase from 1000 to 8000, AUC values for all of SUPC, UPUP, KS increase, with UPUP and KS reaching just about 0.7 and SUPC reaching the AUC score 1.00. The boxplot on the right shows the distribution of SUPC scores for the manifolds and stratified spaces at sample size 8000, demonstrating a clean separation between the two types of data. This indicates that the manifold hypothesis can be effectively tested with SUPC.
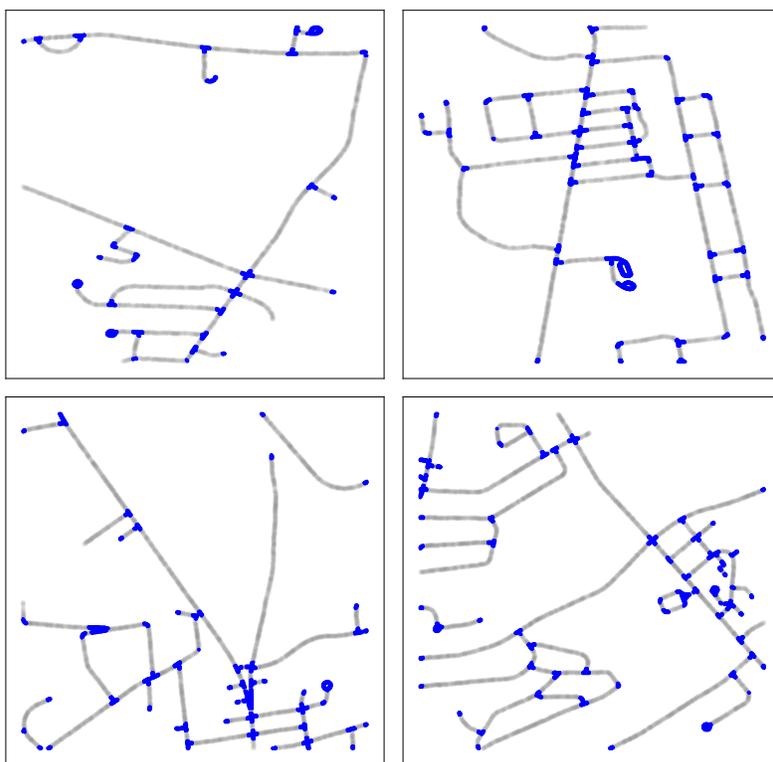
FIGURE 6. Singularities discovered by HADES marked blue in the Massachusetts Roads Dataset.

5.3. Real data: Road network. We apply HADES to the Massachusetts Roads Dataset [36], a dataset consisting of pixelised images of road networks in Massachusetts. Each road network is mathematically a planar embedding of a graph. Intersections and sharp corners of the road are singular points, and everything else is locally a straight line, and thus are smooth points. From Figure 6, visual inspection reveals that singularities are accurately detected. Each image had 1500×1500 resolution, containing 45,000 to 200,000 pixels with non-zero brightness values. The time taken to run each dataset ranges from 6 to 31 seconds. Expanding this analysis, the same computational experiment can be performed to other datasets that can be modeled as (1-dimensional) graphs, including images of neurons, and filamentary structures formed by galaxies.

5.4. Real data: Cyclo-octane conformation. We apply HADES to the dataset of cyclo-octane conformations. This dataset, introduced in [34], consists of 6040 points on the 24-dimensional space $\mathbb{R}^{24}$ that parametrises 3D positions of 8 carbon molecules in the cyclo-octane C_8H_{16}. The space of cyclo-octane was previously identified to be the union of a Klein bottle and a sphere, intersecting at two circles [34]. These two circles are
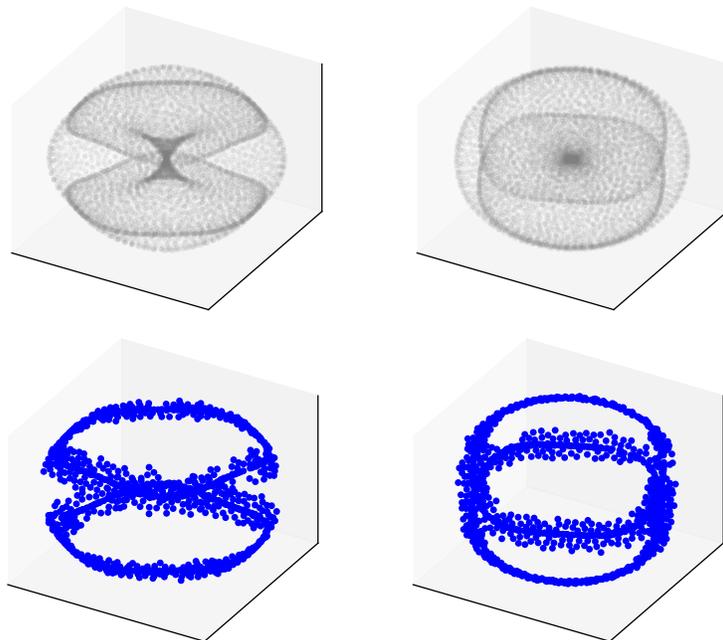
FIGURE 7. Singularities discovered by HADES marked blue in a cyclo-octane conformation dataset, which are union of two circles. Each row shows rotations of the same 3D Isomap projection of the 24-dimensional dataset. The first row shows the whole dataset and the second row shows singularities.

singularities of the space of conformations, and indeed they are correctly detected by HADES, as seen in Figure 7. The 3D projections of the conformation dataset, obtained using the dimensionality reduction algorithm Isomap [46], is displayed in Figure 7; we emphasise that the computation wasn't done on the 3D projection, and instead done directly on the original 24-dimensional data.

Running HADES on the entire conformation dataset took 5 seconds on a standard laptop. This shows great improvement from the previous benchmark for this dataset, in [45], in which their singularity detection algorithm *Geometric Anomaly Detection* took at least several hours on parallel processing, as informed by the first author on private communication.

5.5. Real data: Images of handwritten digits and clothings. We apply HADES to image datasets, of handwritten digits (MNIST) and clothing items (Fashion-MNIST), and find that images with high singularity scores are visibly more anomalous. MNIST is a standard dataset of images of handwritten digits [29] consisting of 60,000 data points, where there are 6,000 data points for each digit from $0, 1, \ldots 9$. Each data point is a $28 \times 28 = 784$-dimensional vector of brightness values between 0 and 1, where each entry of the vector indicates the brightness value of each pixel in the image. Similarly, the Fashion-MNIST

dataset consists of 28×28 images of 10 classes of clothing items[6], where there are 6,000 data points per class.

We applied HADES on MNIST and Fashion-MNIST datasets on each class of 6,000 images[7], and sorted the images according to their singularity scores. Prior to applying HADES, each 784-dimensional image vector was reduced to 100-dimensional vector by applying Discrete Cosine Transform. Figure 8 (MNIST) and Figure 9 (Fashion-MNIST) show the result, where the left half of each Figure displays images with the lowest singularity scores and the right half displays images with the highest singularity scores.

Images on the right half have irregular characteristics when compared to images on the left. This is explained from the fact that HADES assesses *local uniformity*. Indeed, images on the left look similar to each other, indicating that there are a lot more of similar images of small, subtle variations, thus locally constituting a more uniform distribution with a well-behaved variation. On the other hand, images on the right arise from irregular handwritings and clothing items. This means that there wouldn't be a uniform distribution of similar variations of the images, and thus picked up by HADES as highly singular. The computation time for running HADES on 6,000 images corresponding to each digit spanned 30 seconds to 45 seconds.

---

[6]T-shirt, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot

[7]Similar results were obtained from running HADES on the entire dataset of 60,000 datasets.

FIGURE 8. Images with the lowest singularity scores (left half) and the highest singularity scores (right half), upon applying HADES to the MNIST hand-written digits dataset.

Least Singular   Most Singular

FIGURE 9. Images with the lowest singularity scores (left half) and the highest singularity scores (right half), upon applying HADES to the Fashion-MNIST dataset.
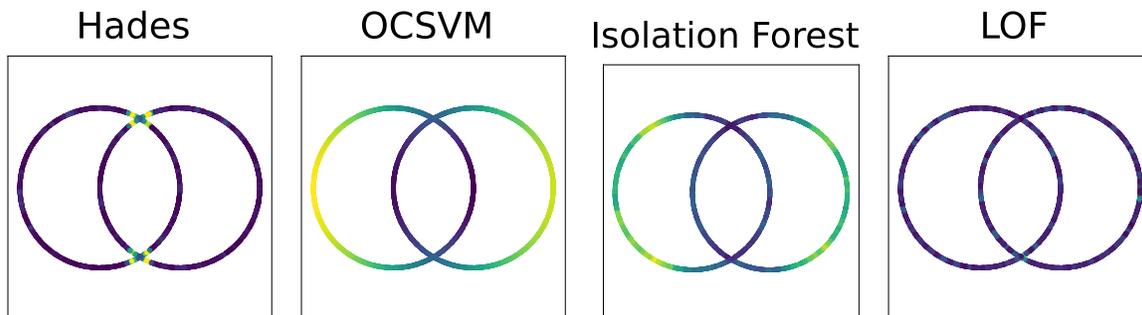
FIGURE 10. HADES is different from the *anomaly detection* algorithms; points marked as highly anomalous are marked in yellow.

5.6. Anomaly detection. We remark that HADES has a different objective to existing anomlay detection algorithms. Whereas HADES detects anomaly in local geometry, existing anomaly detection algorithms detect outliers. Along with HADES, three anomaly detection algorithms were tesed in Figure 10 (One-Class SVM [41], Isolation Forest [33], Local Outlier Factor [16]). The points with high anomaly score are marked in yellow (viridis colormap).

6. Conclusion

We introduced and studied HADES, an unsupervised learning algorithm that assigns a singularity score to data points. This is done by measuring how much the local geometry deviates from a manifold using a goodness-of-fit. The strengths of the algorithm are firstly its speed, in particular compared with recent topological approaches, and secondly that it can be seen as first step toward learning the full stratified space. The main disadvantage is that the goodness-of-fit algorithm simply detects what is *not* like a disk, and doesn't give a further details about the local geometry. This is where future research may blossom by using the richer information of local geometry provided by topological methods; for example one may compute persistent homology only at points declared to be singular by HADES. These research works together aim to create a computational toolbox for modeling general data using stratified spaces.

References

[1] M. Adamaszek and H. Adams. The vietoris–rips complexes of a circle. *Pacific Journal of Mathematics*, 290(1):1–40, 2017.

[2] E. Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *IEEE Transactions on Information Theory*, 57(3):1692–1706, 2011.

[3] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local pca. *The Journal of Machine Learning Research*, 18(1):253–309, 2017.

[4] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local pca. *The Journal of Machine Learning Research*, 18(1):253–309, 2017.

[5] A. Babaeian, A. Bayestehtashk, and M. Bandarabadi. Multiple manifold clustering using curvature constrained path. *PloS one*, 10(9):e0137986, 2015.

[6] P. Bartlett. Theoretical statistics; stat 210b. 2013.

[7] U. Bauer. Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

[8] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[9] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Inferring local homology from sampled stratified spaces. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 536–546. IEEE, 2007.

[10] P. Bendich and J. Harer. Persistent intersection homology. *Foundations of Computational Mathematics*, 11(3):305–336, 2011.

[11] P. Bendich, B. Wang, and S. Mukherjee. Local homology transfer and stratification learning. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1355–1370. SIAM, 2012.

[12] T. Birdal, A. Lou, L. J. Guibas, and U. Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021.

[13] E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

[14] Y. Bokor, K. Turner, and C. Williams. Reconstructing linearly embedded graphs: A first step to stratified space learning. *Foundations of Data Science*, 4(4):537–561, 2022.

[15] Y. Bokor Bleile. Towards stratified space learning: 2-complexes. *arXiv e-prints*, pages arXiv–2305, 2023.

[16] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[17] B. C. Brown, A. L. Caterini, B. L. Ross, J. C. Cresswell, and G. Loaiza-Ganem. The union of manifolds hypothesis. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.

[18] B. C. Brown, A. L. Caterini, B. L. Ross, J. C. Cresswell, and G. Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2022.

[19] F. Camastra and A. Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.

[20] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on pattern analysis and machine intelligence*, 24(10):1404–1407, 2002.

[21] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern recognition*, 47(8):2569–2581, 2014.

[22] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

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

[24] A. M. Farahmand, C. Szepesvári, and J.-Y. Audibert. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 265–272, 2007.

[25] G. Haro, G. Randall, and G. Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *Advances in Neural Information Processing Systems*, 19, 2006.

[26] M. Helmer and V. Nanda. Conormal spaces and whitney stratifications. *Foundations of Computational Mathematics*, pages 1–36, 2022.

[27] 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.

[28] K. Johnsson, C. Soneson, and M. Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):196–202, 2014.

[29] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

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

[31] U. Lim. Strange random topology of the circle. *arXiv preprint arXiv:2305.16270*, 2023.

[32] U. Lim, H. Oberhauser, and V. Nanda. Tangent space and dimension estimation with the wasserstein distance. *arXiv preprint arXiv:2110.06357*, 2021.

[33] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[34] S. Martin, A. Thompson, E. A. Coutsias, and J.-P. Watson. Topology of cyclo-octane energy landscape. *The journal of chemical physics*, 132(23):234115, 2010.

[35] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[36] V. Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.

[37] V. Nanda. Local cohomology and stratification. *Foundations of Computational Mathematics*, 20:195–222, 2020.

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

[39] H. Rootzen. Extreme value theory for moving average processes. *The Annals of Probability*, pages 612–652, 1986.

[40] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

[41] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[42] R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

[43] K. Slavakis, S. Salsabilian, D. S. Wack, S. F. Muldoon, H. E. Baidoo-Williams, J. M. Vettel, M. Cieslak, and S. T. Grafton. Riemannian multi-manifold modeling and clustering in brain networks. In *Wavelets and Sparsity XVII*, volume 10394, pages 9–24. SPIE, 2017.

[44] M. Soltanolkotabi and E. J. Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[45] B. J. Stolz, J. Tanner, H. A. Harrington, and V. Nanda. Geometric anomaly detection in data. *Proceedings of the National Academy of Sciences*, 117(33):19664–19669, 2020.

[46] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[47] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[48] L. Van Der Maaten, E. O. Postma, H. J. van den Herik, et al. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13, 2009.

[49] V. Vasudevan and M. Ramakrishna. A hierarchical singular value decomposition algorithm for low rank matrices. *arXiv preprint arXiv:1710.02812*, 2017.

[50] J. Von Rohrscheidt and B. Rieck. Topological singularity detection at multiple scales. 2023.

[51] L. Waas and T. Mäder. From samples to persistent stratified homotopy types. *arXiv preprint arXiv:2206.08926*, 2022.

[52] X. Wang, K. Slavakis, and G. Lerman. Multi-manifold modeling in non-euclidean spaces. In *Artificial Intelligence and Statistics*, pages 1023–1032. PMLR, 2015.

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

[54] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing*, 26(1):313–338, 2004.

Appendix

Appendix A. Experimental details

We implemented HADES in Python. Computational experiments were done with a standard laptop: Macbook Pro 2018 with the 2.6 GHz Intel Core i7 processor and the 16 GB 2400 MHz DDR4 memory.

Synthetic data: Low-dimensional. Details of the datasets used are in Table A.1. Hyperparameters were not specified to run each dataset; they were identified automatically by the algorithm. This means that only the point clouds were inputted to HADES to produce Figure 3.

Dataset	Description	Size	Time	T/S
Circle	One circle	5000	48	0.010
Sphere	One 2-dimensional sphere	10000	140	0.014
Two circles	Two intersecting circles	5000	78	0.016
Venus	One disk and two intersecting line segments attached	10000	147	0.015
Two spheres	Two intersecting spheres	10000	195	0.020
Cone	Two joined cones that are cut along top and bottom	10000	181	0.018
Skewer	One saddle surface skewered by a line segment	10000	175	0.018
Pinch torus	One torus, pinched along a neck	10000	184	0.018

TABLE A.1. Details of synthetic low-dimensional datasets, with the number of data points (size) and the time taken to run HADES on each data, and the time taken divided by size, in seconds (T/S).

Synthetic data: High-dimensional. There are three families of datasets here, constructed for values of $d = 1, 2, 3, 4, 5$:

(1) **Solid ball.** One d-dimensional unit ball (with filled interior).
(2) **Two Disks.** Two $2d$-dimensional unit disks intersecting at a d-dimensional disk. Constructed by taking two $2d$-dimensional unit disks in $\mathbb{R}^{3d}$, where the first disk spans axes $1, 2, \ldots 2d$ and the second disk spans axes $d + 1, d + 2, \ldots 3d$.
(3) **Two Spheres.** Two d-dimensional spheres intersecting at a $(d - 1)$-dimensional sphere. Constructed by taking two d-dimensional unit spheres in $\mathbb{R}^{d+1}$, whose centres are spaced apart by distance 1.

To calculate ROC curves, binary labels constituting ground truth are required. We define the ground truth label to depend on the local radius used for neighborhood isolation. This is because singular locus in a stratified space has measure zero, so that there is in fact

0 probability that a randomly sampled point from a stratified space is singular. However, the measure is positive when we thicken the singular locus by a radius, which is relevant to experimental setting where local neighborhoods used for data analysis may intersect the singular locus sufficiently closely.

We therefore define a binary label on a stratified space M with singular locus M_{sing} by declaring that $x \in M$ is s-close to singularity if the distance from x to M_{sing} is within s. Furthermore, when a local radius parameter r is used for HADES, we set $s = r/2$, so that a data point is declared singular iff it is within the distance $r/2$ from the singular locus. The scores used for HADES' classification is $\log(1/p_i)$, where p_i is the goodness-of-fit p-value of the i-th data point.

The volume of a d-dimensional disk of radius r is $\omega_d r^d$, where ω_d is a constant. Therefore, when $r < 1$, the volume diminishes exponentially in d. To account for this, we increased the radius parameter and the sample size as the dimension increased. We used the radius parameters $r_d = r_0^{1/d}$ and sample sizes $N_d = N_0 \alpha^d$ for constants r_0, N_0, α. For the Solid Ball dataset, we used $(r_0, N_0, \alpha) = (0.02, 15000, 1.5)$. For the Two Disks dataset, we used $(r_0, N_0, \alpha) = (0.1, 15000, 1.5)$ and for the Two Spheres dataset, we used $(r_0, N_0, \alpha) = (0.03, 15000, 1.5)$. The threshold parameter η was fixed at $\eta = 0.95$.

Synthetic data: Manifold hypothesis. The manifolds consisted of spheres, ellipsoids, Cartesian products of spheres, and torus. The stratified spaces consisted of unions of two spheres, cones, hollow cubes, and a union of three disks. For each specified sample size $N \in \{1000, 2000, 4000, 8000\}$, 20 copies of each point cloud were randomly generated. Thus for each sample size, a synthetic dataset consisting of 160 point clouds sampled from various manifolds and 120 point clouds sampled from various stratified spaces were generated. Then SUPC, UPUP, KS scores were calculated for each point cloud, and we tested their efficacy in distinguishing manifolds from stratified spaces.

Road networks. Starting from 1500×1500 images of aerial photographs of road networks, we extracted pixels that contain non-zero brightness values. Due to the uniform, clean nature of the images, we used fixed hyperparameters $(r, \eta) = (0.012, 0.8)$ (each image was normalised to fit in a unit square). Due to the large number of pixels in the images, only 10% of the pixels were used for singularity score calculations, and the singularity scores obtained here were extrapolated to the rest of the pixels. This significantly reduced computation time while still cleanly detecting singularities. The number of points contained in the images ranged from 45,000 to 200,000, and the time taken to run HADES on each image ranged from 6 to 31 seconds.

Cyclo-octane conformation. The consists of 6040 points on the 24-dimensional space $\mathbb{R}^{24}$ that parametrises 3D positions of 8 carbon molecules in the cyclo-octane C_8H_{16}. This

was taken from the publicly available repository of [45]. HADES was run directly on the 24-dimensional dataset, with the fixed hyperparameters $(r, \eta) = (0.35, 0.95)$. This took 6 seconds to run.

Image datasets. We first applied Discrete Cosine Transform to each image and reduced the 784-dimensional image vector into a 100-dimensional vector. This reduces the data dimension whilst retaining shape information of each digit. On this transformed dataset of 100-dimensional vectors, we ran HADES with the fixed hyperparameters of $(k, \eta) = (200, 0.95)$, where k is the number of nearest neighboring points used for local neighborhood isolation, and η is the PCA threshold parameter. Nevertheless the algorithm returns similar results when we change the hyperparameters. HADES was run separately on each class of images, although we observed similar results when we ran the algorithm on the whole dataset.

Appendix B. Explicit computation of the kernel MMD

The content of this section is the proof of Theorem 1, which is the formula we use to evaluate the MMD used in the Uniformity Test.

The following formula holds in general for kernel MMD:

$$
\Delta_\kappa(\mu, \nu) = \iint \kappa(x, y) \, d(\mu - \nu)(x) \, d(\mu - \nu)(y)
$$
$$
= \iint \kappa(x, y) \, d\mu(x) \, d\mu(y) + \iint \kappa(x, y) \, d\nu(x) \, d\nu(y) - 2 \iint \kappa(x, y) \, d\mu(x) \, d\nu(y)
$$

PROPOSITION 8. *Suppose κ is a kernel on $\mathbb{R}^d$. Let $U_d = \omega_d^{-1} \mathcal{H}^d|_{\mathbb{D}_d}$ be the uniform distribution over the unit d-dimensional ball where $\omega_d = \pi^{d/2}/\Gamma\left(1 + \frac{d}{2}\right)$ is the volume of the unit d-dimensional ball. Let $\hat{\mu}_n = \frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n})$ be the iid sample drawn uniformly from U_d. Then we have the following.*

$$
\Delta_\kappa(\mu, \nu) = \frac{1}{\omega_d^2} \bar{F}_{\kappa,d} + \frac{1}{n^2} \sum_{i,j} \kappa(x_i, x_j) - \frac{2}{n\omega_d} \sum_{i=1}^{n} F_{\kappa,d}(x_i)
$$

where

$$
F_{\kappa,d}(x) = \int_{\|y\| \leq 1} \kappa(x, y) \, dy, \quad \bar{F}_{\kappa,d} = \int_{\|x\| \leq 1} F_{\kappa,d}(x) \, dx
$$

PROOF. We separately evaluate the terms:

$$
\iint \kappa(x, y) \, dU_d(x) \, dU_d(y) = \frac{1}{\omega_d^2} \iint_{\|x\| \leq 1, \|y\| \leq 1} \kappa(x, y) \, dx \, dy = \frac{1}{\omega_d^2} \int_{\|x\| \leq 1} F_{\kappa,d}(x) \, dx
$$

$$
\iint \kappa(x, y) \, d\hat{\mu}_n(x) \, d\hat{\mu}_n(y) = \frac{1}{n^2} \sum_{i,j} \kappa(x_i, x_j)
$$

$$
\iint \kappa(x, y) \, d\hat{\mu}_n(x) \, dU_d(y) = \frac{1}{n\omega_d} \sum_{i=1}^{n} F_{\kappa,d}(x_i)
$$

$\square$

The volume of $\mathbb{S}^{d-1}$ is equal to $d\omega_d$. The volume of a d-dimensional ball is equal to $\pi^{d/2}/\Gamma(\frac{d}{2} + 1)$.

LEMMA 9. *Let $\kappa(x, y)$ be a kernel on $\mathbb{R}^d$ invariant under rotation, i.e. for any orthogonal transform $A \in O(d)$, we have that $\kappa(Ax, Ay) = \kappa(x, y)$. Then whenever $\|x_1\| = \|x_2\|$, we have $F_{\kappa,d}(x_1) = F_{\kappa,d}(x_2)$. Furthermore, the following identity holds:*

$$
\bar{F}_{k,d} = d\omega_d \int_0^1 F_{\kappa,d}(r) r^{d-1} \, dr
$$

PROOF. Rotational invariance of the unit ball directly implies that $\|x_1\| = \|x_2\|$ gives $F_{\kappa,d}(x_1) = F_{\kappa,d}(x_2)$. We also evaluate:

$$\bar{F}_{k,d} = \int_{\|x\|\leq 1} F_{\kappa,d}(x)\,\mathrm{d}\,x = \mathcal{H}^{d-1}(\mathbb{S}^{d-1}) \int_0^1 F_{\kappa,d}(r) r^{d-1}\,\mathrm{d}\,r = d\omega_d \int_0^1 F_{\kappa,d}(r) r^{d-1}\,\mathrm{d}\,r$$

$\square$

We also note the simple expression for the expected value.

PROPOSITION 10. *We have:*

$$\mathbb{E}_{\hat{\mu}_n}\Delta_\kappa^2(\hat{\mu}_n, \mu) = \frac{1}{n} \cdot \left(\int \kappa(x,x)\,\mathrm{d}\,\mu(x) - \iint \kappa(x,y)\,\mathrm{d}\,\mu(x)\,\mathrm{d}\,\mu(y) \right)$$

PROOF.

$$\mathbb{E}_{\hat{\mu}_n}\Delta_\kappa^2(\hat{\mu}_n, \mu) = \iint \kappa\,\mathrm{d}\,x\,\mathrm{d}\,y - 2 \iint \kappa\,\mathrm{d}\,x\,\mathrm{d}\,y + \frac{C}{n} + \frac{n(n-1)}{n^2}\mathbb{E}\sum_{i\neq j}\kappa(X_iX_j)$$

$$= \frac{C}{n} - \frac{1}{n}\mathbb{E} \iint \kappa\,\mathrm{d}\,x\,\mathrm{d}\,y$$

$\square$

PROPOSITION 11. *For $k \geq 0$, let $\kappa(x,y) = \langle x, y\rangle^k$ be a monomial kernel. Then we have that:*

$$F_{\kappa,d}(r) = \omega_{d-1}\,\mathrm{B}\left(\frac{k+1}{2}, \frac{d+1}{2}\right) = \frac{\pi^{(d-1)/2}\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k+d}{2}+1)} \cdot r^k \qquad \textit{, if k is even}$$

$$\bar{F}_{\kappa,d} = \frac{d}{d+k}\omega_d\omega_{d-1}\,\mathrm{B}\left(\frac{k+1}{2}, \frac{d+1}{2}\right) = \frac{2^d\pi^{d-1}}{(d-1)!} \cdot \frac{1}{d+k}\,\mathrm{B}\left(\frac{k+1}{2}, \frac{d+1}{2}\right) \quad \textit{, if k is even}$$

and both expressions are zero if k is odd.

PROOF. We directly evaluate:

$$F_{\kappa,d}(r) = \int_{\|y\|\leq 1} \kappa(r \cdot e_1, y)^k\,\mathrm{d}\,y$$

$$= \int_{-1}^{+1} \int_{\|z\|\leq\sqrt{1-s^2}, z\in\mathbb{R}^{d-1}} (rs)^k\,\mathrm{d}\,z\,\mathrm{d}\,s$$

$$= \omega_{d-1}r^k \int_{-1}^{+1} s^k(1-s^2)^{(d-1)/2}\,\mathrm{d}\,s$$

32

By symmetry, odd k implies that $F_{\kappa,d}(r) = 0$. For even k, we may write:

$$
\begin{aligned}
\int_{-1}^{+1} s^k (1-s^2)^{(d-1)/2} \, d s &= 2 \int_0^1 (s^2)^{k/2} (1-s^2)^{(d-1)/2} \, d s \\
&= 2 \int_0^1 t^{k/2} (1-t)^{(d-1)/2} (2\sqrt{t})^{-1} \, d t \\
&= \int_0^1 t^{(k-1)/2} (1-t)^{(d-1)/2} \, d t \\
&= B\left(\frac{k+1}{2}, \frac{d+1}{2}\right)
\end{aligned}
$$

where $B(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1} \, d t = \Gamma(u)\Gamma(v)/\Gamma(u+v)$ is the Beta function. Thus for even k we get that:

$$
F_{\kappa,d}(r)/r^k = \omega_{d-1} B\left(\frac{k+1}{2}, \frac{d+1}{2}\right) = \frac{\pi^{(d-1)/2}}{\Gamma(1+\frac{d-1}{2})} \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{d+1}{2})}{\Gamma(\frac{k+d}{2}+1)} = \frac{\pi^{(d-1)/2}\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k+d}{2}+1)}
$$

For $\bar{F}_{\kappa,d}$ and even k,

$$
\begin{aligned}
\bar{F}_{\kappa,d} &= d\omega_d \int_0^1 F_{\kappa,d}(r) r^{d-1} \, d r \\
&= d\omega_d \omega_{d-1} B\left(\frac{k+1}{2}, \frac{d+1}{2}\right) \int_0^1 r^{k+d-1} \, d r \\
&= \frac{d}{d+k} \cdot \omega_d \omega_{d-1} \cdot B\left(\frac{k+1}{2}, \frac{d+1}{2}\right)
\end{aligned}
$$

We note that:

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

where we used the Lagrange duplication formula $\Gamma(z)\Gamma(z+\frac{1}{2}) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$ for $z = (d+1)/2$. This gives:

$$
\bar{F}_{\kappa,d} = \frac{2^d \pi^{d-1}}{(d-1)!} \cdot \frac{1}{d+k} B\left(\frac{k+1}{2}, \frac{d+1}{2}\right)
$$

$\square$

PROPOSITION 12. *Let $\kappa(x,y) = \sum_{k=0}^{\infty} a_k \langle x, y \rangle^k$ be a power series kernel with $a_k \geq 0$. Then we have that:*

$$F_{\kappa,d}(r) = \omega_{d-1} \sum_{k=0}^{\infty} B\left(k + \frac{1}{2}, \frac{d+1}{2}\right) a_{2k} r^{2k}$$

$$= \pi^{(d-1)/2} \sum_{k=0}^{\infty} \frac{\Gamma(k + \frac{1}{2})}{\Gamma(k + \frac{d}{2} + 1)} \cdot a_{2k} r^{2k}$$

$$\bar{F}_{\kappa,d} = d\omega_d \omega_{d-1} \sum_{k=0}^{\infty} B\left(k + \frac{1}{2}, \frac{d+1}{2}\right) \frac{a_{2k}}{d + 2k}$$

$$= \frac{2^d \pi^{d-1}}{(d-1)!} \sum_{k=0}^{\infty} B\left(k + \frac{1}{2}, \frac{d+1}{2}\right) \frac{a_{2k}}{d + 2k}$$

PROOF. This is proven by interchanging sum and integral, using the previous proposition for monomial kernel, and noting that only the even indexed terms survive. The exchange of sum and integral is justified because of absolute convergence, following from $a_k \geq 0$. $\qquad \square$

We now prove Theorem 1.

THEOREM 13. *Let $\hat{\mu}_n = \frac{1}{n}(\delta_{x_1} + \cdots + \delta_{x_n})$ be a discrete (non-random) measure and let $\mathfrak{u}_d$ be the uniform distribution over the unit d-dimensional disk in $\mathbb{R}^d$. Let κ be a kernel given by $\kappa(x,y) = \sum_{k=0}^{\infty} a_k \langle x, y \rangle^k$. Then we have:*

$$\Delta_\kappa^2(\hat{\mu}_n, \mathfrak{u}_d) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(x_i, x_j) + \sum_{k=0}^{\infty} a_{2k} \beta_{d,k} \left(\frac{d}{d + 2k} - \frac{2}{n} \sum_{i=1}^{n} \|x_i\|^{2k}\right)$$

where

$$\beta_{d,k} = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{d}{2} + 1)\Gamma(k + \frac{1}{2})}{\Gamma(k + \frac{d}{2} + 1)}$$

and $\Gamma(t)$ is the Gamma function.

PROOF. The claim can be restated as $\Delta_\kappa(\hat{\mu}_n, U_d) = A + B - C$, where

$$A = \frac{1}{n^2} \sum_{i,j} \kappa(x_i, x_j)$$

$$B = \frac{\Gamma(\frac{d}{2} + 1)}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{\Gamma(k + \frac{1}{2}) a_{2k}}{\Gamma(k + \frac{d}{2} + 1)} \frac{d}{d + 2k}$$

$$C = \frac{2}{n} \cdot \frac{\Gamma(\frac{d}{2} + 1)}{\sqrt{\pi}} \sum_{k=1}^{n} \frac{\Gamma(k + \frac{1}{2}) a_{2k}}{\Gamma(k + \frac{d}{2} + 1)} \sum_{i=1}^{n} \|x_i\|^{2k}$$

We firstly know that:

$$\Delta_\kappa(\hat{\mu}_n, U_d) = A + B - C$$

34

$$A = \frac{1}{n^2} \sum_{i,j} \kappa(x_i, x_j)$$

$$B = \frac{\omega_{d-1}}{\omega_d} \sum_{k=0}^{\infty} \frac{d}{d+2k} \, \mathrm{B}\left(k + \frac{1}{2}, \frac{d+1}{2}\right) a_{2k}$$

$$C = \frac{2}{n} \frac{\omega_{d-1}}{\omega_d} \sum_{k=1}^{n} \mathrm{B}\left(k + \frac{1}{2}, \frac{d+1}{2}\right) a_{2k} \sum_{i=1}^{n} \|x_i\|^{2k}$$

We compute:

$$B = \frac{\omega_{d-1}}{\omega_d} \sum_{k=0}^{\infty} \frac{d}{d+2k} \, \mathrm{B}\left(k + \frac{1}{2}, \frac{d+1}{2}\right) a_{2k}$$

$$= \pi^{-1/2} \frac{\Gamma(\frac{d+2}{2})}{\Gamma(\frac{d+1}{2})} \sum_{k=0}^{\infty} \frac{d}{d+2k} \frac{\Gamma(k+\frac{1}{2})\Gamma(\frac{d+1}{2})}{\Gamma(k+\frac{d}{2}+1)} a_{2k}$$

$$= \pi^{-1/2} \Gamma\left(\frac{d}{2} + 1\right) \sum_{k=0}^{\infty} \frac{d}{d+2k} \frac{\Gamma(k+\frac{1}{2})a_{2k}}{\Gamma(k+\frac{d}{2}+1)}$$

and

$$C = \frac{2}{n} \frac{\omega_{d-1}}{\omega_d} \sum_{k=1}^{n} \mathrm{B}\left(k + \frac{1}{2}, \frac{d+1}{2}\right) a_{2k} \sum_{i=1}^{n} \|x_i\|^{2k}$$

$$= \frac{2}{n} \pi^{-1/2} \Gamma\left(\frac{d}{2} + 1\right) \sum_{k=1}^{n} \frac{\Gamma(k+\frac{1}{2})a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \sum_{i=1}^{n} \|x_i\|^{2k}$$

$\square$

Error rate. The MMD expression above involves the following infinite series:

$$B = \frac{\Gamma(\frac{d}{2}+1)}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{\Gamma(k+\frac{1}{2})a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \frac{d}{d+2k}$$

$$C = \frac{2}{n} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\sqrt{\pi}} \sum_{k=1}^{n} \frac{\Gamma(k+\frac{1}{2})a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \sum_{i=1}^{n} \|x_i\|^{2k}$$

Up to constant, the series are:

$$\sum_{k=0}^{\infty} \frac{\Gamma(k+\frac{1}{2})a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \frac{d}{d+2k}, \quad \sum_{k=1}^{\infty} \frac{\Gamma(k+\frac{1}{2})a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \sum_{i=1}^{n} \|x_i\|^{2k}$$

Observe that the summands $d/(d+2k)$ and $\sum_i \|x_i\|^{2k}$ are non-increasing in k. Also, we substitute in $a_{2k} = \gamma^{2k}$. We are interested in the relative error of estimation, so that we are then further simply interested in:

$$\sum_{k=0}^{\infty} \frac{\Gamma(k+\frac{1}{2})\gamma^{2k}}{\Gamma(k+\frac{d}{2}+1)}$$

At $d = 1$, this is

$$\sum_{k=0}^{\infty} \frac{\Gamma(k + \frac{1}{2})\gamma^{2k}}{\Gamma(k + \frac{3}{2})} \leq \sum_{k=1}^{\infty} \frac{\gamma^{2k}}{k} = -\log(1 - \gamma^2)$$

Thus a crude error bound is given by considering the convergence rate of the function $\log(1 - \gamma^2)$.

By direct evaluation, the evaluation up to $k = 10$ of the Taylor series and $\gamma \leq 0.9$ gives relative error ≤ 0.03 and $\gamma \leq 0.5$ gives relative error $\leq 10^{-6}$.

Appendix C. Proof of the theoretical guarantee

This section, spanning across the rest of the paper, contains the proof of our main theoretical guarantee, Theorem 7. We develop several technical tools step-by-step and proceed towards the final subsection. As such, the reader is advised to start from the last subsection and work backwards to understand details of the proof.

C.1. Wasserstein inequalities. We prove several inequalities involving the Wasserstein distance.

Concentration

In this subsection we study how the empirical measure approximates the underlying measure, in the sense of Wasserstein distance. The main objective of this subsection is Proposition 20, which is derived by simplifying Corollary 1.2 from [13]. We use the notion of *covering number* for this:

$$N_{\text{cover}}(M, r) = \min \left\{ m \,\middle|\, \exists x_1, \ldots x_m \in M, \cup_{i=1}^m \mathcal{B}(x_i, r) \supseteq M \right\}$$

THEOREM 14 (Boissard-Le Gouic). *Let* (M, d, μ) *be a measured Polish space of a finite diameter* R. *Suppose that there exist* $\alpha > 2p, \beta > 0$ *so that the following holds for for every* $0 < r < R/4$:

$$N_{\text{cover}}(M, r) \leq \beta \left(\frac{R}{r} \right)^\alpha$$

Then the following holds:

$$\mathbb{E}\left[W_p(\hat{\mu}_m, \mu) \right] \leq \frac{64R}{3} \cdot \left(\frac{2p}{\alpha - 2p} \right)^{2p/\alpha} \cdot \left(\frac{\beta}{m} \right)^{1/\alpha}$$

To apply this to compact subsets of a Euclidean space, we use a lemma from [6]:

LEMMA 15. *The D-dimensional unit ball* $\mathcal{B}_D$ *satisfies:*

$$N_{\text{cover}}(\mathcal{B}_D, r) \leq (1 + 2r^{-1})^D$$

PROOF. It is easy to see that a *maximal packing* by N' balls of radii $r/2$ is also a covering by balls of radii r[8], and so we have $N_{\text{cover}}(\mathcal{B}_D, r) \leq N'$. Now consider a maximal packing by balls of radii $r/2$ centered at $x_1, \ldots x_{N'}$. Then,

$$\cup_i \mathcal{B}(x_i, r/2) \subseteq (1 + r/2) \cdot \mathcal{B}_D$$
$$\implies N' \cdot (r/2)^D \leq (1 + r/2)^D$$

$\square$

---

[8]Suppose that balls of radii $r/2$ centered at $x_1, \ldots x_{N'}$ is a maximal packing, but it's not a covering if we chose radii r. Then there exists a point y that is away by the distance r from $x_1, \ldots x_{N'}$, which means that balls of radii $r/2$ centered at $N' + 1$ points $\{y, x_1, \ldots x_{N'}\}$ is also a packing. This contradicts maximality.

COROLLARY 16. *Let μ be a Borel probability measure valued in $\mathbb{R}^D$ whose support has the diameter of R, and suppose $D \geq 3$. Then we have:*

$$\mathbb{E}\left[W(\hat{\mu}_m, \mu) \right] \leq \frac{c_1}{m^{1/D}} \leq \frac{c_2}{m^{1/D}}$$

$$\text{where } c_1 = 32R \cdot \left(\frac{2}{D-2} \right)^{2/D}, \ c_2 = 51R$$

Also, if $D \geq 4$, we may take $c_2 = 32R$.

PROOF. Whenever $r < R/2$, the following holds:

$$N_{\text{cover}}(M, r) \leq N_{\text{cover}}((R/2) \cdot \mathcal{B}_D, r) = N_{\text{cover}}(\mathcal{B}_D, 2r/R) \leq \left(1 + \frac{2}{2r/R} \right)^D \leq (1.5R/r)^D$$

(The assumption $r < R/2$ is only used in the last inequality above) Thus we may apply the previous theorem by taking $\alpha = D$, $\beta = 1.5^D$, and $p = 1$, from which we get that:

$$\mathbb{E}\left[W(\hat{\mu}_m, \mu) \right] \leq 32R \cdot f(D/2) \cdot m^{-1/D}, \text{where } f(t) = (t-1)^{-1/t}$$

The derivative of $f(t)$ has the same sign as $(t-1)\log(t-1) - t$[9], which is an increasing function that takes a zero value at some $t \in (4.5, 5)$ and nowhere else. Thus $f(t)$ at $[1.5, \infty)$ is bounded above by $f(1.5) = 2^{2/3} \leq 1.6$ and the limit value of f at infinity, which is ≤ 1; we have $\lim_{t\to\infty}(t-1)^{1/t} \leq \lim_{t\to\infty} t^{1/t} = \exp(\lim_{t\to\infty}(\log t)/t) = 1$. Therefore, we get:

$$32R \cdot f(D/2) \cdot m^{-1/D} \leq 32R \cdot 2^{2/3} \cdot m^{-1/D} \leq 51R \cdot m^{-1/D}$$

Note also that $f(2) = 1$, so that $D \geq 4$ implies the tighter bound. $\qquad\square$

To obtain a concentration inequality, we use the Proposition A2 from [13]:

PROPOSITION 17. *Let (E, d, μ) be a measured Polish metric space of a finite diameter R and suppose that μ has a finite p-th moment. Then we have:*

$$\Pr\left(W_p(\hat{\mu}_m, \mu) \geq t + \mathbb{E}[W_p(\hat{\mu}_m, \mu)] \right) \leq \exp\left(-\frac{mt^{2p}}{2R^{2p}} \right)$$

Combining the above with Corollary 16, we obtain that:

PROPOSITION 18 (Global concentration). *Let μ be a Borel probability measure valued in $\mathbb{R}^D$ whose support has the diameter of R, and suppose $D \geq 3$. For any $t > 0$, the following holds whenever $m \geq f(t)$:*

$$\Pr\left(W(\hat{\mu}_m, \mu) \geq t \right) \leq \exp\left(-\frac{mt^2}{8R^2} \right)$$

where $f(t) = (102R/t)^D$.

---

[9]The derivative is $\frac{(t-1)\log(t-1)-t}{t^2 \cdot (t-1)^{1+1/t}}$.

PROOF. Follows directly by using Proposition 17 and letting the expected value of the Wasserstein distance be $\leq t/2$ in Corollary 16. □

We modify the above Proposition to study local behaviour of measures.

PROPOSITION 19 (Local concentration). *Let* $\mathbf{X} = (X_1, \ldots X_m)$ *be an i.i.d. sample drawn from* $\mu \in \mathcal{P}(\mathbb{R}^D)$, *where* $D \geq 3$. *Let* $\hat{\mu}_m = \frac{1}{m} \sum_i \delta_{X_i}$ *be the empirical measure constructed from* $\mathbf{X}$. *Let* $U \subseteq \mathbb{R}^D$ *be a Borel set which is contained in a ball of radius* r. *Denote* $u = \mu(U)$. *For any error level* $\epsilon > 0$, *the following holds whenever* $m \geq \max(N, 2u^{-1})$:

$$\Pr\left(W(\hat{\mu}_m|_U, \mu|_U) \geq t \right) \leq c \cdot m^N \gamma^m$$

where

$$c = \left(\frac{u}{1-u} \right)^N, \quad N = \lceil (204r/t)^D \rceil, \quad \gamma = 1 - u(1 - \exp(-t^2/8r^2))$$

In particular, $\gamma \in (0,1)$ *and the probability of error decays exponentially in* m.

PROOF. We condition over points of $\mathbf{X}$ falling into U. Denote by $\mathcal{S}_I$ the event $(X_i \in U \iff i \in I)$, $|I|$ for the cardinality of each I, and $u := \mu(U)$, we have

$$\Pr\left(W(\hat{\mu}_m|_U, \mu|_U) \geq t \right) = \sum_{I \subseteq \{1, \ldots m\}} \Pr(\mathcal{S}_I) \cdot \Pr\left(W(\hat{\mu}_m|_U, \mu|_U) \geq t \,\middle|\, \mathcal{S}_I \right)$$

$$= \sum_{I \subseteq \{1, \ldots m\}} u^{|I|} (1-u)^{m-|I|} \Pr\left(W(\hat{\mu}_m|_U, \mu|_U) \geq t \,\middle|\, \mathcal{S}_I \right)$$

$$= \sum_{k=0}^{m} \binom{m}{k} u^k (1-u)^{m-k} \Pr\left(W(\hat{\mu}_m|_U, \mu|_U) \geq t \,\middle|\, \mathcal{S}_{\{1, \ldots k\}} \right)$$

$$\text{(C.1)}$$

Now we apply Proposition 18 to the conditional probabilities above. This only applies to $k \geq N = \lceil (204r/t)^D \rceil$, thus we split the sum for $k < N$ and $k \geq N$. Writing $\xi =$

$\exp(-t^2/8r^2)$, Equation (C.1) is thus bounded by:

$$\leq \sum_{k=0}^{N-1} \binom{m}{k} u^k (1-u)^{m-k} + \sum_{k=N}^{m} \binom{m}{k} u^k (1-u)^{m-k} \xi^k$$

$$\leq (1-u)^m \cdot \sum_{k=0}^{N-1} \left(\frac{mu}{1-u}\right)^k + \sum_{k=0}^{m} \binom{m}{k} (u\xi)^k (1-u)^{m-k}$$

$$= (1-u)^m \frac{\left(\frac{mu}{1-u}\right)^N - 1}{\frac{mu}{1-u} - 1} + (1-u+u\xi)^m$$

$$\leq (1-u)^m \left(\left(\frac{mu}{1-u}\right)^N - 1\right) + (1-u+u\xi)^m$$

$$\leq \left(\frac{mu}{1-u}\right)^N \cdot (1-u+u\xi)^m$$

where in the second to last inequality we used the assumption $mu/(1-u) \geq 2$ and in the last inequality we used $(1-u)^m \leq (1-u+u\xi)^m$. $\qquad\square$

We further modify the above into a simultaneous concentration inequality, which is the main result of the section.

PROPOSITION 20 (Local simultaneous concentration). *Let* $\mathbf{X} = (X_1, \dots X_m)$ *be an i.i.d. sample drawn from* $\mu \in \mathcal{P}(\mathbb{R}^D)$, *where* $D \geq 3$. *Let* $\hat{\mu}_m = \frac{1}{m}\sum_i \delta_{X_i}$ *be the empirical measure constructed from* $\mathbf{X}$. *Also let* $r, t > 0$ *and* $U_i = \mathcal{B}(X_i, r)\backslash\{X_i\}$. *Then the following holds whenever* $m \geq \max(N, 2/u_-)$:

$$\Pr\left(\max_i W(\hat{\mu}_m|_{U_i}, \mu|_{U_i}) \leq t\right) \geq 1 - \delta_m$$

where $\lim_{m\to\infty} \delta_m = 0$ *exponentially fast, given explicitly as:*

$$\delta = c \cdot m^{N+1} \gamma^m$$

where

$$c = \left(\frac{u_+}{1-u_+}\right)^N, \quad N = \left\lceil\left(\frac{204r}{t}\right)^D\right\rceil, \quad \gamma = 1 - u_-(1-\xi), \quad \xi = \exp\left(\frac{-t^2}{8r^2}\right)$$

$$u_- = \inf_{x\in\text{supp}\,\mu} \mu(\mathcal{B}(x,r)), \quad u_+ = \sup_{x\in\text{supp}\,\mu} \mu(\mathcal{B}(x,r))$$

PROOF. We use union bound for different $i = 1, \dots m$. Let $\mu^m = \mu \times \cdots \times \mu$ be the product measure on $(\mathbb{R}^D) \times \cdots \times (\mathbb{R}^D)$. Define the set $E_i \subseteq (\mathbb{R}^D)^m$ as the set where the exception event occurs for U_i:

$$E_i = \left\{\mathbf{x} = (x_1, \dots x_m) \,\Big|\, W(\delta_\mathbf{x}|_{V_i}, \mu|_{V_i}) \geq t\right\}, \text{ where } V_i = \mathcal{B}(x_i, r)$$

Then we have:

$$\Pr\left(\max_i W(\hat{\mu}_m|_{U_i}, \mu|_{U_i}) \le t\right) = 1 - \mu^m(E_1 \cup \cdots \cup E_m)$$

We then apply the union bound:

$$\mu^m(E_1 \cup \cdots \cup E_m) \le \mu^m(E_1) + \cdots + \mu^m(E_m)$$

$$= m \cdot \int \mu^{m-1}\left((x_2, \ldots x_m) \Big| (x_1, x_2, \ldots x_m) \in E_1\right) dx_1$$

$$\le m \cdot \int \left(\frac{u_x}{1 - u_x}\right)^N (m-1)^N \gamma_x^{m-1} dx$$

where $u_x = \mu(\mathcal{B}(x, r))$ and $\gamma_x = 1 - u_x(1 - \xi)$. Then $u_x \le u_+$ and $\gamma_x \le \gamma$, so that we have:

$$\mu^m(E_1 \cup \cdots \cup E_m) \le c \cdot m^{N+1} \gamma^m$$

and the claim is shown. $\qquad\square$

Lipschitz continuity

LEMMA 21. *Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^D)$ and $\pi \in \mathrm{Gr}(k, D)$. Denoting $\mu_i' = \Pi(\mu_i, \pi)$, the following holds:*

$$W(\mu_1', \mu_2') \le W(\mu_1, \mu_2)$$

PROOF. For every transportation plan from μ_1 to μ_2, we can construct a less costly transportation plan from $\Pi(\mu_1, \pi)$ to $\Pi(\mu_2, \pi)$ by simply pushforwarding across projection.

Denote by $\mu_1^\perp = \Pi(\mu_1, \pi)$ and similarly $\mu_2^\perp$. Denote by p_π the orthogonal projection map to π. By definition we have $\mu_1^\perp(U) = \mu(p_\pi^{-1}U)$ for every open $U \subseteq \pi$ and similarly for $\mu_2^\perp$.

Given μ_{12}, a coupling of μ_1 and μ_2, we may define $\mu_{12}^\perp$ as the pushforward along $p_\pi \times p_\pi$, as follows:

$$\mu_{12}^\perp(U \times V) = \mu_{12}(p_\pi^{-1}U \times p_\pi^{-1}V)$$

for each open $U, V \subseteq \pi$. $\mu_{12}^\perp$ is a coupling of $\mu_1^\perp, \mu_2^\perp$ because:

$$\mu_{12}^\perp(U \times \pi) = \mu_{12}(p_\pi^{-1}U \times p_\pi^{-1}\pi) = \mu_{12}(p_\pi^{-1}U \times \mathbb{R}^D) = \mu_1(p_\pi^{-1}U) = \mu_1^\perp(U)$$

and similarly for μ_2. Now,

$$\int_{\pi \times \pi} \|x - y\| \, d\mu_{12}^\perp(x, y) = \int_{\mathbb{R}^D \times \mathbb{R}^D} \|p_\pi(x) - p_\pi(y)\| \, d\mu_{12}(x, y) \le \int_{\mathbb{R}^D \times \mathbb{R}^D} \|x - y\| \, d\mu_{12}(x, y)$$

where the first equality is due to the general fact that, for $f : X \to Y$,

$$\int_Y \phi(y) \, d f_*\mu(y) = \int_X \phi(f(x)) \, d\mu(x)$$

41

where in our case, $\phi(x,y) = \|x - y\|$, $f(x,y) = (p_\pi(x), p_\pi(y))$, $\mu = \mu_{12}$, and $f_*\mu = \mu_{12}^\perp$. □

LEMMA 22. *Let $\mu \in \mathcal{P}(\mathbb{R}^D)$ and $\pi_1, \pi_2 \in \mathrm{Gr}(k, D)$. Assume that the support of μ is bounded in the ball of radius 1, centered at the origin. Denoting $\mu_i = \Pi(\mu, \pi_i)$, we have:*

$$W(\mu_1, \mu_2) \leq \sqrt{\sin^2 \theta_1 + \cdots + \sin^2 \theta_d}$$

where $(\theta_1, \ldots \theta_d)$ are the principal angles between (π_1, π_2).

PROOF. Denote the orthogonal projection map to π_i by p_i. We define a coupling μ_{12} of (μ_1, μ_2):

$$\mu_{12}(U \times V) = \mu(p_1^{-1} U \cap p_2^{-1} V)$$

It is a coupling since $\mu(U \times \mathbb{R}^D) = \mu(p_1^{-1} U \cap \mathbb{R}^D) = \mu(p_1^{-1} U) = \mu_1(U)$ and similarly for μ_2.

For each x, consider the following sets:

$$S_x = p_2(p_1^{-1}(x) \cap \mathcal{B}_1), \quad S = \{(x,y) \mid x \in \pi_1, y \in S_x\} \subseteq \pi_1 \times \pi_2$$
$$S_x' = (\pi_2 \cap \mathcal{B}_1) \backslash S_x, \quad S' = \{(x,y) \mid x \in \pi_1, y \in S_x'\} \subseteq \pi_1 \times \pi_2$$

where $\mathcal{B}_1 = \mathcal{B}(0,1) \subseteq \mathbb{R}^D$ is the unit ball centered at origin. Also let $\theta = \angle(\pi_1, \pi_2)$. We claim that $\mu_{12}|_{S'} \equiv 0$ and $S_x \subseteq \mathcal{B}(x, \sin \theta)$. The proposition follows from these assumptions:

$$\int_{\pi_1 \times \pi_2} \|x - y\| \, d\,\mu_{12}(x,y) = \int_S \|x - y\| \, d\,\mu_{12}(x,y) \leq \int_S \sin \theta \, d\,\mu_{12}(x,y) = \sin \theta$$

It remains to prove the postponed claims. First we show $\mu_{12}|_{S'} \equiv 0$. Suppose $U \times V \subseteq S'$. By definition of S', for each $(x,y) \in U \times V$, we have $y \notin p_2(p_1^{-1}(x) \cap \mathcal{B}_1)$, i.e. $p_1^{-1}(x) \cap p_2^{-1}(y) \cap \mathcal{B}_1 = \emptyset$. Therefore we have $p_1^{-1}(U) \cap p_2^{-1}(V) \cap \mathcal{B}_1 = \emptyset$. Since the support of μ is in $\mathcal{B}_1$, we have that $\mu_{12}(U \times V) = \mu(p_1^{-1}(U) \cap p_2^{-1}(V)) = \mu(\emptyset) = 0$. Therefore $\mu_{12}|_{S'} \equiv 0$.

Now we show $S_x \subseteq \mathcal{B}(x, \sin \theta)$. Suppose $y \in S_x = p_2(p_1^{-1}(x) \cap \mathcal{B}_1)$, so that there is $z \in \mathcal{B}_1$ such that $p_1(z) = x, p_2(z) = y$. Let $d_0 = \dim(\pi_1 \cap \pi_2)$. Define $\pi_i^\perp \subseteq \pi_i$ to be the orthogonal complement of $\pi_1 \cap \pi_2$, so that $\pi_i = \pi_i^\perp + (\pi_1 \cap \pi_2)$ is an orthogonal decomposition for $i = 1, 2$. By Corollary 37, we obtain an orthonormal basis $\{u_1, \ldots u_{d_0}\} \cup \{v_1, w_1, \ldots v_{d-d_0}, w_{d-d_0}\}$ of $\mathrm{span}(\pi_1, \pi_2)$ so that:

$$\pi_1 \cap \pi_2 = \mathrm{span}(u_1, \ldots u_{d_0})$$
$$\pi_1^\perp = \mathrm{span}(v_1, \ldots v_{d-d_0})$$
$$\pi_2^\perp = \mathrm{span}(v_1', \ldots v_{d-d_0}')$$
$$\text{where } v_i' = (\cos \theta_i) v_i + (\sin \theta_i) w_i$$

with $(\theta_1, \ldots \theta_{d-d_0})$ being the nonzero principal angles of (π_1, π_2). We now attempt to understand (x,y) through their 2-dimensional projections. For each i, define $\rho_i = \mathrm{span}(v_i, w_i)$

and $z_i = \Pi(z, \rho_i)$. Then for any $u \in \rho_i$, we have that $\Pi(z, u) = \Pi(z_i, u)$ [10]. This gives an orthogonal decomposition:

$$x = \Pi(z, (\pi_1 \cap \pi_2) + \pi_1^{\perp}) = \Pi(z, \pi_1 \cap \pi_2) + \sum_{i=1}^{d-d_0} \Pi(z_i, v_i)$$

$$y = \Pi(z, (\pi_1 \cap \pi_2) + \pi_1^{\perp}) = \Pi(z, \pi_1 \cap \pi_2) + \sum_{i=1}^{d-d_0} \Pi(z_i, v_i')$$

Therefore,

$$\|x - y\|^2 = \sum_{i=1}^{d-d_0} \|\Pi(z_i, v_i) - \Pi(z_i, v_i')\|^2 = \sum_{i=1}^{d-d_0} (\sin^2 \theta_i) \cdot \|z_i\|^2 \leq \sum_{i=1}^{d-d_0} \sin^2 \theta_i$$

where the second equality follows from elementary Euclidean geometry on each 2-dimensional plane ρ_i. The claim thus follows (by padding the zero principal angles back in, for which $\sin 0 = 0$.) $\qquad \square$

PROPOSITION 23. *Let $M = M_1 \cup M_2 \subset \mathbb{R}^D$ be the union of two d-dimension submanifolds. Let $\tau = \min(\tau_1, \tau_2)$, where τ_i is the reach of the manifold M_i. Let $x \in M_1 \cap M_2$ and $y \in \mathbb{R}^D$. Let $r > 0$ be a number and let $s = \|y - x\|/r$. Then there exist constants $c_5, c_6 > 0$ depending only on d such that the following holds:*

$$\rho, s \leq c_5 \implies W(\mu_{x,r}, \mu_{y,r}) \leq c_6(\rho + s), \text{ where } \rho = r/\tau$$

PROOF. We construct a 3-step transportation plan for the bound. The first two steps redistribute masses, where in the first step the claim reduces to the case of a single manifold. The third step moves two parts of mass through translation and relocation. We begin by defining the following notations:

$$B_x = \mathcal{B}(x, r), \quad \alpha_{i,x} = \mathcal{H}^d(M_i \cap B_x)$$

Step 1. We redistribute mass equally for two manifolds. For $i = 1, 2$, define $\mu_{x,r}^{(i)}$ to be the normalised restriction of $\mu_{x,r}$ to $g_{x,r}(M_i)$, where $g_{x,r}(z) = (z - x)/r$. Then we see that:

$$\mu_{x,r} = \frac{\alpha_{1,x} \cdot \mu_{x,r}^{(1)} + \alpha_{2,x} \cdot \mu_{x,r}^{(2)}}{\alpha_{1,x} + \alpha_{2,x}}$$

Define the following:

$$\mu_{x,r}' = \frac{1}{2}\left(\mu_{x,r}^{(1)} + \mu_{y,r}^{(2)}\right)$$

---

[10]More generally, for any pair of subspaces $\pi' \subseteq \pi$, we have that $\Pi(\Pi(z, \pi), \pi') = \Pi(z, \pi')$, as it can be checked by directly writing down the projection matrices.

so that the masses are equally distributed on the two manifolds. We have:

$$W(\mu_{x,r}, \mu_{y,r}) \leq \left(W(\mu_{x,r}, \mu'_{x,r}) + W(\mu_{y,r}, \mu'_{y,r}) \right) + W(\mu'_{x,r}, \mu'_{y,r})$$

$$\leq \left(W(\mu_{x,r}, \mu'_{x,r}) + W(\mu_{y,r}, \mu'_{y,r}) \right) + \frac{1}{2} \sum_{i=1}^{2} W(\mu_{x,r}^{(i)}, \mu_{y,r}^{(i)})$$

where the first inequality follows from the triangle inequality and the second inequality is due to the fact that the support of μ lies on $M_1 \cup M_2$.

If $\alpha_{1,x} \geq \alpha_{2,x}$, then a transportation plan from $\mu_{x,r}$ to $\mu'_{x,r}$ can be constructed by moving $u \cdot \mu_{x,r}^{(1)}$ to the origin and back to $u \cdot \mu_{x,r}^{(2)}$, where $u = |\alpha_{1,x} - \alpha_{2,x}| / (2\alpha_{1,x} + 2\alpha_{2,x})$. Distance of masses moved by the transportation plan is at most 2. If $\alpha_{1,x} \leq \alpha_{2,x}$, there is a completely analogous transportation plan. The transportation cost is thus bounded by:

$$W(\mu_{x,r}, \mu'_{x,r}) \leq 2u = \frac{|\alpha_{1,x} - \alpha_{2,x}|}{\alpha_{1,x} + \alpha_{2,x}}$$

and similarly for $W(\mu_{y,r}, \mu'_{y,r})$.

Step 2. From the previous step, we are interested in bounding $W(\mu_{x,r}^{(i)}, \mu_{y,r}^{(i)})$ for $i = 1, 2$. Due to symmetry in consideration of M_1 and M_2, let us simply write $N = M_1$ and also:

$$\nu_x = \mu_{x,r}^{(1)}, \ \nu_y = \mu_{y,r}^{(1)}$$

We are interested in bounding $W(\nu_x, \nu_y) = W(\mu_{x,r}^{(1)}, \mu_{y,r}^{(1)})$. Define:

$$U_x = g_{x,r}(B_x \backslash B_y), \ V_x = g_{x,r}(B_x \cap B_y)$$

where $g_{x,r}(z) = (z - x)/r$. Since $g_{x,r}(B_x)$ is the unit ball of radius 1 at the origin of $\mathbb{R}^D$, we see that (U_x, V_x) partition that ball into two regions. We will use them to divide ν_x into two parts. Also define:

$$\beta_x = \mathcal{H}^d(N \cap B_x \backslash B_y), \ \gamma = \mathcal{H}^d(N \cap B_x \cap B_y), \ \alpha_x = \alpha_{1,x}$$

so that $\beta_x + \gamma = \alpha_x$. With this notation we have:

$$\nu_x = \frac{\beta_x \nu'_x + \gamma \nu''_x}{\beta_x + \gamma}, \text{ where } \nu'_x = \nu_x \|_{U_x}, \ \nu''_x = \nu_x \|_{V_x}$$

and similarly for ν_y. We'd like to compare the pairs (ν'_x, ν'_y) and (ν''_x, ν''_y) separately, but the ratios β_x / γ and β_y / γ are different. To match the ratios, define $\nu_y^\dagger$ as the linear combination of (ν'_y, ν''_y) that has the same ratio as that of (ν'_x, ν''_x) in ν_x:

$$\nu_y^\dagger = \frac{\beta_x \nu'_y + \gamma \nu''_y}{\beta_x + \gamma}$$

Now we construct a transportation plan from ν_y to $\nu_y^\dagger$; they are both linear combinations of ν_y' and ν_y'' with different ratios. If $\beta_x \geq \beta_y$, then we transport ν_y into $\nu_y^\dagger$ by moving $u_2 \cdot \nu_y''$ to the origin and then to $u_2 \cdot \nu_y'$, where $u_2 = \gamma \cdot |(\beta_y + \gamma)^{-1} - (\beta_x + \gamma)^{-1}|$. If $\beta_x \leq \beta_y$, we move $u_2 \cdot \nu_y'$ to the origin and then to $u_2 \cdot \nu_y''$. Distance of masses moved around in this process is at most 2. Therefore,

$$W(\nu_y, \nu_y^\dagger) \leq 2 \cdot u_2 = 2\gamma \cdot \left| \frac{1}{\beta_y + \gamma} - \frac{1}{\beta_x + \gamma} \right| = 2\gamma \cdot \left| \frac{1}{\alpha_y} - \frac{1}{\alpha_x} \right| \leq 2 \cdot \left| \frac{\alpha_y}{\alpha_x} - 1 \right|$$

Therefore,

$$W(\nu_x, \nu_y) \leq W(\nu_x, \nu_y^\dagger) + W(\nu_y^\dagger, \nu_y)$$
$$= W\left(\frac{\beta_x \nu_x' + \gamma \nu_x''}{\beta_x + \gamma}, \frac{\beta_x \nu_y' + \gamma \nu_y''}{\beta_x + \gamma} \right) + W(\nu_y^\dagger, \nu_y)$$
$$\leq \frac{\beta_x W(\nu_x', \nu_y') + \gamma W(\nu_x'', \nu_y'')}{\beta_x + \gamma} + 2 \cdot \left| \frac{\alpha_y}{\alpha_x} - 1 \right|$$

Step 3. At this point our task is reduced to bounding both $W(\nu_x', \nu_y')$ and $W(\nu_x'', \nu_y'')$. We simply relocate all mass of ν_x' to the origin and bring it back to ν_y', so that we use the trivial bound $W(\nu_x', \nu_y') \leq 2$. To bound $W(\nu_x'', \nu_y'')$, we observe that:

$$\nu_x'' = g(\mathcal{H}^d) \|_{g(M_1) \cap g(B_x \cap B_y)} = g(\mathcal{H}^d \|_{M_1 \cap B_x \cap B_y})$$

where $g = g_{x,r}$. Therefore,

$$W(\nu_x'', \nu_y'') = W\left(g_{x,r}(\mathcal{H}^d \|_{M_1 \cap B_x \cap B_y}), g_{y,r}(\mathcal{H}^d \|_{M_1 \cap B_x \cap B_y}) \right)$$

Since for any w, $g_{y,r}(w) - g_{x,r}(w) = (x - y)/r$, we obtain ν_y'' from ν_x'' by pushforwarding the measure by translation through $(x - y)/r$. Thus $W(\nu_x'', \nu_y'') \leq s := \|x - y\|/r$. Therefore,

$$\frac{\beta_x W(\nu_x', \nu_y') + \gamma W(\nu_x'', \nu_y'')}{\beta_x + \gamma} \leq \frac{\beta_x \cdot 2 + \gamma \cdot s}{\beta_x + \gamma} \leq \frac{2\beta_x}{\alpha_x} + s$$

This thus shows that:

$$W(\nu_x, \nu_y) \leq \frac{2\beta_x}{\alpha_x} + \frac{2|\alpha_y - \alpha_x|}{\alpha_x} + s$$

Total bound. We now collect the terms from above and bound them using s and r. The main tool here is Corollary 33, which gives bounds for the volume of a manifold cut out by a ball. To apply it to balls of radii r centered at x and y, we assume that $r/\tau < c_3$ and

$s < 1$. Collecting the terms from above, we get the following inequality:

$$W(\mu_{x,r}, \mu_{y,r}) \leq E_1 + E_2 + E_3$$

$$E_1 = \frac{|\alpha_{1,x} - \alpha_{2,x}|}{\alpha_{1,x} + \alpha_{2,x}} + \frac{|\alpha_{1,y} - \alpha_{2,y}|}{\alpha_{1,y} + \alpha_{2,y}}$$

$$E_2 = \frac{|\alpha_{1,y} - \alpha_{1,x}|}{\alpha_{1,x}} + \frac{|\alpha_{2,y} - \alpha_{2,x}|}{\alpha_{2,x}}$$

$$E_3 = \frac{\beta_{1,x}}{\alpha_{1,x}} + \frac{\beta_{2,x}}{\alpha_{2,x}} + s$$

where $\beta_{i,x} = \mathcal{H}^d(M_i \cap B_x \backslash B_y)$. Corollary 33 implies that for $i = 1, 2$, we have:

$$\frac{\alpha_{i,x}}{\omega_d r^d} \in \left[1 - c_4\rho, 1 + c_4\rho\right], \frac{\alpha_{i,y}}{\omega_d r^d} \in \left[1 - c_4(\rho + s), 1 + c_4(\rho + s)\right]$$

where $\rho = r/\tau$ [11]. Here we used the fact that $d(y, M_i) \leq \|x - y\| \leq s \cdot r$. To work with β_x, we use the triangle inequality to see that $\mathcal{B}(x, r - sr) \subseteq \mathcal{B}(y, r) = B_y$, and thus:

$$\beta_{i,x} = \mathcal{H}^d(M_i \cap B_x \backslash B_y) \leq \mathcal{H}^d(M_i \cap B_x \backslash \mathcal{B}(x, r - sr))$$

Thus Corollary 33 again implies:

$$\beta_{i,x} \leq \omega_d r^d(1 + c_4\rho) - \omega_d(r - sr)^d(1 - c_4(r - sr)/\tau)$$

$$\leq \omega_d r^d\left(1 + c_4\rho - (1 - s)^d(1 - c_4\rho)\right)$$

$$\leq \omega_d r^d\left(1 - (1 - s)^d + 2c_4\rho\right)$$

$$\leq \omega_d r^d(d \cdot s + 2c_4\rho)$$

where in the last inequality we used the fact that $(1 - t)^d \geq 1 - d \cdot t$ for $t \in [0, 1]$ and $d \geq 1$ [12] Therefore, we see that:

$$E_1 \leq \frac{c_4\rho}{1 - c_4\rho} + \frac{c_4(\rho + s)}{1 - c_4(\rho + s)}, \quad E_2 \leq 2 \cdot \frac{c_4(2\rho + s)}{1 - c_4\rho}, \quad E_3 \leq 2 \cdot \frac{d \cdot s + 2c_4\rho}{1 - c_4\rho} + s$$

---

[11] Defining $\tau = \min(\tau_1, \tau_2)$, where τ_i is the reach of M_i, makes these bounds work simultaneously for a single value of τ.

[12] This is because the second derivative of $(1 - t)^d$ is non-negative.

Therefore, assuming that $\rho + s \le 1/(2c_4)$ and thus $1 - c_4(\rho + s) \ge 1/2$, we produce the following linear bound:

$$\begin{aligned}
&W(\mu_{x,r}, \mu_{y,r}) \\
&\le E_1 + E_2 + E_3 \\
&\le \frac{c_4(2\rho + s) + 2c_4(2\rho + s) + 2d \cdot s + 4c_4\rho}{1 - c_4(\rho + s)} + s \\
&\le \frac{10c_4\rho + (3c_4 + 2d)s}{1 - c_4(\rho + s)} + s \\
&\le 20c_4\rho + (6c_4 + 4d + 1)s
\end{aligned}$$

Therefore we may set $c_5 = \min(\frac{1}{4c_4}, c_3, 1)$ and $c_6 = \max(20c_4, 6c_4 + 4d + 1)$ to obtain our claim. $\qquad \square$

The following is a simplified version of the main result in Section 4 of [32]:

PROPOSITION 24. *Let μ be the uniform distribution over $M \subset \mathbb{R}^D$, which is a d-dimensional submanifold. Then for every $x \in M$, $\mu_{x,0}$ is the uniform distribution over[13] $T_x M \cap \mathcal{B}(0,1)$. Furthermore, we have the following bound:*

$$r/\tau \le c_7 \implies W(\mu_{x,r}, \mu_{x,0}) \le c_8 r/\tau$$

Using the above and repeating the mass-redistribution argument in Step 1 of the proof in Proposition 23 verbatim, we obtain the following:

PROPOSITION 25. *Let μ be the uniform distribution over $M = M_1 \cup M_2 \subset \mathbb{R}^D$, which is a union of two d-dimensional submanifolds. Then for every $x \in M_1 \cap M_2$, $\mu_{x,0}$ is the uniform distribution over $(T_x M_1 \cup T_x M_2) \cap \mathcal{B}(0,1)$. Furthermore, we have the following bound:*

$$r/\tau \le c_9 \implies W(\mu_{x,r}, \mu_{x,0}) \le c_{10} r/\tau$$

We prove a result that allows us to work with one manifold at a time when dealing with the smooth region of $M_1 \cup M_2 \cup \mathbb{R}^D$, where each M_i is a submanifold.

PROPOSITION 26. *Let $M = M_1 \cup M_2 \subset \mathbb{R}^D$ be a union of two d-dimensional submanifolds, such that $M_1 \cap M_2$ is nonempty. Suppose that for every $x \in M_1 \cap M_2$, we have $\dim(T_x M_1 \cap T_x M_2) = d_0$ for a fixed d_0, and that principal angles of $(T_x M_1, T_x M_2)$ are all $\ge \phi$. Then we have:*

$$h(M_1, M_2) = \inf_{x \in M_1} \frac{d(x, M_2)}{d(x, M_1 \cap M_2)} \in (0, 1]$$

In particular, for any $r > 0$ and $x \in M_1$, we have:

$$x \notin \mathcal{B}(M_1 \cap M_2, h^{-1} \cdot r) \implies \mathcal{B}(x, r) \cap M_2 = \emptyset$$

---

[13]To be precise, it is the intersection $T'_x M \cap \mathcal{B}(0,1)$, where $T'_x M = T_x M - x$ is the linear subspace of $\mathbb{R}^D$ obtained by translating $T_x M$ by $(-x)$. The $\mathcal{B}(0,1)$ here refers to the unit ball centred at the origin in $\mathbb{R}^D$.

where $h = h(M_1, M_2)$.

PROOF. Firstly $h \leq 1$ holds trivially since $d(x, M_2) \leq d(x, M_1 \cap M_2)$. Let $r > 0$ be a number satisfying:

$$\frac{r}{\tau} < \min\left(\frac{\sqrt{2}-1}{2}, \frac{1-\cos\phi}{12}\right) \tag{C.2}$$

where $\tau = \min(\tau_1, \tau_2)$ and τ_i is the reach of M_i. The angle condition involving $\cos\phi$ will be used in the final steps of the proof. Since M_1 partitions into the disjoint union of $M_1 \cap B$ and $M_1 \backslash B$ where $B = \mathcal{B}(M_1 \cap M_2, r)$, we may write:

$$h = \min\left(\inf_{x \in M_1 \cap B} \frac{d(x, M_2)}{d(x, M_1 \cap M_2)}, \quad \inf_{x \in M_1 \backslash B} \frac{d(x, M_2)}{d(x, M_1 \cap M_2)}\right)$$

The second term is easily seen to be positive:

$$\inf_{x \in M_1 \backslash B} \frac{d(x, M_2)}{d(x, M_1 \cap M_2)} \geq \frac{\inf_{x \in M_1 \backslash B} d(x, M_2)}{\sup_{x \in M_1 \backslash B} d(x, M_1 \cap M_2)} > 0$$

where the numerator is positive since $M_1 \backslash B$ is a compact set and $x \mapsto d(x, M_1)$ is positive and continuous, and the denominator is finite since M_1, M_2 are bounded.

Reduction to linear algebra. We now examine the fraction $d(x, M_2) / d(x, M_1 \cap M_2)$ when $x \in M_1 \cap B$. Denote $r_0 = d(x, M_1 \cap M_2)$. By compactness of $M_1 \cap M_2$, we see that there is a point $x_0 \in M_1 \cap M_2$ such that $r_0 = d(x, x_0)$. Also denote $\pi_i = T_{x_0} M_i$ and:

$$x_1 = \Pi(x, \pi_1), \quad x_2 = \Pi(x_1, M_2)$$

Then we apply Lemma 28 due Equation C.2 with $\|x - x_0\| = r_0 < (\sqrt{2}-1)\tau$ and [14] $\|x_2 - x_0\| \leq 2r_0 < (\sqrt{2}-1)\tau$,

$$d(x, \pi_1) \leq r_0^2/\tau, \quad d(x_2, \pi_2) \leq (2r_0)^2/\tau$$

Therefore:

$$\begin{aligned} d(x, M_2) &\geq d(x_1, M_2) - d(x, x_1) \\ &= d(x_1, x_2) - d(x, \pi_1) \\ &\geq d(x_1, \pi_2) - d(x_2, \pi_2) - d(x, \pi_1) \\ &\geq d(x_1, \pi_2) - 5\tau^{-1} r_0^2 \end{aligned} \tag{C.3}$$

Linear algebra. Now we are interested in controlling $d(x_1, \pi_2)$, and this is an exercise of linear algebra since $x_1 \in \pi_1$. Write $x_1^{\perp} = \Pi(x_1, \pi_2)$ and $(z, z_\perp) = (x_1 - x_0, x_1^{\perp} - x_0)$,

---

[14]In detail: $\|x_1 - x_2\| = \inf_{y \in M_2} \|x_1 - y\| \leq \|x_1 - x_0\| \leq \|x - x_0\| = r_0$ and thus $\|x_2 - x_0\| \leq \|x_2 - x_1\| + \|x_1 - x_0\| \leq 2r_0$.

so that $d(x_1, \pi_2) = d(x_1, x_1^\perp) = d(z, z_\perp)$. Let $\pi_i' = \pi_i - x_0$ be a vector space (satisfying $0 \in \pi_i'$), and let $A_i \in \mathbb{R}^{D \times d}$ be a matrix whose columns are an orthonormal basis of π_i'. Then $z \in \pi_1'$ and $z_\perp = \Pi(z, \pi_2') = A_2 A_2^\top z$ and

$$d(x_1, \pi_2) = d(z, z_\perp)^2 = \|z\|^2 - \|z_\perp\|^2 \tag{C.4}$$

by Pythagoras' theorem. Now for some $u \in \mathbb{R}^d$ we may write $z = A_1 u$, so that $\|z_\perp\| = \|A_2 A_2^\top z\| = \|A_2 A_2^\top A_1 u\|$. Using the fact that the map $w \mapsto A_i w$ is distance-preserving and the assumption that principal angles between tangent spaces (π_1, π_2) are $\geq \phi$, we get:

$$\|z_\perp\| = \|A_2 A_2^\top A_1 u\| = \|A_2^\top A_1 u\| \leq \|A_2^\top A_1\| \cdot \|u\| = \|A_2^\top A_1\| \cdot \|z\| \leq (\cos\phi) \cdot \|z\| \tag{C.5}$$

and we also note that:

$$\|z\| = d(x_1, x_0) \geq d(x, x_0) - d(x, x_1) = d(x, x_0) - d(x, \pi_1) \geq r_0 - r_0^2/\tau \tag{C.6}$$

Combining the bound. We plug Equations (C.4), (C.5), (C.6) into Equation (C.3):

$$
\begin{aligned}
d(x, M_2) &\geq d(x_1, \pi_2) - 5r_0^2/\tau \\
&\geq \sqrt{1 - \cos^2\phi} \cdot \|z\| - 5r_0^2/\tau \\
&\geq (1 - \cos\phi)(r_0 - r_0^2/\tau) - 5r_0^2/\tau \\
&\geq r_0 - \left((\cos\phi)r_0 + 6r_0^2/\tau\right) \\
&\geq \frac{1 - \cos\phi}{2} \cdot r_0 > 0
\end{aligned}
$$

where in the last equality, we used the assumption $r_0/\tau \leq (1 - \cos\phi)/12$ in Equation (C.2). Therefore, recalling that $r_0 = d(x, M_1 \cap M_2)$, the following holds for all $x \in M_1 \backslash B$:

$$\frac{d(x, M_2)}{d(x, M_1 \cap M_2)} \geq \frac{1 - \cos\phi}{2} > 0$$

and the claim is proven. $\qquad\square$

The MMD for the Gaussian kernel can be controlled with the Wasserstein distance as follows:

LEMMA 27. *Let $\kappa(x, y) = e^{-\gamma\|x-y\|^2}$ be a Gaussian kernel. Then whenever $\|f\|_\kappa \leq 1$, f is a $\sqrt{2\gamma}$-Lipschitz function. Therefore, for any probability measures μ, ν valued in $\mathbb{R}^D$ with finite first moment, we have:*

$$\Delta_\kappa(\mu, \nu) \leq \sqrt{2\gamma} \cdot W(\mu, \nu)$$

PROOF.

$$\begin{aligned}
|f(x) - f(y)| &= |\langle f, \kappa(x,-) - \kappa(y,-)\rangle| \\
&\le \|f\|_\kappa \cdot \|\kappa(x,-) - \kappa(y,-)\|_\kappa \\
&= \|f\|_\kappa \cdot \sqrt{\kappa(x,x) + \kappa(y,y) - 2\kappa(x,y)} \\
&= \sqrt{2}\|f\|_\kappa \cdot \sqrt{1 - e^{-\gamma\|x-y\|^2}} \\
&\le \sqrt{2\gamma}\|f\|_\kappa \cdot \|x - y\|
\end{aligned}$$

where in the last inequality we used $\sqrt{1 - e^{-s^2}} \le s$. Therefore any function with $\|f\|_\kappa \le 1$ is also a $\sqrt{2\gamma}$-Lipschitz function. The conclusion follows by the integral probability metric definition. $\square$

C.2. Measures on Manifolds. We use the following lemma from [32], which is a simple extension of Proposition 6.3 of [38]. It controls the deviation of geodesic from the first order approximation:

LEMMA 28. *Let M be a smooth compact n-manifold embedded in $\mathbb{R}^D$ with reach τ. Suppose that x, y are connected by a (unit speed) geodesic $\gamma : [0, \tilde{r}] \to M$ of length $\tilde{r}$ with $\gamma(0) = x, \gamma(\tilde{r}) = y$, and denote $r = \|x - y\|$. Then the following inequalities hold:*

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

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

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

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

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

LEMMA 29. *Let $M \subset \mathbb{R}^D$ be a compact set and let τ be its reach. Let π_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 $0 < r < \tau$, $\pi_M|_{\mathcal{B}(M,r)}$ is a single-valued continuous function.

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

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

Since $d(y, M) = \|y - \pi(y)\|$ for each $y \in \mathcal{B}(M, r)$, 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 $\inf\{\|x - y\| \mid y \in M \backslash \mathcal{B}(\pi(x), s)\} = \|x - \pi(x)\| = d(x, M)$.

This is a contradiction. Since $M \backslash \mathcal{B}(\pi(x), s)$ is a compact set, the distance function $y \mapsto \|y - x\|$ attains a minimum on some $z \in M \backslash \mathcal{B}(\pi(x), s)$. 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. $\qquad \square$

LEMMA 30. *Let $M \subset \mathbb{R}^D$ be a compact path-connected set and let τ 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 τ 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:

$$\|z - x\| \le \|z - \bar{z}\| + \|\bar{z} - x\| \le \|y - \bar{z}\| + \|\bar{z} - x\| = \|y - x\|$$

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

PROPOSITION 31. *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 $\tilde{\pi}_{x,r} := \pi_x|_{M \cap \mathcal{B}(x,r)}$. 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 $J_y \tilde{\pi}_x = \det(A_x^\top A_y)$, 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:*

$$\cos \theta_{x,y} \ge 1 - \frac{d_M(x, y)}{\tau}$$

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

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

(5) *Suppose that $r < (\sqrt{2} - 1)\tau$. We have*

$$\frac{\mathcal{H}^d(M \cap \mathcal{B}(x, r))}{\omega_d r^d} \in \left[(1 - \rho^2/4)^{d/2}, (1 - \rho - \rho^2)^{-d} \right]$$

where $\rho = r/\tau$.

PROOF. (1) The nonsingularity is Lemma 5.4 from [38]. 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 $d\,\tilde{\pi}_x(v) = A_x^\top v$ for each (embedded) tangent vector $v \in T_y M$.

(3) This is Proposition 6.2 from [38].

(4) This folllows from (2) and the definition of principal angles.

(5) The lower bound is Lemma 5.3 from [38]. To see the upper bound, we note by the Area Formula of geometric measure theory that the volume $\mathcal{H}^d(M \cap \mathcal{B}(x,r))$ is the integral of Jacobian of inverse-projection in $\pi(M \cap \mathcal{B}(x,r))$, i.e.

$$\mathcal{H}^d(M \cap \mathcal{B}(x,r)) = \int_{\pi(M \cap \mathcal{B}(x,r))} (J_{z'}\tilde{\pi}_x)^{-1}\, d\,z$$

where $z' \in M \cap \mathcal{B}(x,r)$ is the unique point such that $\pi_x(z') = z$. Now note that $\pi(M \cap \mathcal{B}(x,r))$ is contained in a ball of radius r in $T_x M$, so that its measure is at most $\omega_d r^d$. Furthermore, the inverse of Jacobian in the integrand is at most $(1 - d_M(x,y)/\tau)^{-d}$ by (3) and (4). By the bound on geodesic length (28), we have $d_M(x,y)/\tau \leq \rho + \rho^2$ and thus obtain the claim. $\qquad\square$

COROLLARY 32. *Let $M \subset \mathbb{R}^D$ be a d-dimensional submanifold. There exist constants $c_1 > 0, c_2 \geq 0$ depending only on d such that the following hold.*

$$r < c_1\tau \implies \frac{\mathcal{H}^d(M \cap \mathcal{B}(x,r))}{\omega_d r^d} \in \left[1 - \frac{c_2 r}{\tau}, 1 + \frac{c_2 r}{\tau}\right]$$

PROOF. Let's first assume that $r < (\sqrt{2} - 1)\tau$. Then we can relax the upper bound of (5) of the previous Proposition into $(1 - \sqrt{2}\rho)^{-d}$. Now we further relax our lower and upper bounds, which are given by:

$$f_1(t) = (1 - t^2/4)^{d/2}, \quad f_2(t) = (1 - \sqrt{2}t)^{-d}$$

Their second derivatives are given by:

$$f_1''(t) = d \cdot (1 - t^2/4)^{d/2} \cdot \frac{(d-1)t - 4}{(t^2 - 4)^2}, \quad f_2''(t) = 2d \cdot (d+1) \cdot (1 - \sqrt{2}t)^{-d-2}$$

Then we see that $f_1''(t) \leq 0$ for $t \in [0, 4/(d-1)]$ and $f_2''(t) \geq 0$ for $t \in [0, 1/\sqrt{2}]$.

Therefore, if we let $c_1 = \min(\sqrt{2} - 1, 4/(d-1))$, then we see that $1 - c_2 t \leq f_1(t)$ and $1 + c_2 t \geq f_2(t)$, where $c_2 \geq 0$ is given by:

$$c_2 = \max\left(1 - (3/4)^{d/2}, (\sqrt{2} - 1)^{-d}\right)$$

which are values obtained from slopes of $f_1(t), f_2(t)$ by plugging in $t = 1$ and $t = \sqrt{2} - 1$ respectively. $\qquad\square$

COROLLARY 33. *Let $M \subset \mathbb{R}^D$ be a d-dimensional submanifold, and let $r > 0$. Suppose $x \in \mathbb{R}^D$ is a point satisfying $\mathrm{d}(x, M) = s \cdot r$. There exists constants $c_3, c_4 \geq 0$ depending only on d such that the following holds.*

$$r < c_3\tau, \ s < 1 \implies \frac{\mathcal{H}^d(M \cap \mathcal{B}(x, r))}{\omega_d r^d} \in \left[1 - c_4(s + r/\tau), 1 + c_4(s + r/\tau)\right]$$

PROOF. We start by assuming that $s < 1$ and $r < \tau$, so that there is a unique point of projection $y \in M$ minimising distance from x, so that $\|x - y\| = sr$. By the triangle inequality, we have the inclusions:

$$\mathcal{B}(y, r - sr) \subseteq \mathcal{B}(x, r) \subseteq \mathcal{B}(y, r + sr)$$

which implies:

$$\mathcal{H}^d(M \cap \mathcal{B}(y, r - sr)) \leq \mathcal{H}^d(M \cap \mathcal{B}(x, r)) \leq \mathcal{H}^d(M \cap \mathcal{B}(y, r + sr))$$

Applying the previous Corollary, we get the lower bound:

$$\mathcal{H}^d(M \cap \mathcal{B}(y, r - sr)) \geq (1 - c_2(r - sr)/\tau) \cdot \omega_d(r - sr)^d \geq \omega_d r^d \cdot (1 - c_2 r/\tau)(1 - s)^d$$

and the upper bound:

$$\mathcal{H}^d(M \cap \mathcal{B}(y, r + sr)) \leq (1 + c_2(r + sr)/\tau) \cdot \omega_d(r + sr)^d \leq \omega_d r^d \cdot (1 + 2c_2 r/\tau)(1 + s)^d$$

where we are assuming that $r < (c_1/2)\tau$, so that $r - sr \leq r + sr \leq c_1\tau$ and the previous Corollary applies. Assuming $d \geq 1$ and $t \in [0, 1]$, the functions $t \mapsto (1 - t)^d$ and $t \mapsto (1 + t)^d$ both have non-negative second derivative, so that we have $(1 - t)^d \geq 1 - d \cdot t$ and $(1 + t)^d \leq 1 + 2^d \cdot t$. Letting $c_4' = \max(2^d, 2c_2)$, we get:

$$\frac{\mathcal{H}^d(M \cap \mathcal{B}(x, r))}{\omega_d r^d} \in \left[(1 - c_4's)(1 - c_4' r/\tau), (1 + c_4's)(1 + c_4' r/\tau)\right]$$

Expanding the brackets, we get:

$$(1 + c_4's)(1 + c_4' r/\tau) = 1 + c_4's + c_4' r/\tau + c_4' sr/\tau \leq 1 + c_4's + c_4' \cdot 2r/\tau \leq 1 + 2c_4'(s + r/\tau)$$

and similarly $(1 - c_4's)(1 - c_4' r/\tau) \geq 1 - 2c_4'(s + r/\tau)$. We thus obtain the claim by setting $c_3 = c_1/2$ and $c_4 = 2c_4'$. $\qquad\square$

C.3. Principal angle. 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 34. Given $\pi_1, \pi_2 \in \mathrm{Gr}(d, D)$, let $A_i \in \mathbb{R}^{D \times d}$ be a matrix with orthonormal columns that span π_i. Denote by $\measuredangle(\pi_1, \pi_2) \in [0, 1]^d$ the singular values of the matrix $A_1^\top A_2$, arranged in the descending order. The *principal angles* of (π_1, π_2) are defined as

the angles $(\theta_1, \ldots \theta_d) \in [0, \pi/2]^d$ such that $(\cos \theta_1, \ldots \cos \theta_d) = \measuredangle(\pi_1, \pi_2)$, which satisfy $\theta_1 \leq \cdots \leq \theta_d$.

We note in particular that $\theta_1 = \cdots = \theta_{d_0} = 0 < \theta_{d_0+1}$, where $d_0 = \dim(\pi_1) = \dim(\pi_2)$. The largest principal angle has a simple interpretation:

LEMMA 35. *If $\measuredangle(\pi_1, \pi_2) = (\cos \theta_1, \ldots \cos \theta_d)$ for $\pi_1, \pi_2 \in \mathrm{Gr}(d, D)$, then:*

$$\theta_d = \max_{x \in \pi_1} \min_{y \in \pi_2} \measuredangle(x, y) = \mathrm{d}_H(\pi_1 \cap \mathbb{S}, \pi_2 \cap \mathbb{S})$$

Here $\measuredangle(x, y) = \cos^{-1}(\langle x, y \rangle / (\|x\| \cdot \|y\|))$, $\mathrm{d}_H(A, B) = \inf\{r \mid \mathcal{B}(A, r) \supseteq B, \mathcal{B}(B, r) \supseteq A\}$ is the Hausdorff distance between two sets A, B, and $\mathbb{S}$ is the unit $(D - 1)$-dimensional sphere.

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

$$\cos \theta_D = \min_{\|z\|=1} \|A_1^\top A_2 z\| = \min_{\|y\|=1, y \in \Pi_2} \|A_1^\top y\| = \min_{\|y\|=1, y \in \Pi_2} \langle y_1, y \rangle$$

where y_1 is the unit vector in the direction of $A_1 A_1^\top y$. Noting that $\langle y_1, y \rangle = \max_{\|x\|=1, x \in \pi_1} \langle x, y \rangle$, we have $\cos \theta_D = \min_{\|y\|=1, y \in \pi_2} \max_{\|x\|=1, x \in \pi_1} \langle x, y \rangle$. $\square$

Principal angles characterise pairs of subspaces up to rotation.

PROPOSITION 36. *$\measuredangle$ induces the following bijection:*

$$\measuredangle : \frac{\mathrm{Gr}(d, D) \times \mathrm{Gr}(d, D)}{O(D)} \longrightarrow S(d, \max(0, 2d - D))$$

where $S(k, j) = \{(t_1, \ldots t_k) \mid 1 \geq t_1 \geq \cdots \geq t_k \geq 0, t_1 = \cdots = t_j = 1\}$, which is a set homeomorphic to the standard $(k - j)$-simplex.

Explicitly, we have the following. If $(\pi_1, \pi_2), (\pi_1', \pi_2') \in \mathrm{Gr}(d, D) \times \mathrm{Gr}(d, D)$ satisfy $\measuredangle(\pi_1, \pi_2) = \measuredangle(\pi_1', \pi_2')$, then there exists an element $A \in O(D)$ such that $(A\pi_1, A\pi_2) = (\pi_1', \pi_2')$. Furthermore, if $(t_1, \ldots t_d) \in [0, 1]^d$ satisfies $t_1 \geq \cdots \geq t_d$ and $t_1 = \cdots = t_j = 1$ with $j = d - \max(0, 2d - D)$, then there exists $(\pi_1, \pi_2) \in \mathrm{Gr}(d, D) \times \mathrm{Gr}(d, D)$ such that $\measuredangle(\pi_1, \pi_2) = (t_1, \ldots t_d)$.

PROOF. We prove the explicit version. Suppose that $(\pi_1, \pi_2), (\pi_1', \pi_2') \in \mathrm{Gr}(d, D) \times \mathrm{Gr}(d, D)$ with $\measuredangle(\pi_1, \pi_2) = \measuredangle(\pi_1', \pi_2')$. Let $A_i \in \mathbb{R}^{D \times d}$ be a matrix with orthonormal columns spanning π_i, and similarly define A_i'. Without loss of generality, we may assume that $A_1 = A_1' = J$, since by Gram-Schmidt there are matrices $H, H' \in O(D)$ such that $HA_1 = H'A_1' = J$, where $J = [I_d, 0_{D-d,d}] \in \mathbb{R}^{D \times d}$ has 1 on the diagonal and zero elsewhere. Let's relabel $B = A_2, B' = A_2'$. Also write $B^\top = [B_1^\top, B_2^\top]$ and $(B')^\top = [(B_1')^\top, (B_2')^\top]$, where B_1, B_1' are both $(d \times d)$-matrices.

Since $\angle(\pi_1, \pi_2) = \angle(\pi_1', \pi_2')$, the singular values of $(d \times d)$-matrices $J^\top B = B_1$ and $J^\top B' = B_1'$ are equal. Therefore there exist $U, V \in O(d)$ such that $B_1' = UB_1V^\top$. Then:

$$\begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} V^\top = \begin{bmatrix} UB_1V^\top \\ B_2V^\top \end{bmatrix} = \begin{bmatrix} B_1' \\ B_3 \end{bmatrix}, \text{ where } B_3 = B_2 V^\top$$

The right hand side also has orthonormal columns, so that we have $(B_1')^\top B_1' + B_3^\top B_3 = I_d$. Since B' also have orthonormal columns, we also have $(B_1')^\top B_1' + (B_2')^\top B_2' = I_d$. Therefore, $B_3^\top B_3 = (B_2')^\top B_2'$. This guarantees the existence of $W \in O(D - d)$ such that $WB_3 = B_2'$. Therefore, for $Z = \begin{bmatrix} U & 0 \\ 0 & W \end{bmatrix}$, we have:

$$ZBV^\top = \begin{bmatrix} U & 0 \\ 0 & W \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} V^\top = \begin{bmatrix} UB_1V^\top \\ WB_2V^\top \end{bmatrix} = \begin{bmatrix} B_1' \\ WB_3 \end{bmatrix} = \begin{bmatrix} B_1' \\ B_2' \end{bmatrix} = B'$$

Therefore $Z\pi_2 = \pi_2'$. The block diagonal form of Z also ensures that Z leaves $\pi_1 = \pi_1' = \mathbb{R}^k$ invariant. Therefore, we have $(Z\pi_1, Z\pi_2) = (\pi_1', \pi_2')$ as desired. $\square$

COROLLARY 37. *Given $\pi_1, \pi_2 \in \mathrm{Gr}(d, D)$, suppose that $A_i \in \mathbb{R}^{D \times d}$ has columns forming an orthonormal basis of π_i. Let $d_0 = \dim(\pi_1 \cap \pi_2) \geq 2d - D$ and let $d_1 = D - 2d + d_0$. Then there exist matrices $U \in O(D)$ and $V_1, V_2 \in O(d)$ such that $UA_1V_1 = \tilde{A}_1$ and $UA_2V_2 = \tilde{A}_2$, where*

$$\tilde{A}_1 = \begin{bmatrix} I_d \\ 0 \end{bmatrix}, \quad \tilde{A}_2 = \begin{bmatrix} I_{d_0} & 0 \\ 0 & \cos\Theta \\ 0 & \sin\Theta \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{D \times d}$$

where $\Theta = \mathrm{diag}(\theta_{d_0+1}, \ldots \theta_d) \in \mathbb{R}^{(d-d_0) \times (d-d_0)}$ is the diagonal matrix of nonzero principal angles $\angle(\pi_1, \pi_2) = (\theta_1, \ldots \theta_d)$.

PROOF. $\tilde{A}_1, \tilde{A}_2$ have orthonormal columns and furthermore $\tilde{A}_1^\top \tilde{A}_2$ and $A_1^\top A_2$ have the same singular values. Therefore the previous proposition applies, and the claim follows. $\square$

We recall the following Lemma from [**4**], dealing with covariance of one disk:

LEMMA 38. *Let μ be the uniform distribution over $\pi \cap \mathcal{B}(0, 1)$, where π is a d-dimensional subspace π of $\mathbb{R}^D$. Then,*

$$\vec{\lambda}\Sigma[\mu] = \frac{1}{d+2}(\underbrace{1, \ldots 1}_{d}, \underbrace{0, \ldots 0}_{D-d})$$

We also need the covariance of *two* disks:

PROPOSITION 39. *Let $\pi_1, \pi_2 \in \mathrm{Gr}(d, D)$ with $\dim(\pi_1 \cap \pi_2) = d_0 \geq 2d - D$. Define $\mu = \frac{1}{2}(\mu_1 + \mu_2)$, where μ_i is the uniform measure over $\pi_i \cap \mathcal{B}(0, 1)$. Then eigenvalues of the covariance of μ are:*

$$\vec{\lambda}\Sigma[\mu] = \frac{1}{(d+2)}(\underbrace{1, \ldots 1}_{d_0}, \cos^2(\theta_{d_0+1}/2), \ldots \cos^2(\theta_d/2), \sin^2(\theta_d/2), \ldots \sin^2(\theta_{d_0+1}/2), \underbrace{0, \ldots 0}_{D-2d+d_0})$$

where $\theta_1 \leq \cdots \leq \theta_d$ are the principal angles between (π_1, π_2) with $\theta_1 = \ldots = \theta_{d_0} = 0 < \theta_{d_0+1}$.

PROOF. By starting from the matrix form in Corollary 37 and then by applying multiple 2-dimensional rotations to the standard matrix form of (π_1, π_2), the following can be proven. There exists an orthogonal matrix $V \in O(D)$ and matrices A_1, A_2 such that columns of each VA_i is an orthonormal basis of π_i:

$$A_1 = \begin{bmatrix} I_{d_0} & 0 \\ 0 & \cos\frac{1}{2}\Theta \\ 0 & \sin\frac{1}{2}\Theta \\ 0 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} I_{d_0} & 0 \\ 0 & \cos\frac{1}{2}\Theta \\ 0 & -\sin\frac{1}{2}\Theta \\ 0 & 0 \end{bmatrix}$$

where $\Theta \in \mathbb{R}^{(d-d_0) \times (d-d_0)}$ is the diagonal matrix of nonzero principal angles. Therefore $A_1 = U_1 J, A_2 = U_2 J$ where J is given by $J^\top = [I_d, 0] \in \mathbb{R}^{d \times D}$, and

$$U_1 = \begin{bmatrix} I_{d_0} & & & \\ & \cos\frac{1}{2}\Theta & -\sin\frac{1}{2}\Theta & \\ & \sin\frac{1}{2}\Theta & \cos\frac{1}{2}\Theta & \\ & & & I_{D-2d+d_0} \end{bmatrix}, U_2 = \begin{bmatrix} I_{d_0} & & & \\ & \cos\frac{1}{2}\Theta & \sin\frac{1}{2}\Theta & \\ & -\sin\frac{1}{2}\Theta & \cos\frac{1}{2}\Theta & \\ & & & I_{D-2d+d_0} \end{bmatrix}$$

Here the matrix $V \in O(D)$ simply plays the role of an orthonormal coordinate transform and can be safely ignored in calculating the eigenvalues of $\Sigma[\mu]$. Indeed, orthonormal coordinate transformation induces a conjugation on the covariance matrix, and leaves its eigenvalues invariant. Thus without loss of generality, assume that columns of each A_i is an orthonormal basis of π_i.

Let $Z \in \mathbb{R}^D$ be a random vector, drawn from the uniform distribution over the unit d-dimensional disk that spans the first d canonical basis vectors of $\mathbb{R}^D$. Then for each $X_i \sim \mu_i$, we have $X_i = U_i Z$. This implies that:

$$\Sigma[\mu_i] = \mathbb{E}[X_i X_i^\top] = U_i \mathbb{E}[ZZ^\top]U_i^\top = \frac{1}{(d+2)}U_i \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} U_i^\top$$

Thus we write U_1, U_2 in block diagonal forms:

$$U_i = \begin{bmatrix} U_i^{(11)} & U_i^{(12)} \\ U_i^{(21)} & U_i^{(22)} \end{bmatrix}$$

where

$$U_i^{(11)} = \begin{bmatrix} I_{d_0} & 0 \\ 0 & \cos \frac{1}{2}\Theta \end{bmatrix}, U_i^{(12)} = \begin{bmatrix} 0 & 0 \\ (-1)^i \sin \frac{1}{2}\Theta & 0 \end{bmatrix}$$

$$U_i^{(21)} = \begin{bmatrix} 0 & (-1)^{i+1} \sin \frac{1}{2}\Theta \\ 0 & 0 \end{bmatrix}, U_i^{(22)} = \begin{bmatrix} \cos \frac{1}{2}\Theta & 0 \\ 0 & I_{D-2d+d_0} \end{bmatrix}$$

Thus we compute:

$$\Sigma[\mu_1] = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_{11}^\top & U_{21}^\top \\ U_{12}^\top & U_{22}^\top \end{bmatrix} = \begin{bmatrix} U_{11} & 0 \\ U_{21} & 0 \end{bmatrix} \begin{bmatrix} U_{11}^\top & U_{21}^\top \\ U_{12}^\top & U_{22}^\top \end{bmatrix} = \begin{bmatrix} U_{11}U_{11}^\top & U_{11}U_{21}^\top \\ U_{21}U_{11}^\top & U_{21}U_{21}^\top \end{bmatrix}$$

Thus

$$U_{11}U_{11}^\top = \begin{bmatrix} I_{d_0} & 0 \\ 0 & \cos^2 \frac{1}{2}\Theta \end{bmatrix}, U_{11}U_{21}^\top = \begin{bmatrix} 0 & 0 \\ \cos \frac{1}{2}\Theta \sin \frac{1}{2}\Theta & 0 \end{bmatrix}, U_{21}U_{21}^\top = \begin{bmatrix} \sin^2 \frac{1}{2}\Theta & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$\Sigma[\mu_1] = \frac{1}{(d+2)} \begin{bmatrix} I_{d_0} & 0 & 0 & 0 \\ 0 & \cos^2 \frac{1}{2}\Theta & \cos \frac{1}{2}\Theta \sin \frac{1}{2}\Theta & 0 \\ 0 & \cos \frac{1}{2}\Theta \sin \frac{1}{2}\Theta & \sin^2 \frac{1}{2}\Theta & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Doing the calculation verbatim for $\Sigma[\mu_2]$ gives flipped sign for off-diagonl entries. Thus:

$$\Sigma[\mu] = \Sigma[\frac{1}{2}(\mu_1 + \mu_2)] = \frac{1}{(d+2)} \begin{bmatrix} I_{d_0} & 0 & 0 & 0 \\ 0 & \cos^2 \frac{1}{2}\Theta & 0 & 0 \\ 0 & 0 & \sin^2 \frac{1}{2}\Theta & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

This is already a diagonal matrix, and we directly take the diagonal entry to obtain the claim. □

C.4. Singularity Score. Eigenvalue control

Before deriving results on singularity score, we first need to derive results on dimension estimation and linear approximation, which are used to define the singularity score. In this section we derive results for controlling the change of eigenvalues of a real symmetric matrix. The real symmetric matrix of interest for us is the covariance matrix, from which we get eigenvalues for dimension estimation. We introduce the following notations.

DEFINITION 40. Given a symmetric real matrix $A \in \mathbb{R}^{D \times D}$, we use the following notation for the vector of eigenvalues of A, arranged in the decreasing order:

$$\vec{\lambda}A = (\lambda_1 A, \ldots \lambda_D A) \in \mathbb{R}^D$$

We also denote:

$$\lambda_k^{\text{gap}} A = \lambda_k A - \lambda_{k+1} A$$

$$\text{Tail}_k A = \lambda_{k+1} A + \cdots + \lambda_D A$$

$$\text{TQ}_k A = \frac{\text{Tail}_k A}{\text{Tail}_0 A}$$

where TQ stands for *tail quotient*. For a measure μ on $\mathbb{R}^D$, we will use a slight abuse of notation, and denote the spectral gap of its covariance matrix as:

$$\lambda_k^{\text{gap}} \mu = \lambda_k^{\text{gap}} \Sigma \mu$$

The Hoffman-Wielandt theorem [27] will be our main tool of controlling the eigenvalues:

THEOREM 41 (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*

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

where $\|A\|_{\text{F}} := \sqrt{\text{Tr}(A^\top A)}$ denotes the Frobenius norm. In particular, if A, A' are real symmetric matrices, then:

$$\|\vec{\lambda}(A) - \vec{\lambda}(A')\| \le \|A - A'\|_{\text{F}}$$

We will also use the following Lipschitz continuity relation from [32]:

PROPOSITION 42. *Let $\mu, \nu \in \mathcal{P}$ such that the support of each measure is contained in a ball of radius r. Then,*

$$\|\Sigma \mu - \Sigma \nu\| \le 8r \cdot W(\mu, \nu)$$

We then get the following bound on the variation of the spectral gap:

LEMMA 43. *Let $\mu, \nu \in \mathcal{P}$ be such that the support of each measure is contained in a ball of radius 1. Then,*

$$|\lambda_k^{\text{gap}} \mu - \lambda_k^{\text{gap}} \nu| \le 16D \cdot W(\mu, \nu)$$

PROOF. Let $A = \Sigma \mu$, $A' = \Sigma \nu$. The Hoffman-Wielandt theorem implies the following for all k:

$$D^{-1/2} \cdot |\lambda_k(A) - \lambda_k(A')| \le D^{-1/2} \cdot \|\vec{\lambda}(A) - \vec{\lambda}(A')\|_1 \le \|\vec{\lambda}(A) - \vec{\lambda}(A')\|_2 \le \|A - A'\|_{\text{F}}$$

where the second inequality is due to the fact that $D^{-1/2} \cdot \|x\|_1 \le \|x\|_2$ for any $x \in \mathbb{R}^D$. The triangle inequality then implies:

$$|\lambda_k^{\text{gap}}(A) - \lambda_k^{\text{gap}}(A')| \le |\lambda_k(A) - \lambda_k(A')| + |\lambda_{k+1}(A) - \lambda_{k+1}(A')| \le 2\sqrt{D} \cdot \|A - A'\|_{\text{F}}$$

Now the claim follows by applying Proposition 42 and the fact that Frobenius norm satisfies $\|B\|_{\text{F}} \leq \sqrt{D} \cdot \|B\|$ generally for any $B \in \mathbb{R}^{D \times D}$.

$$\|A - A'\|_{\text{F}} \leq \sqrt{D} \cdot \|A - A'\| \leq 8\sqrt{D} \cdot W(\mu, \nu)$$

$\square$

The variation of tail quotient can be controlled as follows:

LEMMA 44. *Let $\mu, \nu \in \mathcal{P}$ be such that the support of each measure is contained in a ball of radius 1. Assume that $W(\mu, \nu) \leq \beta/(16D)$, where $\beta = \|\vec{\lambda}\Sigma\mu\|_1$. Then the following holds for all k:*

$$|TQ_k(\mu) - TQ_k(\nu)| \leq 32D\beta^{-1} \cdot W(\mu, \nu)$$

PROOF. Denote $A = \Sigma\mu$, $A' = \Sigma\nu$. The Hoffman-Wielandt theorem and Proposition 42 imply:

$$|\text{Tail}_k A - \text{Tail}_k A'| = \left|\sum_{i>k} \lambda_k A - \lambda_k A'\right| \leq \sum_{i>k} |\lambda_k A - \lambda_k A'| \leq \|\vec{\lambda}A - \vec{\lambda}A'\|_1$$

$$\leq \sqrt{D} \cdot \|\vec{\lambda}A - \vec{\lambda}A'\|_2 \leq \sqrt{D} \cdot \|A - A'\|_{\text{F}} \leq 8D \cdot W(\mu, \nu)$$

where we also used the fact that $D^{-1/2} \cdot \|x\|_1 \leq \|x\|_2$ generally for any $x \in \mathbb{R}^D$ and $\|B\|_{\text{F}} \leq \sqrt{D} \cdot \|B\|$ generally for any $B \in \mathbb{R}^{D \times D}$. Define the following notations:

$$\alpha = \text{Tail}_k(\Sigma\mu), \quad \beta = \text{Tail}_0(\Sigma\mu)$$

$$t_1 + \alpha = \text{Tail}_k(\Sigma\nu), \quad t_2 + \beta = \text{Tail}_0(\Sigma\nu), \quad t = 8D \cdot W(\mu, \nu)$$

From the above we know that $|t_1|, |t_2| \leq t$ and by assumption $t \leq \beta/2$. By simple calculation the claim follows:

$$\left|\frac{\alpha + t_1}{\beta + t_2} - \frac{\alpha}{\beta}\right| = \left|\frac{t_1\beta - t_2\alpha}{\beta(\beta + t_2)}\right| \leq \frac{t(\beta + \alpha)}{\beta(\beta - t)} \leq \frac{2\beta t}{\beta^2/2} = 4\beta^{-1}t$$

$\square$

Stability

In this section we show that when spectral gap is bounded from below and the estimated dimension are constant, then the singularity score obeys a Lipschitz continuity relation. We first note the following variant of the Davis-Kahan theorem [23, 53]:

THEOREM 45 (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 $\min(\lambda^{\text{gap}}_{d_1-1}A, \lambda^{\text{gap}}_{d_2}A) > 0$. Let π_A be the span of the eigenspaces corresponding to $\lambda_{d_1}, \lambda_{d_1+1}, \ldots \lambda_{d_2}$, and let $\theta_1 \leq \ldots \leq \theta_d$ be the principal angles between (π_A, π_B). Then we have:*

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

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

$$\sqrt{\sin^2 \theta_1 + \cdots + \sin^2 \theta_d} \leq \frac{2}{\lambda_d^{\text{gap}} A} \cdot \min\left(\|A - B\|_{\text{F}}, \sqrt{d}\|A - B\| \right)$$

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

We then prove the following Lipschitz continuity relation for the singularity score:

PROPOSITION 46. *Let* $\mu, \nu \in \mathcal{P}$ *be measures whose supports are contained in the ball of radius 1 centered at the origin of* $\mathbb{R}^D$. *Assume that for some* $\eta \in (0, 1)$ *and* $k > 0$, *we have* $\hat{d}_\eta(\mu) = \hat{d}_\eta(\nu)$, *and define* $s = \max(\lambda_k^{\text{gap}} \Sigma\mu, \lambda_k^{\text{gap}} \Sigma\nu)$. *Then the following hold:*

$$\mathfrak{S}(\mathcal{L}_\eta \mu, \mathcal{L}_\eta \nu) \leq 8s^{-1} \cdot W(\mu, \nu)$$

$$|\sigma_\eta(\mu) - \sigma_\eta(\nu)| \leq (1 + 8s^{-1}) W(\mu, \nu)$$

where $\mathfrak{S}(\pi_1, \pi_2) = \sqrt{\sin^2 \theta_1 + \cdots + \sin^2 \theta_d}$, *with* $(\theta_1, \ldots \theta_d)$ *being the principal angles between* (π_1, π_2).

PROOF. Let's abbreviate $\sigma = \sigma_\eta, \mathcal{L} = \mathcal{L}_\eta$. By assumption $\dim \mathcal{L}\mu = \dim \mathcal{L}\nu = k$ and we now apply the Davis-Kahan theorem and Proposition 42:

$$\mathfrak{S}(\mathcal{L}\mu, \mathcal{L}\nu) \leq \frac{1}{s}\|\Sigma\mu - \Sigma\nu\| \leq \frac{8}{s} W(\mu, \nu)$$

For the singularity score, we get:

$$|\sigma(\mu) - \sigma(\nu)| = |\Delta(\mu_\perp, \mathfrak{u}_k) - \Delta(\nu_\perp, \mathfrak{u}_k)| \leq \Delta(\mu_\perp, \nu_\perp) \leq W(\mu_\perp, \nu_\perp)$$

where $\mu_\perp = \Pi(\mu, \mathcal{L}\mu), \nu_\perp = \Pi(\nu, \mathcal{L}\nu)$. Furthermore,

$$\begin{aligned}
W(\mu^\perp, \nu^\perp) &= W\left(\Pi(\mu, \mathcal{L}\mu), \Pi(\nu, \mathcal{L}\nu) \right) \\
&\leq W\left(\Pi(\mu, \mathcal{L}\mu), \Pi(\nu, \mathcal{L}\mu) \right) + W\left(\Pi(\nu, \mathcal{L}\mu), \Pi(\nu, \mathcal{L}\nu) \right) \\
&\leq W(\mu, \nu) + \mathfrak{S}(\mathcal{L}\mu, \mathcal{L}\nu) \\
&\leq \left(1 + \frac{8}{s} \right) W(\mu, \nu)
\end{aligned}$$

where in the second to last inequality we applied Lemmas 21 and 22. □

In the following Proposition, spectral gap, tail quotient, dimension estimate, and the singularity score are simultaneously controlled, for a measure ν sufficiently close to a given measure μ. In our application of the Proposition, ν will be an empirical measure, from which the empirical singularity score will be calculated.

PROPOSITION 47. *Let $\mu, \nu \in \mathcal{P}$ be measures supported on the unit ball $\mathcal{B}(0,1) \subset \mathbb{R}^D$. Let $a \in [0,1]$ and $k \geq 0$ be an integer. Suppose that $W(\mu, \nu)$ is sufficiently small; explicitly, assume that:*

$$W(\mu, \nu) \leq \frac{\min(4as, 4\beta, a\beta G)}{64D}$$

where $\beta = \|\vec{\lambda}\Sigma\mu\|_1$, $G_k = TQ_{k-1}(\mu) - TQ_k(\mu)$, and $s = \lambda_k^{\text{gap}}(\mu)$. Then the following hold:
(1) There is a bound for spectral gaps:

$$\lambda_k^{\text{gap}}(\nu) \geq (1-a)\lambda_k^{\text{gap}}(\mu)$$

(2) There are inclusions of intervals of tail quotients:

$$J_{k,0}(\nu) \supseteq J_{k,a}(\mu), \quad J_{k,0}(\mu) \supseteq J_{k,a}(\mu)$$

where $J_{k,a}(\mu) = [TQ_k(\mu) + \frac{a}{2}G_k, TQ_{k-1}(\mu) - \frac{a}{2}G_k]$.
(3) For a choice of $\eta \in J_{k,a}(\mu)$, the following hold:

$$\hat{d}_\eta(\mu) = \hat{d}_\eta(\nu) = k$$

$$\mathfrak{S}(\pi_1, \pi_2) \leq 8s^{-1} \cdot W(\mu, \nu)$$

$$|\sigma_\eta\mu - \sigma_\eta\nu| \leq (1 + 8s^{-1}) \cdot W(\mu, \nu)$$

where $\mathfrak{S}(\pi_1, \pi_2) = \sqrt{\sin^2\theta_1 + \cdots + \sin^2\theta_d}$, with $(\theta_1, \ldots \theta_d)$ being the principal angles between (π_1, π_2).

PROOF. (1) By Lemma 43 and the assumption on $W(\mu, \nu)$, we have:

$$|\lambda_k^{\text{gap}}\mu - \lambda_k^{\text{gap}}\nu| \leq 16D \cdot W(\mu, \nu) \leq a \cdot s$$

(2) By Lemma 44 and the assumption on $W(\mu, \nu)$, for each j we have:

$$|TQ_j(\mu) - TQ_j(\nu)| \leq \frac{32D}{\beta} \cdot W(\mu, \nu) \leq \frac{a}{2} \cdot G_k(\mu)$$

This implies the first inclusion, and the second inclusion holds trivially.

(3) By (2), $\eta \in J_{k,a}(\mu)$ implies both $\eta \in J_{k,0}(\nu)$ and $\eta \in J_{k,0}(\mu)$, so that $\hat{d}_\eta(\mu) = \hat{d}_\eta(\nu) = k$ by the definition of $\hat{d}_\eta$. The rest of the claims follow from Proposition 46. $\qquad \square$

Limit behaviour

PROPOSITION 48. *Let $M \subset \mathbb{R}^D$ be a d-dimensional submanifold. Suppose $\eta \in (0, (d+2)^{-1})$. Then for any $x \in M$, $\sigma_\eta(\mu_{x,0}) = 0$.*

PROOF. By Proposition 24, we have $\mu_{x,0} = \mathcal{H}^d|_{T_x^\circ M}$, where $T_x^\circ M = T_x M \cap \mathcal{B}(0,1)$. Also noting that the spectral gap of $\mu_{x,0}$ is $(d+2)^{-1}$, we obtain the claim. $\qquad \square$

PROPOSITION 49. *Let $M = M_1 \cup M_2 \subset \mathbb{R}^D$ be a union of two d-dimensional submanifolds such that for any $x \in M_1 \cap M_2$, we have $\dim(T_x M_1 \cap T_x M_2) = d_0$, and all principal angles of $(T_x M_1, T_x M_2)$ are bounded above a fixed constant $\phi > 0$. Suppose $\eta \in (0, (d+2)^{-1} \cdot \sin^2(\phi/2))$. Then the function $x \mapsto \sigma_\eta(\mu_{x,0})$ is continuous on $M_1 \cap M_2$, and takes positive values. In particular, we have $\inf_{x \in M_1 \cap M_2} \sigma_\eta(\mu_{x,0}) > 0$.*

PROOF. Due to the eigenvalue computation in Proposition 39, the condition $\eta < d^{-1} \cdot \sin^2(\phi/2)$ implies

$$x \in M_1 \cap M_2 \implies \hat{d}_\eta(\mu_{x,0}) = 2d - d_0$$

Also, we have a lower bound on the $(2d - d_0)$-th spectral gap:

$$\lambda_{2d-d_0}^{\text{gap}}(\Sigma \mu_{x,0}) \geq \frac{\sin^2(\phi/2)}{d+2}$$

Therefore we can apply Proposition 46, and see that the function $x \mapsto \sigma_\eta(\mu_{x,0})$ is (Lipschitz) continuous on $M_1 \cap M_2$.

The projected measure $(\mu_{x,0})_\perp$ is the (pushforward along) projection of $\mu_{x,0}$ to the $(2d - d_0)$-dimensional space spanned by $T_x M_1 + T_x M_2$. This measure, supported along the union of two d-dimensional disks, is clearly not equal to the $(2d - d_0)$-dimensional uniform measure $\mathfrak{u}_{2d-d_0}$. Then the universality of kernel MMD implies that:

$$\sigma_\eta(\mu_{x,0}) = \Delta\left((\mu_{x,0})_\perp, \, \mathfrak{u}_{2d-d_0} \right) > 0$$

Therefore the function $x \mapsto \sigma_\eta(\mu_{x,0})$ is continuous and positive on a compact set $M_1 \cap M_2$, so that its infimum is also positive. $\qquad\square$

C.5. Proof of the main theorem. In this section we prove Theorem 7, the main mathematical result of this article.

Definitions.
For the logical clarity of the proof, we will first define some constants, and postpone the explanation for their choice to later parts of the proof.
τ, ψ, ζ, s_0 are defined as:

$$\tau = \min(\tau_1, \tau_2), \quad \psi = \frac{\sin^2(\phi/2)}{d+2}, \quad \zeta = \frac{\psi \cdot d}{128(d+2)}, \quad s_0 = 1 + \frac{8}{\psi}$$

η_-, η_+ are defined as:

$$\eta_- = \frac{1}{4}\psi, \quad \eta_+ = \frac{3}{4}\psi$$

ξ, ξ_0 are defined as:

$$3\xi = \inf_{x \in M_1 \cap M_2} \sigma_{\psi/2}(\mu_{x,0}), \quad \xi_0 = \min\left(\zeta, \frac{\xi}{s_0} \right)$$

r_0, c_A, c_B are defined as:

$$r_0 = \min\left(c_5, c_7, c_9, \frac{\xi_0}{2c_8}, \frac{\xi_0}{8c_6}, \frac{\xi_0}{4c_{10}}\right) \cdot \tau$$

$$c_A = \min\left(c_5, \frac{\xi_0}{8c_6}\right)$$

$$c_B = \max\left(h(M_1, M_2)^{-1}, h(M_2, M_1)^{-1}\right)$$

$$\text{where } h(M_1, M_2) = \inf_{x \in M_1} \frac{d(x, M_2)}{d(x, M_1 \cap M_2)}$$

where the constants $c_5, \ldots c_{10}$, which depend only on d, are defined in Propositions 24, 23, and 25.

Finally, we *fix a choice* of η, r as any number in the range:

$$\eta \in [\eta_-, \eta_+], \quad r \in (0, r_0]$$

We remark that if $\nu \in \mathcal{P}$ and $\hat{d}_\eta(\nu) = \hat{d}_{\eta'}(\nu)$ for some threshold values η, η', then we have $\sigma_\eta(\nu) = \sigma_{\eta'}(\nu)$. In particular, due to Propositions 39, 49, $x \in M_1 \cap M_2$ implies $\hat{d}_{\psi/2}(\mu_{x,0}) = \hat{d}_\eta(\mu_{x,0})$, and thus:

$$3\xi = \inf_{x \in M_1 \cap M_2} \sigma_{\psi/2}(\mu_{x,0}) = \inf_{x \in M_1 \cap M_2} \sigma_\eta(\mu_{x,0})$$

Outline.

We will first describe the non-random situation in detail and then describe the randomness using the Wasserstein concentration inequality. Let $\mathbf{x} = (x_1, \ldots x_n) \subset M$. Define the (non-random) singularity scores:

$$\sigma_i = \sigma_i(\mathbf{x}, r, \eta) = \sigma_\eta(\hat{\mu}_i)$$

where $\hat{\mu}_i$ is defined using $(\mathbf{x}, r)$ as described in the Introduction.

Our strategy of proof involves the following successive approximations:

Singular part: $d(x_i, M_1 \cap M_2) \leq c_A r \implies \sigma(\hat{\mu}_i) \approx \sigma(\mu_{x_i,r}) \approx \sigma(\mu_{y_i,r}) \approx \sigma(\mu_{y_i,0}) \geq 3\xi$

Smooth part: $d(x_i, M_1 \cap M_2) \geq c_B r \implies \sigma(\hat{\mu}_i) \approx \sigma(\mu_{x_i,r}) \approx \sigma(\mu_{x_i,0}) = 0$

where y_i is the projection from x_i to $M_1 \cap M_2$. By the choice of parameters made before, the approximations will each amount to at most ξ of error, so that in the smooth case we have $\sigma(\hat{\mu}_i) \leq \xi$ and in the singular case we have $\sigma(\hat{\mu}_i) \geq 2\xi$. We now describe the proof precisely.

Limit behaviour. We apply Propositions 49 and 48, and see that:

$$x \in (M_1 \cup M_2)\backslash(M_1 \cap M_2) \implies \sigma(\mu_{x,0}) = 0$$

$$x \in M_1 \cap M_2 \implies \sigma(\mu_{x,0}) \geq 3\xi > 0$$

Singular part. When $\mathrm{d}(x_i, M_1 \cap M_2) \leq c_A r$, the following holds:

$$\sigma_i \geq \sigma(\mu_{y,0}) - |\sigma(\hat{\mu}_i) - \sigma(\mu_{y_i,0})| = 3\xi - |\sigma(\hat{\mu}_i) - \sigma(\mu_{y_i,0})| \tag{C.7}$$

where $y_i \in M_1 \cap M_2$ is a point satisfying $\mathrm{d}(x_i, y_i) = \mathrm{d}(x_i, M_1 \cap M_2)$[15].

Smooth part. When $\mathrm{d}(x_i, M_1 \cap M_2) \geq c_B r$, the following holds:

$$\sigma_i \leq \sigma(\mu_{x_i,0}) + |\sigma(\hat{\mu}_i) - \sigma(\mu_{x_i,0})| = 0 + |\sigma(\hat{\mu}_i) - \sigma(\mu_{x_i,0})| \tag{C.8}$$

Even though Equations (C.7) and (C.8) didn't use anything specific about the distance $\mathrm{d}(x_i, M_1 \cap M_2)$, this will be used while controlling the error terms.

From singularity score to Wasserstein distance.

Differences of singularity scores are controlled using Proposition 47, which is a Lipschitz continuity relation with respect to the Wasserstein distance. Our definition of ζ is obtained by setting $a = 1/2$, $\beta = d/(d+2)$, $G = s = \psi$ in the condition in Proposition 47. This then implies that for all $x \in M$ and $\nu \in \mathcal{P}$,

$$W(\mu_{x,0}, \nu) \leq \zeta \implies |\sigma(\mu_{x,0}) - \sigma(\nu)| \leq s_0 \cdot W(\mu_{x,0}, \nu)$$

where we recall our definition $s_0 = 1 + 8\psi^{-1}$. The definition of ξ_0 allows us to make a more straightforward inference:

$$W(\mu_{x,0}, \nu) \leq \xi_0 \implies |\sigma(\mu_{x,0}) - \sigma(\nu)| \leq \xi \tag{C.9}$$

Therefore we can control error terms in Equation (C.7), (C.8) using the Wasserstein distance.

Wasserstein distance control.

Singular part. Suppose that x satisfies $\mathrm{d}(x, M_1 \cap M_2) \leq c_A r$. Let $y \in M_1 \cap M_2$ satisfy $\mathrm{d}(x, y) = \mathrm{d}(x, M_1 \cap M_2)$. We denote $\rho = r/\tau$ and also set $s = \|x - y\|/r$. Then Propositions 24 and 23 imply that:

$$\rho, s \leq c_5 \implies W(\mu_{x,r}, \mu_{y,r}) \leq c_6(\rho + s)$$

$$\rho \leq c_9 \implies W(\mu_{y,r}, \mu_{y,0}) \leq c_{10}\rho$$

Our definitions of r_0, c_A allows us to apply the bounds above, and we obtain:

$$W(\mu_{x,r}, \mu_{y,0}) \leq W(\mu_{x,r}, \mu_{y,r}) + W(\mu_{y,r}, \mu_{y,0}) \leq \frac{\xi_0}{2} \tag{C.10}$$

---

[15]Compactness of $M_1 \cap M_2$ and continuity of the distance function implies that such a y exists.

Smooth part. Suppose that x satisfies $d(x, M_1 \cap M_2) \geq c_B r$. Here our choice of c_B allows us to apply Proposition 26, so that $\mathcal{B}(x, r)$ intersects *either only one* of M_1 or M_2. Thus we only need to work with one manifold at a time here. Thus Proposition 25 implies:

$$\rho \leq c_7 \implies W(\mu_{x,r}, \mu_{x,0}) \leq c_8 \rho$$

and yet again by our definition of r_0, this bound implies:

$$W(\mu_{x,r}, \mu_{x,0}) \leq \frac{\xi_0}{2} \tag{C.11}$$

Empirical estimation.

Almost all of the puzzle pieces have been fit together to complete the proof. It now remains to control the probability of empirical esimtation.

We reintroduce randomness, and let $\mathbf{X}_n = (X_1, \ldots X_n)$ be an iid sample drawn uniformly from $M_1 \cup M_2$. The choice of all other parameters remain the same as before. We plug in the error level of $t = r\xi_0/2$ to Proposition 20[16], and obtain the following. Whenever $n \geq \max(N, 2/u_-)$, we have:

$$\Pr\left(\max_i W(\hat{\mu}_i, \mu_{X_i, r}) \leq \frac{\xi_0}{2}\right) \geq 1 - \delta_m \tag{C.12}$$

where $\lim_{m \to \infty} \delta_m = 0$ exponentially fast, given explicitly as:

$$\delta = c \cdot n^{N+1} \gamma^n$$

where

$$c = \left(\frac{u_+}{1 - u_+}\right)^N, \quad N = \left\lceil \left(\frac{408}{\xi_0}\right)^D \right\rceil, \quad \gamma = 1 - u_-(1 - e^{-\xi_0^2/32})$$

$$u_- = \inf_{x \in \mathrm{supp}\, \mu} \mu(\mathcal{B}(x, r)), \quad u_+ = \sup_{x \in \mathrm{supp}\, \mu} \mu(\mathcal{B}(x, r))$$

Therefore there exists some $n_0 > 0$ such that, for the $\delta > 0$ given in our theorem, $n \geq n_0$ implies $\delta_n \leq \delta$. Note that this n_0 depends on δ, μ, r, ξ_0, which have already been fixed in the beginning of the proof.

Combining the bound.

We now complete the proof. When $n \geq n_0$, the following holds for every i, with probability at least $1 - \delta$:

$$W(\hat{\mu}_i, \mu_{X_i, r}) \leq \frac{\xi_0}{2}$$

---

[16]Instead of $t = \xi_0/2$, we plug in $t = r\xi_0/2$ because we are controlling the Wasserstein distance between measures that have been rescaled by the factor of r^{-1}.

Equations (C.10) and (C.11) apply verbatim for the random setting:

$$d(X_i, M_1 \cap M_2) \leq c_A r \implies W(\mu_{X_i, r}, \mu_{Y_i, 0}) \leq \frac{\xi_0}{2}$$

$$d(X_i, M_1 \cap M_2) \geq c_B r \implies W(\mu_{X_i, r}, \mu_{X_i, 0}) \leq \frac{\xi_0}{2}$$

where $Y_i \in M_1 \cap M_2$ is a point satisfying $d(X_i, Y_i) = d(X_i, M_1 \cap M_2)$. Therefore by the triangle inequality, the above two equations imply that:

$$d(X_i, M_1 \cap M_2) \leq c_A r \implies W(\hat{\mu}_i, \mu_{Y_i, 0}) \leq \xi_0$$

$$d(X_i, M_1 \cap M_2) \geq c_B r \implies W(\hat{\mu}_i, \mu_{X_i, 0}) \leq \xi_0$$

This precisely fits the condition in Equation (C.9), from which we obtain that:

$$d(X_i, M_1 \cap M_2) \leq c_A r \implies |\sigma(\hat{\mu}_i) - \sigma(\mu_{Y_i, 0})| \leq \xi$$

$$d(X_i, M_1 \cap M_2) \geq c_B r \implies |\sigma(\hat{\mu}_i) - \sigma(\mu_{X_i, 0})| \leq \xi$$

Plugging them into Equations (C.7) and (C.8), we obtain the conclusion of the theorem:

$$d(X_i, M_1 \cap M_2) \leq c_A r \implies \sigma(\hat{\mu}_i) \geq 2\xi$$

$$d(X_i, M_1 \cap M_2) \geq c_B r \implies \sigma(\hat{\mu}_i) \leq \xi$$