# Geometric and Topological Inference from Random Samples


Sung Hyun Lim (Uzu Lim)

Merton College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Trinity 2023

# Acknowledgements

This thesis marks a key point of my mathematical life, and I would like to thank the people who made this possible.

To my family: 즐거울 때에도 너덜너덜할 때에도 함께해준 우리 가족. 타지생활을 오래 하면서 힘이 되어줘서 고마워요. 밥 잘 먹었냐고 챙겨주는 엄마, 조용하지만 든든하게 지원해주는 아빠. 그리고 힘든 얘기 들어주면서 이것저것 이야기 나눈 지수. 나는 이렇게 30년동안의 수학 여행에서 쉼표를 찍고 갑니다.

To Vidit and Harald. Thank you for guiding my mathematical journey for the past 4 years. While navigating the many different directions of my research, your advices helped me re-center myself and made my mathematics mature better.

To Shriram. Thank you for always being there to listen. It's great to banter and learn things with you. You've changed a lot about how I see the world, and I will continue paying attention to deeper undercurrents of life.

To Yonatan Fruchter. I'm glad that we maintained our friendship. Hanging out with you made the DPhil journey more fascinating, when we played that oud-cutlery-bluegrass arrangement or when you cooked huge barbecue feasts at midnight.

To Juniper Nights - Zach, James, and Alex. We built something special musically. The Tuesday nights during the band's years have been a very special chance to dedicate my musical energy.

And lastly, some small but important things-

- 침착맨님, 잔잔하게 항상 웃겨주셔서 감사합니다. 보면서 마음도 많이 따뜻해졌어요.
- Celeste: My soft queer heart glowed from this amazing game.
- Arknights: Intricate lore and engaging gameplay, which always stayed with me.
- Kill Six Billion Demons: This unique series taught me many life lessons.
- Lao Gan Ma: The best sauce ever that helped me throughout the DPhil life.

# Abstract

We study statistical inference problems in geometry and topology inspired by data science. Generally we consider an independently, identically distributed random sample $\mathbf{X} = (X_1, \dots X_n)$ drawn from a measure $\mu$ over a set $M \subseteq \mathbb{R}^D$. Then we consider the class of problem of inferencing a property $\mathcal{P}$ of $M$ only using $\mathbf{X}$. The following pairs of $(M, \mathcal{P})$ are considered:

**(1)** $M$ **= Manifold,** $\mathcal{P}$ **= Dimension.** We locally apply PCA (principal components analysis), and infer the dimension of $M$ by counting how many of the variances fall under a threshold. While this method is widely used, a rigorous mathematical theorem guaranteeing its correctness was not established. We prove such a theorem.

**(2)** $M$ **= Manifold,** $\mathcal{P}$ **= Tangent spaces.** The local PCA algorithm above can also be used to infer tangent spaces: a tangent space is estimated as a linear span of top principal components. Again for this standard well-known algorithm, we prove theorems guaranteeing the correctness of the algorithm.

**(3)** $M$ **= Stratified space,** $\mathcal{P}$ **= Singular points.** A stratified space generalises manifolds, and possesses singularities at which there is no local resemblance to a Euclidean space. We present a fast algorithm that detects singularities using local hypothesis testing. The kernel method used in the algorithm is significantly faster than previous topological methods. Experimental results on both synthetic and real data are presented. Furthermore, we prove a theorem that guarantees the algorithm's correctness in the case of union of two manifolds.

**(4)** $M$ **= Manifold,** $\mathcal{P}$ **= Homotopy type.** A standard way to infer homotopy type of a manifold from a finite sample is via constructing a simplicial complex at a small distance threshold. However instead of stopping at a small threshold, we consider arbitrarily large connectivity thresholds and study anomalous topology arising from this. In particular, we study a very specific case of circle $M = \mathbb{S}^1$, and show that Cech complexes arising from finite samples on $M$ are homotopic to bouquets of high-dimensional spheres with high probability.

# Contents


**1 Introduction** **1**

**2 Differential geometry** **7**

2.1 Geometry of embedded manifolds . . . . . 7

2.2 Calculus . . . . . 17

**3 Statistical distance** **20**

3.1 Wasserstein distance . . . . . 20

3.2 Kernel distance . . . . . 30

**4 Concentration inequality** **39**

4.1 Covariance matrix . . . . . 39

4.2 Wasserstein distance . . . . . 44

**5 Tangent space and dimension estimation** **49**

5.1 Introduction . . . . . 49

5.2 Lipschitz property of covariance matrix . . . . . 56

5.3 Flattening a measure on manifold . . . . . 59

5.4 Principal angles . . . . . 68

5.5 Tangent space and dimension estimation . . . . . 70

**6 Singularity detection - Theory** **78**

6.1 Introduction . . . . . 78

6.2 Eigenvalue control . . . . . 81

6.3 Covariance of two disks . . . . . 83

6.4 Singularity score . . . . . 85
6.5 Proof of the main theorem . . . . . 88

**7 Singularity detection - Experiments** **95**
7.1 Introduction . . . . . 95
7.2 Algorithm . . . . . 97
7.3 Comparison with other methods . . . . . 106
7.4 Experiments . . . . . 110
7.5 Experimental details . . . . . 119

**8 Strange random topology of the circle** **124**
8.1 Introduction . . . . . 124
8.2 Expected Euler characteristic . . . . . 128
8.3 Limit behaviour of Euler characteristic . . . . . 131
8.4 Random homotopy types . . . . . 135
8.5 Odd spheres . . . . . 141

**Bibliography** **144**

# List of Figures


5.1 Local PCA on a manifold . . . . . 51
5.2 Transportation plan for locally flattening a manifold . . . . . 62
5.3 Local view of a noisy measure on a manifold . . . . . 63

7.1 Algo Schematic . . . . . 98
7.2 Comparison of Hades and Ripser . . . . . 107
7.3 Diminishing persistence of the main topological signal of spheres . . . . . 108
7.4 Comparison of Hades and anomaly detection algorithms . . . . . 110
7.5 Hades applied to synthetic datasets, visualised . . . . . 112
7.6 Hades applied to synthetic datasets, quantified with ROC and AUC . . . 113
7.7 Hades applied to test the manifold hypothesis . . . . . 114
7.8 Hades applied to road network datasets . . . . . 116
7.9 Hades applied to the cyclo-octane conformation dataset . . . . . 116
7.10 Hades applied to the MNIST dataset . . . . . 118
7.11 Hades applied to the Fashion-MNIST dataset . . . . . 119

8.1 Expected Euler characteristics of random Cech complexes on a circle . . 125

# Chapter 1

# Introduction

A standard form of data is given by tuples of numbers. Mathematically, this is a *point cloud*, defined as a finite subset of a Euclidean space $\mathbf{x} = \{x_1, \dots x_n\} \subset \mathbb{R}^D$. We may imagine $\mathbf{x}$ to be a set of points floating in a $D$-dimensional space, outlining a shape. A simple probabilistic model for $\mathbf{x}$ is that the points $x_1, \dots x_n$ are iid (independent, identically distributed) samples drawn from a probability measure $\mu$ over $\mathbb{R}^D$. Then we may ask: how do we infer the *shape* of $\mu$ from $\mathbf{x}$?

This vague question can be made specific, leading us to topics at the intersection of data science, geometry, and topology. There are natural geometric objects we can use from the mathematical perspective: graphs, simplicial complexes, manifolds, stratified spaces, and so on. Furthermore, standard toolboxes include tangent spaces, curvature, and homology groups. Statistical inference of these properties is the subject of this thesis. Our problems all take the following form:

**Meta-Question.** *Let $\mathbf{X} = (X_1, \dots X_n)$ be an iid sample drawn from a measure $\mu$ on a subset $M \subseteq \mathbb{R}^D$. Let $\mathcal{P}$ be a property of $M$. Estimate $\mathcal{P}$ from $\mathbf{X}$ without $M$.*

**Main topics.** The following pairs of $(M, \mathcal{P})$ in the Meta-Question are considered.

- $M$ = Manifold, $\mathcal{P}$ = Dimension (Chapter 5).

  We locally apply PCA (principal component analysis) on a manifold to estimate its intrinsic dimension. This is done by estimating covariance matrix locally, and then counting the number of principal components required to explain a specified percentage (e.g. 95%). This is a standard algorithm known as Local PCA, yet a mathematical theorem establishing the correctness of this algorithm was not proven before. We prove such a theorem, for points sampled from a noisy non-uniform measure on a manifold. Content of this chapter is published on *SIAM Journal on Applied Algebra and Geometry* [64].

- $M$ = Manifold, $\mathcal{P}$ = Tangent spaces (Chapter 5).

  We locally apply PCA on a manifold to estimate tangent spaces at data points. The Local PCA algorithm is almost same as above, where we use the span of principal components to estimate tangent spaces. Unlike dimension estimation, there are previous mathematical literature that proves that Local PCA-based tangent space estimation on a manifold works correctly. However the theorem proven in this chapter is more general and explicitly computes all constants appearing in the theorem. Content of this chapter is published on *SIAM Journal on Applied Algebra and Geometry* [64].

- $M$ = Stratified space, $\mathcal{P}$ = Singular points (Chapter 6, 7).

  We detect singularities in a stratified space using local measure comparison. A stratified space generalises manifolds, and possesses singularities at which there is no local resemblance to a Euclidean space. Our algorithm detects this singularity by locally reducing dimension of data points and comparing them to a uniform distribution. This comparison is done using kernel MMD, a statistical distance for comparing two measures. This kernel method is much faster than the previous line of research that attempted the same task with topological methods. We prove that this algorithm correctly detects singularities with high probability, in the case of a stratified space given by a union of two manifolds.

  We furthermore implement the algorithm on Python and demonstrate its efficacy

using various computational experiments. On synthetic datasets, the algorithm cleanly extracts singularities from datasets of known geometry, and attains high AUC scores. By aggregating local information of singularity score, the algorithm is also able to perform a *global* test of whether the underlying dataset is a manifold. Furthermore, the algorithm recovers known singularities in real datasets: road networks and cyclo-octane conformations. On image datasets, the algorithm separates locally well-behaved regular images from the anomalous images that deviate from its class. Content and code of these two chapters are published on *SIAM Journal on Mathematics of Data Science* [65] and GitHub (github.com/uzulim/hades).

- $M$ = Manifold, $\mathcal{P}$ = Homotopy type (Chapter 8).

  We quantify high-dimensional topology arising from the Cech complex constructed on a random sample drawn from a *circle*. The investigations done here deviate from the standard methods used in topological data analysis, in which homotopy type of a manifold from a finite sample is "recovered" by constructing a simplicial complex at a small distance threshold. Instead of stopping at a small threshold, we consider arbitrarily large connectivity thresholds and study anomalous topology arising from this. We only study the case of random samples drawn from a circle, which already yields a rich array of strange topological phenomena. We show that odd-dimensional spheres and bouquets of even-dimensional spheres appear at specific parts of filtration radii with high probability. Content of this chapter is published on *Discrete & Computational Geometry* [63].

**Technical tools.** Tools used to analyse these problems are drawn from differential geometry and statistics. While the above topics were separate, the toolbox used to understand them developed on a common thread. We collect technical results developed for this purpose on Chapters 2 - 4. These chapters are meant as reference materials for later chapters, and therefore the reader may benefit from first reading Chapters 5 - 8 for overviews of the main results of the thesis, and then revisit the earlier chapters for reference. These early chapters are composed as follows:

- Differential geometry (Chapter 2).

  We study the geometry of manifolds that are smoothly embedded in an ambient Euclidean space. Specifically, we study behaviour of geodesics and volumes of this embedded manifold.

- Statistical distance (Chapter 3).

  We study two statistical distances, the Wasserstein distance and the kernel MMD. These distances are used to quantify how different two probability distributions are. Many theoretical arguments in the thesis boil down to proving that two probability distributions have a small distance between them. These are then used to estimate quantities using Lipschitz continuity relations, which are proven separately in later Chapters.

- Concentration inequality (Chapter 4).

  A concentration inequality asserts that a random sample approximates a target quantity with high probability. We study two types of concentration inequalities: that of covariance matrix and the Wasserstein distance. We modify known results that perform *global* estimation into *simultaneous local* estimations.

**Notations and conventions.** The reader will benefit from skimming over the list below, before venturing deeper into the thesis.

| Notation | Meaning |
|---|---|
| $\#A$ | The number of elements in a set $A$ |
| $\mathbb{E}[X]$ | Expected value of a random variable $X$ |
| $\mathbb{P}[E]$ | Probability of an event $E$ |
| $\Sigma[\mu]$ | The $D \times D$ covariance matrix of a measure $\mu$ on $\mathbb{R}^D$ |
| $\mathcal{B}_r(x) = \mathcal{B}(x, r)$ | Open ball of radius $r$, centered at $x$ |
| $\mathrm{d}_M(x, y)$ | Distance between $x$ and $y$ in a metric space $M$ |
| $\mathrm{d}_M(A, B)$ | Distance between $A$ and $B$, defined as $\mathrm{d}_M(A, B) = \inf_{x \in A, y \in B} \mathrm{d}_M(x, y)$ |
| $\|A\| = \|A\|_2$ | The operator norm of a matrix $A$ |
| $\|A\|_{\mathrm{F}}$ | The Frobenius norm of a matrix $A$ |
| $\mathrm{W}_p(\mu, \nu)$ | The $p$-Wasserstein distance between $\mu$ and $\nu$ |
| $\mathrm{W}(\mu, \nu)$ | $\mathrm{W}(\mu, \nu) = \mathrm{W}_1(\mu, \nu)$ |
| $\vec{\lambda}A$ | Vector of eigenvalues of a real symmetric matrix $A$ in the decreasing order |
| $\lambda_k A$ | The $k$-th entry of $\vec{\lambda}A$ |
| $\lambda_k^{\mathrm{gap}} A$ | The $k$-th spectral gap: $\lambda_k A - \lambda_{k+1} A$ |
| $\Pi(x, A)$ | Projection of a point $x$ to a set $A$ |
| $\Pi(\mu, A)$ | Pushforward of a measure $\mu$ along the projection $\Pi(-, A)$ |
| $J_x f$ | The Jacobian determinant of a function at $x$. |
| $\mathcal{H}^d$ | The $d$-dimensional Hausdorff measure |
| $\mu\|_U(V)$ | Normalised restriction of a measure, defined as $\mu\|_U(V) = \mu(U \cap V)/\mu(U)$ |
| $\mathcal{P}(M)$ | The set of Borel probability measures on a topological space $M$ |
| $\mathcal{P}$ | $\mathcal{P} = \mathcal{P}(\mathbb{R}^D)$ |
| $\tau_M$ | Reach of a set $M$ |
| $\omega_d$ | $\omega_d = \pi^{d/2}/\Gamma(\frac{d}{2} + 1)$, volume of the unit $d$-dimensional ball |
| $\delta_x$ | Dirac delta measure centred at a point $x$ |
| $\delta_{\mathbf{x}}$ | For $\mathbf{x} = \{x_1, \dots x_n\}$, $\delta_{\mathbf{x}} = \frac{1}{n}(\delta_{x_1} + \dots + \delta_{x_n})$. |
| $g_{x,r}(y)$ | Affine linear map $r^{-1}(y - x)$ |

The following is the choice of symbols we will be making throughout the thesis.

| **Notation** | **Meaning** |
|---|---|
| $M$ | Manifold or a stratified space, embedded in a Euclidean space $\mathbb{R}^D$ |
| $d$ | The (intrinsic) dimension of the manifold or a stratified space $M$. |
| $D$ | The ambient dimension of the Euclidean space $\mathbb{R}^D$ in which $M$ is embedded. |
| $\mu$ | A Borel probability measure on $\mathbb{R}^D$ whose support is contained in $M$. |
| $\mathbf{X}$ | i.i.d. sample drawn from a measure $\mu$ over $M$. |
| $n$ or $m$ | Number of sample points in $\mathbf{X}$. |
| $r$ | Local radius used to isolate points of $\mathbf{X}_n$ inside a ball of radius $r$. |

We make note of some acronyms we use throughout the thesis.

- MMD: Maximum mean discrepancy
- PCA: Principal components analysis

Chapters 2, 3, 4 provide preliminary material for the main results of the thesis, which are in Chapters 5, 6, 7, 8. To aid the reader's understanding, we will mark descriptions of the preliminary results in the context of the whole thesis with the symbol (★) .

# Chapter 2

# Differential geometry

We remind the reader that descriptions of the preliminary results in the context of the whole thesis will be marked with the symbol (★) .

## 2.1 Geometry of embedded manifolds

We prove some results in differential geometry that are relevant to us. Most of the results concern metric properties of manifolds smoothly embedded in a higher-dimensional Euclidean space.

We introduce the following notation for an open ball:

**Definition 2.1.1.** For $x \in \mathbb{R}^D$ and $r > 0$, the open ball of radius $r$ centred at $x$ is denoted as follows:

$$\mathcal{B}_r(x) = \mathcal{B}(x, r) = \{y \mid \mathrm{d}(x, y) < r\}$$

We introduce the notion of *reach* $\tau$. Its significance in this thesis is that $1/\tau$ is an upper bound for acceleration of any geodesic of a compact Riemannian manifold embedded in a Euclidean space.

**Definition 2.1.2.** Given a set $A \subseteq \mathbb{R}^D$, its reach $\tau_A$ is defined as follows:

$$\tau_A = \sup \left\{ t \;\middle|\; \text{If } \mathrm{d}(x, A) < t, \text{ then there is a unique } x' \in A \text{ such that } \mathrm{d}(x, x') = \mathrm{d}(x, A). \right\}$$

**Proposition 2.1.3** (Proposition 6.1, [72])**.** *Let $M$ be a smooth compact manifold embedded in $\mathbb{R}^D$ with reach $\tau$. If $\gamma : [0, 1] \to M$ is a geodesic on $M$, then the acceleration of $\gamma$ is bounded above by $1/\tau$.*

(★) The following is a simple extension of Proposition 6.3 of [72], and it controls the deviation of geodesic from the first order approximation. It is used in Proposition 5.3.4, and accounts for most of the ways that Riemannian curvature appears in this thesis.

**Lemma 2.1.4.** *[Proposition 6.3, [72]] Let $M$ be a smooth compact manifold embedded in $\mathbb{R}^D$ with reach $\tau$. Suppose that $x, y$ are connected by a (unit speed) geodesic $\gamma : [0, \tilde{r}] \to M$ of length $\tilde{r}$ with $\gamma(0) = x, \gamma(\tilde{r}) = y$, and denote $r = \|x-y\|$. Then the following inequalities hold:*

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

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

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

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

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

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

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

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

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

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

and thus

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

where the last inequality holds because for any $t$, $\|\ddot{\gamma}(t)\| \leq \tau^{-1}$ (the norm of the second fundamental form is bounded above by $\tau^{-1}$. See Proposition 6.1 of [72]).

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

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

□

The following Lemma concerns the nearest-distance projection map.

**Lemma 2.1.5.** *Let $M \subset \mathbb{R}^D$ be a compact set and let $\tau$ be its reach. Let $\pi_M$ be the projection map to $M$, such that for any $x \in \mathbb{R}^D$, $\pi_M(x)$ is the set of points on $M$ that minimises the distance to $M$. The following hold:*

1. *The distance function $x \mapsto \mathrm{d}(x, M) = \inf\{\|y - x\| \mid y \in M\}$ 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 $\mathrm{d}(-, M)$ is a Lipschitz function; we have that: $|\,\mathrm{d}(x, M) - \mathrm{d}(x', M)| \leq \|x - x'\|$.

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

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 $z$. □

This allows us to prove that a manifold is locally connected by short segments.

**Lemma 2.1.6.** *Let $M \subset \mathbb{R}^D$ be a compact path-connected set and let $\tau$ be its reach. If $x, y \in M$ satisfies $\|x - y\| < \tau$, then there exists a continuous path on $M$ that connects $(x, y)$ such that every point on the path has distance at most $\|x - y\|$ from both $x$ and $y$.*

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

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

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

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

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

(★) We derive bounds on Jacobian of the tangential projection map and also volume of manifold cut out by a ball. For the notion of *principal angle* appearing below, see Definition 5.4.1. The expression $\mathcal{H}^d(A)$ is the $d$-dimensional Hausdorff measure of a set $A$; see Definition 5.3.1. The following results appear in Proposition 5.3.4 and 3.1.7.

**Proposition 2.1.7.** *Let $M \subset \mathbb{R}^D$ be a $d$-dimensional submanifold. Let $\pi_x : \mathbb{R}^D \to T_xM$ be the projection map to $T_xM$, and let $\tilde{\pi}_x := \pi_x|_M : M \to T_xM$ 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 $T_xM$.*
3. *For any $y \in M$, the following bound holds:*

   $$\cos\theta_{x,y} \geq 1 - \frac{\mathrm{d}_M(x,y)}{\tau}$$

   *where $\theta_{x,y} = \measuredangle_{\max}(T_xM, T_yM)$ is the largest principal angle.*
4. *For any $y \in M$, the following bound holds:*

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

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

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

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 [73]. 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 $\mathrm{d}\,\tilde{\pi_x}(v) = A_x^\top v$ for each (embedded) tangent vector $v \in T_y M$.

(3) This is Proposition 6.2 from [73].

(4) The first bound folllows from (2) and the definition of principal angles. The second bound follows from (3) and Lemma 2.1.4 (Proposition 6.3, [72]), which implies $\mathrm{d}_M(x,y)/\tau \leq (r/\tau) + (r/\tau)^2 \leq \sqrt{2}r/\tau$.

(5) The lower bound is Lemma 5.3 from [73]. 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.

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

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 $(1 - \mathrm{d}_M(x,y)/\tau)^{-d}$ by (3) and (4). By the bound on geodesic length (Lemma 2.1.4, Proposition 6.3, [72])), we have $\mathrm{d}_M(x,y)/\tau \leq \rho + \rho^2$ and thus obtain the claim. □

**Corollary 2.1.8.** *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:

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

Their second derivatives are given by:

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

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. □

**Corollary 2.1.9.** *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 $c_4 = 2c_4'$. □

**Lemma 2.1.10.** *Let $f_0 : \mathbb{R}^d \to \mathbb{R}^+$ be a function such that $f_0(x) = f_0(\lambda x)$ for any $\lambda > 0$, and that $f_0$ is differentiable when restricted to the unit sphere $S^{d-1}$. Define the scaling map $f(x) = f_0(x)x$ for $x \neq 0$. Then the Jacobian determinant of $f$ is given by:*

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

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

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

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

(★) We introduce the notion of exponential map of a Riemannian manifold and derive a result on the norm of its derivative. This result can be used to obtain an alternative proof of Proposition 5.3.4, by replacing projection map to tangent space by the exponential map.

**Definition 2.1.11.** Let $M$ be a Riemannian manifold and let $x \in M$. Given a tangent vector $v \in T_x M$, let $\gamma_v$ be the unique geodesic such that $\gamma_v(0) = x$ and $\dot{\gamma}_v(0) = v$. Then for each $v$ such that $\gamma_v$ is defined on the interval $[0,1]$, the value of exponential map at $v$ is defined as:

$$\exp_x(v) = \gamma_v(1)$$

Sectional curvature may be used to bound the Jacobian of the exponential map, as follows[61]:

**Theorem 2.1.12.** *Let $M$ be a Riemannian manifold with sectional curvature bounded below and above by $\kappa_-$ and $\kappa_+$. Then for $x \in M$ and $v \in T_xM$, the following holds:*

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

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

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

**Corollary 2.1.13.** *Let $M \subseteq \mathbb{R}^D$ be a smoothly embedded compact Riemannian manifold with reach $\tau$. Then for $x \in M$ and $v \in T_xM$ satisfying $r := \|v\| \leq \pi\tau$, we have:*

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

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

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

*Proof.* Norm of the second fundamental form is bounded above by $\tau^{-1}$ [72], and thus by the Gauss equation applied to sectional curvature (i.e. $K(u,v) = \langle R(u,v)u, v\rangle = \langle \mathbb{I}(u,u), \mathbb{I}(v,v)\rangle - \|\mathbb{I}(u,v)\|^2$ for orthonormal $u, v$), we may take $\kappa_- = -2\tau^{-2}$ and $\kappa_+ = \tau^{-2}$ for the curvature bounds. Thus the radius condition reads $r \leq \pi\tau$. Then we have:

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

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

(★) We prove a result that allows us to work with one manifold at a time when dealing with a union of two manifolds. This Proposition is used in the main theorem of Chapter 6.

**Proposition 2.1.14.** *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*

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

$\dim(T_xM_1 \cap T_xM_2) = d_0$ *for a fixed* $d_0$*, and that principal angles of* $(T_xM_1, T_xM_2)$ *are all* $\geq \phi$*. Then we have:*

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

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

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

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

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

$$\frac{r}{\tau} < \min\left(\frac{\sqrt{2}-1}{2}, \frac{1-\cos\phi}{12}\right) \tag{2.1.1}$$

where $\tau = \min(\tau_1, \tau_2)$ and $\tau_i$ 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:

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

The second term is easily seen to be positive:

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

where the numerator is positive since $M_1 \backslash B$ is a compact set and $x \mapsto \mathrm{d}(x, M_1)$ 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 $\mathrm{d}(x, M_2)/\mathrm{d}(x, M_1 \cap M_2)$ when $x \in M_1 \cap B$. Denote $r_0 = \mathrm{d}(x, M_1 \cap M_2)$. 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 $r_0 = \mathrm{d}(x, x_0)$. 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 2.1.4 (Proposition 6.3, [72]) due Equation 2.1.1 with $\|x - x_0\| = r_0 < (\sqrt{2}-1)\tau$ and [3] $\|x_2 - x_0\| \leq 2r_0 < (\sqrt{2}-1)\tau$,

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

[3] 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 $\|x_2 - x_0\| \leq \|x_2 - x_1\| + \|x_1 - x_0\| \leq 2r_0$.

Therefore:

$$\begin{aligned}\mathrm{d}(x, M_2) &\geq \mathrm{d}(x_1, M_2) - \mathrm{d}(x, x_1) \\ &= \mathrm{d}(x_1, x_2) - \mathrm{d}(x, \pi_1) \\ &\geq \mathrm{d}(x_1, \pi_2) - \mathrm{d}(x_2, \pi_2) - \mathrm{d}(x, \pi_1) \\ &\geq \mathrm{d}(x_1, \pi_2) - 5\tau^{-1} r_0^2 \end{aligned} \tag{2.1.2}$$

**Linear algebra.** Now we are interested in controlling $\mathrm{d}(x_1, \pi_2)$, and this is an exercise of linear algebra since $x_1 \in \pi_1$. Write $x_1^\perp = \Pi(x_1, \pi_2)$ and $(z, z_\perp) = (x_1 - x_0, x_1^\perp - x_0)$, so that $\mathrm{d}(x_1, \pi_2) = \mathrm{d}(x_1, x_1^\perp) = \mathrm{d}(z, z_\perp)$. 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 $\pi_i'$. Then $z \in \pi_1'$ and $z_\perp = \Pi(z, \pi_2') = A_2 A_2^\top z$ and

$$\mathrm{d}(x_1, \pi_2) = \mathrm{d}(z, z_\perp)^2 = \|z\|^2 - \|z_\perp\|^2 \tag{2.1.3}$$

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 $(\pi_1, \pi_2)$ are $\geq \phi$, we get:

$$\|z_\perp\| = \|A_2 A_2^\top A_1 u\| = \|A_2^\top A_1 u\| \leq \|A_2^\top A_1\| \cdot \|u\| = \|A_2^\top A_1\| \cdot \|z\| \leq (\cos\phi) \cdot \|z\| \tag{2.1.4}$$

and we also note that:

$$\|z\| = \mathrm{d}(x_1, x_0) \geq \mathrm{d}(x, x_0) - \mathrm{d}(x, x_1) = \mathrm{d}(x, x_0) - \mathrm{d}(x, \pi_1) \geq r_0 - r_0^2/\tau \tag{2.1.5}$$

**Combining the bound.** We plug Equations (2.1.3), (2.1.4), (2.1.5) into Equation (2.1.2):

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

where in the last equality, we used the assumption $r_0/\tau \leq (1-\cos\phi)/12$ in Equation (2.1.1). Therefore, recalling that $r_0 = \mathrm{d}(x, M_1 \cap M_2)$, the following holds for all $x \in M_1 \backslash B$:

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

and the claim is proven. □

## 2.2 Calculus

(★) We also note some simple calculations we will use later in estimating various quantities and simplifying expressions. Except the last Proposition in this section, the following results can be safely regarded as auxiliary calculation tools.

**Lemma 2.2.1.** *For every $t > 0$ and $s > 1$, the following hold:*

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

*Furthermore, both functions are increasing.*

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

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

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

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

or equivalently,

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

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

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

The derivative is positive iff:

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

which follows from the following:

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

□

**Lemma 2.2.2.** *Suppose $0 < c \leq 1, d \geq 1$ and $t \leq c/d$. Then we have the following linear bound:*

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

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

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

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

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

Take $t_0 = c/d$. Then:

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

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

**Lemma 2.2.3.** *Suppose $d \geq 1, t \in [0,1]$. Then*

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

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

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

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

**Lemma 2.2.4.** *Suppose $a, b, x$ are real where $b > 1$ and $x > e$. Then we have that*

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

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

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

□

(★) The following result describes the covariance matrix of a uniform distribution over a disk. It is crucial to understanding what Local PCA does at the limit of small local neighborhood of a manifold. As such, its usage is propagated into the main theorems of 5 and 6.

**Lemma 2.2.5.** *[Lemma 13, [12]] Given a $d$-dimensional subspace $\Pi$ of $\mathbb{R}^D$, the covariance matrix of the uniform distribution over the disk $\Pi \cap \mathcal{B}_1(0)$ is given by:*

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

*where $\mathrm{P}_\Pi$ is the $D \times D$ projection matrix to $\Pi$. Eigenvalues of this matrix are:*

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

*Proof.* Denote by $\Pi_{d,D}$ the $d$-dimensional subspace of $\mathbb{R}^D$ spanned by the first $d$ canonical basis vectors. The only nontrivial covariance between the marginals of $\mathcal{H}^d|_{\Pi \cap \mathcal{B}_1(0)}$ is:

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

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

$$\Sigma[\mathcal{H}^d|_{\Pi \cap \mathcal{B}_1(0)}] = \frac{1}{d+2} \begin{bmatrix} I_d & 0 \\ 0 & \mathbf{0}_{D-d} \end{bmatrix}$$

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

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

□

# Chapter 3

# Statistical distance

We study two notions of statistical distances in this chapter, the Wasserstein distance and the kernel MMD. A statistical distance assigns a real number to a pair of probability measures, used for quantifying how different two measures are. To aid the reader's understanding, we will mark descriptions of the preliminary results in the context of the whole thesis with the symbol (★) .

## 3.1 Wasserstein distance

Let $(M, \mathrm{d}_M)$ be a Polish metric space equipped with probability measures $\mu$ and $\nu$. For each $p \geq 1$, the *$p$-Wasserstein distance* between $\mu$ and $\nu$ is defined as:

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

where $\Pi(\mu, \nu)$ is the set of measures on $M \times M$ with marginals equal to $\mu$ and $\nu$. Note that whenever $1 \leq p \leq q$, we have $\mathrm{W}_p(\mu, \nu) \leq \mathrm{W}_q(\mu, \nu)$ by the power mean inequality.

(★) The following Lemmas will be the main tools used for Proposition 5.3.4, which is a key ingredient in proving the main theorems of Chapter 5.

**Lemma 3.1.1.** *Let $M$ be a Polish metric space with metric $\mathrm{d}_M$. Suppose $A, B \subseteq M$ are Borel measurable, with inclusion maps $\iota^A : A \hookrightarrow M, \iota^B : B \hookrightarrow M$. Suppose that there is a continuous bijection $f : A \to B$ with a $L \geq 0$ with $d_M(x, f(x)) < L$ for any $x$. Let $\mu$ be a Borel probability measure on $A$. Then for any $p \geq 1$, the Wasserstein distance between*

*pushforwards of $\mu$ and $f_*\mu$ along inclusions are bounded by $L$:*

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

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

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

□

**Lemma 3.1.2.** *Let $M$ be a Polish metric space with metric $\mathrm{d}_M$ and a finite diameter $L := \sup_{x,y\in M} \mathrm{d}_M(x, y)$. For a Borel probability measure $\mu$ on $M$ and a Dirac delta measure $\delta_x$ centered at $x \in M$, we have:*

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

*Proof.* Define the transportation plan $\nu$ on $M \times M$ by

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

whose marginals are $\mu$ and $\delta_x$. The transportation cost is bounded by $L$. □

**Lemma 3.1.3.** *Let $M$ be a Polish metric space with metric $\mathrm{d}_M$ and a finite diameter $L := \sup_{x,y\in M} \mathrm{d}_M(x, y)$. Fix a Borel probability measure $\mu$ on $M$. Let $f$ be a non-negative continuous function on $M$ with $\sup_{x\in M} f(x) - \inf_{x\in M} f(x) \leq C$ and $\int_M f(x) d\mu(x) = 1$. Let $\mu_f$ be the Borel probability measure on $M$ given by taking $f$ as the probability density function. Then for any $p \geq 1$,*

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

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

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

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

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

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

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

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

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

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

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

(★) In the following two Lemmas, we study behaviour of the Wasserstein distance under projection maps. In our context, projection maps arise in the singularity detection algorithm (Chapter 6, 7). The algorithm works by locally projection data points to a lower-dimensional subspace and then measuring how uniformly the data points spread over a disk. Therefore, stability of the Wasserstein distances under projection map is a necessary ingredient to show that the singularity score is also stable under small perturbation of data points (Proposition 6.4.1).

**Lemma 3.1.4.** *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:*

$$\mathrm{W}(\mu_1', \mu_2') \leq \mathrm{W}(\mu_1, \mu_2)$$

*Proof.* For every transportation plan from $\mu_1$ to $\mu_2$, 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 $p_\pi$ the orthogonal projection map to $\pi$. 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 $\mu_{12}$, a coupling of $\mu_1$ and $\mu_2$, 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 $\mu_2$. Now,

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

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

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

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

**Lemma 3.1.5.** *Let $\mu \in \mathcal{P}(\mathbb{R}^D)$ and $\pi_1, \pi_2 \in \mathrm{Gr}(k, D)$. Assume that the support of $\mu$ is bounded in the ball of radius 1, centered at the origin. Denoting $\mu_i = \Pi(\mu, \pi_i)$, we have:*

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

*where $(\theta_1, \ldots \theta_d)$ are the principal angles between $(\pi_1, \pi_2)$.*

*Proof.* Denote the orthogonal projection map to $\pi_i$ by $p_i$. We define a coupling $\mu_{12}$ of $(\mu_1, \mu_2)$:

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

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 $\mu_2$.

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 $\theta = \measuredangle(\pi_1, \pi_2)$. We claim that $\mu_{12}|_{S'} \equiv 0$ and $S_x \subseteq \mathcal{B}(x, \sin\theta)$. The proposition follows from these assumptions:

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

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 $\mu$ 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 5.4.4, we obtain an orthonormal basis $\{u_1, \dots u_{d_0}\} \cup \{v_1, w_1, \dots v_{d-d_0}, w_{d-d_0}\}$ of $\mathrm{span}(\pi_1, \pi_2)$ so that:

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

with $(\theta_1, \dots \theta_{d-d_0})$ being the nonzero principal angles of $(\pi_1, \pi_2)$. 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 $\Pi(z, u) = \Pi(z_i, u)$ [1]. This gives an orthogonal decomposition:

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

Therefore,

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

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

[1] 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.

(★) We define the following notions of localised measures. These are heavily featured in Chapter 6.

**Definition 3.1.6.** Let $\mu$ be a Borel probability measure on $\mathbb{R}^D$. For $x \in \text{supp}(\mu)$ and $r > 0$, define the following:

$$\mu_{x,r} := g_{x,r}(\mu|_{\mathcal{B}(x,r)})$$

where $g_{x,r}(\nu)$ is the pushforward of the measure $\nu$ along the affine linear map $y \mapsto r^{-1}(y - x)$. Furthermore, define the following limit, if it exists:

$$\mu_{x,0} := \lim_{r \to 0} \mu_{x,r}$$

where convergence is measured using the Wasserstein distance. In other words, $\mu_{x,0}$ is the unique measure satisfying $\lim_{r\to 0} \mathrm{W}(\mu_{x,r}, \mu_{x,0}) = 0$.

(★) We bound the change of localised measure under moving around the base point. This is one of the key ingredients of the main theorem in 6.

**Proposition 3.1.7.** *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 $\tau_i$ 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:*

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

*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:

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

**Step 1.** We redistribute mass equally for two manifolds. For $i = 1, 2$, define $\mu^{(i)}_{x,r}$ 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:

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

Define the following:

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

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

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

where the first inequality follows from the triangle inequality and the second inequality is due to the fact that the support of $\mu$ 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 $u \cdot \mu^{(1)}_{x,r}$ to the origin and back to $u \cdot \mu^{(2)}_{x,r}$, 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:

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

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

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

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

We are interested in bounding $\mathrm{W}(\nu_x, \nu_y) = \mathrm{W}(\mu^{(1)}_{x,r}, \mu^{(1)}_{y,r})$. 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 $\nu_x$ 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 $\nu_y$. We'd like to compare the pairs $(\nu'_x, \nu'_y)$ and $(\nu''_x, \nu''_y)$ separately, but the ratios $\beta_x/\gamma$ and $\beta_y/\gamma$ are different. To match the ratios, define $\nu_y^\dagger$ as the linear combination of $(\nu'_y, \nu''_y)$ that has the same ratio as that of $(\nu'_x, \nu''_x)$ in $\nu_x$:

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

Now we construct a transportation plan from $\nu_y$ to $\nu_y^\dagger$; they are both linear combinations of $\nu'_y$ and $\nu''_y$ with different ratios. If $\beta_x \geq \beta_y$, then we transport $\nu_y$ 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,

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

Therefore,

$$\begin{aligned}
\mathrm{W}(\nu_x, \nu_y) \leq& \mathrm{W}(\nu_x, \nu_y^\dagger) + \mathrm{W}(\nu_y^\dagger, \nu_y) \\
=& \mathrm{W}\left( \frac{\beta_x \nu'_x + \gamma \nu''_x}{\beta_x + \gamma}, \frac{\beta_x \nu'_y + \gamma \nu''_y}{\beta_x + \gamma} \right) + \mathrm{W}(\nu_y^\dagger, \nu_y) \\
\leq& \frac{\beta_x \, \mathrm{W}(\nu'_x, \nu'_y) + \gamma \, \mathrm{W}(\nu''_x, \nu''_y)}{\beta_x + \gamma} + 2 \cdot \left| \frac{\alpha_y}{\alpha_x} - 1 \right|
\end{aligned}$$

**Step 3.** At this point our task is reduced to bounding both $\mathrm{W}(\nu'_x, \nu'_y)$ and $\mathrm{W}(\nu''_x, \nu''_y)$. We simply relocate all mass of $\nu'_x$ to the origin and bring it back to $\nu'_y$, so that we use the trivial bound $\mathrm{W}(\nu'_x, \nu'_y) \leq 2$. To bound $\mathrm{W}(\nu''_x, \nu''_y)$, 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,

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

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

$$\frac{\beta_x \, \mathrm{W}(\nu'_x, \nu'_y) + \gamma \, \mathrm{W}(\nu''_x, \nu''_y)}{\beta_x + \gamma} \leq \frac{\beta_x \cdot 2 + \gamma \cdot s}{\beta_x + \gamma} \leq \frac{2\beta_x}{\alpha_x} + s$$

This thus shows that:

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

**Total bound.** We now collect the terms from above and bound them using $s$ and $r$. The main tool here is Corollary 2.1.9, 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:

$$\begin{aligned}
\mathrm{W}(\mu_{x,r}, \mu_{y,r}) &\leq E_1 + E_2 + E_3 \\
E_1 &= \frac{|\alpha_{1,x} - \alpha_{2,x}|}{\alpha_{1,x} + \alpha_{2,x}} + \frac{|\alpha_{1,y} - \alpha_{2,y}|}{\alpha_{1,y} + \alpha_{2,y}} \\
E_2 &= \frac{|\alpha_{1,y} - \alpha_{1,x}|}{\alpha_{1,x}} + \frac{|\alpha_{2,y} - \alpha_{2,x}|}{\alpha_{2,x}} \\
E_3 &= \frac{\beta_{1,x}}{\alpha_{1,x}} + \frac{\beta_{2,x}}{\alpha_{2,x}} + s
\end{aligned}$$

where $\beta_{i,x} = \mathcal{H}^d(M_i \cap B_x \backslash B_y)$. Corollary 2.1.9 implies that for $i = 1, 2$, we have:

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

where $\rho = r/\tau$ [2]. Here we used the fact that $\mathrm{d}(y, M_i) \leq \|x - y\| \leq s \cdot r$. To work with $\beta_x$, 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 2.1.9 again implies:

$$\begin{aligned}
\beta_{i,x} \leq& \omega_d r^d(1 + c_4\rho) - \omega_d(r - sr)^d(1 - c_4(r - sr)/\tau) \\
\leq& \omega_d r^d\Big(1 + c_4\rho - (1 - s)^d(1 - c_4\rho)\Big) \\
\leq& \omega_d r^d\Big(1 - (1 - s)^d + 2c_4\rho\Big) \\
\leq& \omega_d r^d(d \cdot s + 2c_4\rho)
\end{aligned}$$

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

where in the last inequality we used the fact that $(1-t)^d \geq 1 - d \cdot t$ for $t \in [0,1]$ and $d \geq 1$ [3] Therefore, we see that:

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

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

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

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. □

(★) The following is a simplified version of Proposition 5.3.4, which is simply *stated* here for reference. This simplified version will be cited in Chapter 6.

**Proposition 3.1.8.** *Let $\mu$ 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*[4] *$T_xM \cap \mathcal{B}(0,1)$. Furthermore, we have the following bound:*

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

Using the above and repeating the mass-redistribution argument in Step 1 of the proof in Proposition 3.1.7 verbatim, we obtain the following:

**Proposition 3.1.9.** *Let $\mu$ 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*

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

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

*uniform distribution over* $(T_xM_1 \cup T_xM_2) \cap \mathcal{B}(0,1)$. *Furthermore, we have the following bound:*

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

## 3.2 Kernel distance

We briefly outline a basic theory of kernels and derive some results for our usage. For a standard reference for kernel theory, see [81].

**Definition 3.2.1.** Given a set $\mathcal{X}$, a *kernel* $\kappa$ is a symmetric function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ such that for every $x_1, \dots x_n \in \mathcal{X}$, the Gram matrix $G = (\kappa(x_i, x_j))_{1 \leq i,j \leq n}$ is positive semidefinite.

Each kernel corresponds bijectively with a Hilbert space of functions, and this canonical construction is crucial in many applications. We first define the class of Hilbert spaces we are interested in:

**Definition 3.2.2.** A *reproducing kernel Hilbert space* (RKHS) is a set of functions $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ which satisfies the following:

1. $\mathcal{H}$ is complete under an inner product, making it a real Hilbert space.
2. (Reproducing property) For any $f \in \mathcal{H}$ and $x \in \mathcal{X}$, $\langle f, \kappa(x, -) \rangle = f(x)$.
3. The set $\mathcal{H}' = \{a_1\kappa(x_1, -) + \cdots + a_k\kappa(x_k, -) \mid a_i \in \mathbb{R}\}$ is dense in $\mathcal{H}$.

The bijective correspondence between kernels and RKHS is described as follows.

**Proposition 3.2.3** (Moore-Aronszajn)**.** *Given a set* $\mathcal{X}$*, there is a bijection:*

$$\Psi : \{\textit{Kernels on } \mathcal{X}\} \to \{\textit{RKHS on } \mathcal{X}\}$$

*Denoting* $\Phi_x := \kappa(x, -) \in \mathcal{H}$*, the map* $\Psi$ *and* $\Psi^{-1}$ *are given as follows:*

$$\Psi(\kappa) = \textit{Completion of } \left\{ \sum_{j=1}^{N} a_j \Phi_{x_j} \right\} \textit{ under the inner product } \langle \Phi_x, \Phi_y \rangle = \kappa(x, y)$$

$$\Psi^{-1}(\mathcal{H}) = \textit{Unique } \kappa \textit{ such that for any } x \in \mathcal{X}, f \in \mathcal{H}, \langle \Phi_x, f \rangle = f(x)$$

Note that $\Psi^{-1}(\mathcal{H})$ above is well-defined by the Riesz representation theorem, applied to the evaluation functional $\mathrm{ev}_x = (f \mapsto f(x))$.

The RKHS associated to each kernel furnishes a natural and unique domain for performing linear algebra. This becomes a versatile conceptual framework for organising kernel-based calculations in terms of linear algebra in a Hilbert space. For example, the *kernel principal components analysis* performs the principal components analysis of data points $\{x_1, \dots x_n\} \subset \mathcal{X}$ in the embedded points $\{\Phi_{x_1}, \dots \Phi_{x_n}\} \subset \mathcal{H}$, where $\Phi_x = \kappa(x, -)$. Note that by pulling back the metric of $\mathcal{H}$, we also get a metric on $\mathcal{X}$, assuming $\Phi$ is injective. In fact the construction $x \mapsto \Phi_x$ can be made more general, by promoting each point $x \in \mathcal{X}$ into a Dirac-delta measure $\delta_x$, and defining an element of RKHS associated to each *measure*.

**Definition 3.2.4.** 1. Given a probability measure $\mu$, its *kernel mean embedding* is defined as:

$$\Phi_\mu := \int \kappa(x, -) \,\mathrm{d}\, \mu(x) \in \mathcal{H}_\kappa$$

2. The *kernel MMD* (*mean maximum discrepancy*) between two probability measures $\mu, \nu$ is then defined as the real number

$$\Delta_\kappa(\mu, \nu) := \|\Phi_\mu - \Phi_\nu\|_\kappa$$

3. A kernel $\kappa$ is said to be *characteristic* if the mapping $\mu \mapsto \Phi_\mu$ is an injective function, so that $\Delta_\kappa(\mu, \nu) > 0$ implies $\mu \neq \nu$.

The following are two formulas of MMD that follow straightforwardly from its definition:

**Lemma 3.2.5.** *Given a kernel $\kappa$, the following formulas hold for its MMD (maximum mean discrepancy):*

$$\begin{aligned}\Delta^2_\kappa(\mu, \nu) &= \iint \kappa(x, x') \,\mathrm{d}\, \mu(x) \,\mathrm{d}\, \mu(x') + \iint \kappa(y, y') \,\mathrm{d}\, \nu(y) \,\mathrm{d}\, \nu(y') - 2 \iint \kappa(x, y) \,\mathrm{d}\, \mu(x) \,\mathrm{d}\, \nu(y) \\ &= \sup_{\|f\|_\kappa \leq 1} \left| \mathbb{E}_{X \sim \mu} f(X) - \mathbb{E}_{Y \sim \nu} f(Y) \right|^2\end{aligned}$$

*where $\|\cdot\|_\kappa$ is the RKHS norm of the kernel $\kappa$. The second expression is known as the integral probability metric expression.*

The triangle inequality holds for MMD, as does it for any integral probability metric:

$$\begin{aligned}\sup_f |\mathbb{E}f(X) - \mathbb{E}f(Z)| \leq &\sup_f \Big( |\mathbb{E}f(X) - \mathbb{E}f(Y)| + |\mathbb{E}f(Y) - \mathbb{E}f(Z)| \Big) \\ \leq &\sup_f |\mathbb{E}f(X) - \mathbb{E}f(Y)| + \sup_f |\mathbb{E}f(Y) - \mathbb{E}f(Z)|\end{aligned}$$

Two important classes of characteristic kernels on $\mathcal{X} = \mathbb{R}^D$ include the *radial basis function kernel*, given by $\kappa(x, y) = \exp(-\gamma \cdot \|x - y\|^2)$ for $\gamma > 0$, and the *dot product kernel*, given by $\kappa(x, y) = \sum_{m \geq 0} a_m \langle x, y \rangle^m$ with $a_m > 0$.

(★) The MMD for the Gaussian kernel can be controlled with the Wasserstein distance with a Lipschitz continuity relation. This is used in Chapter 6 to replace bounds on Wasserstein distance with bounds on MMD.

**Lemma 3.2.6.** *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 $\mu, \nu$ valued in $\mathbb{R}^D$ with finite first moment, we have:*

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

*Proof.*

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

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

(★) The rest of this section is dedicated to proving the following theorem, which is a key formula used for the singularity detection algorithm in Chapter 7.

**Theorem 3.2.7.** *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 $\kappa$ 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.*

We start with the following proposition.

**Proposition 3.2.8.** *Suppose $\kappa$ is a kernel on $\mathbb{R}^d$. Let $U_d = \omega_d^{-1}\mathcal{H}^d|_{\mathbb{D}_d}$ be the uniform distirbution 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*

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

*Proof.* We separately evaluate the terms:

$$\begin{aligned}
\iint \kappa(x,y)\,\mathrm{d}\,U_d(x)\,\mathrm{d}\,U_d(y) &= \frac{1}{\omega_d^2}\iint_{\|x\|\le 1,\|y\|\le 1}\kappa(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y = \frac{1}{\omega_d^2}\int_{\|x\|\le 1}F_{\kappa,d}(x)\,\mathrm{d}\,x\\
\iint \kappa(x,y)\,\mathrm{d}\,\hat{\mu}_n(x)\,\mathrm{d}\,\hat{\mu}_n(y) &= \frac{1}{n^2}\sum_{i,j}\kappa(x_i,x_j)\\
\iint \kappa(x,y)\,\mathrm{d}\,\hat{\mu}_n(x)\,\mathrm{d}\,U_d(y) &= \frac{1}{n\omega_d}\sum_{i=1}^{n}F_{\kappa,d}(x_i)
\end{aligned}$$

□

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 3.2.9.** *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:*

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

*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:

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

□

We also note the simple expression for the expected value.

**Proposition 3.2.10.** *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.*

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

□

**Proposition 3.2.11.** *For $k \geq 0$, let $\kappa(x, y) = \langle x, y \rangle^k$ be a monomial kernel. Then we have that:*

$$\begin{aligned} F_{\kappa,d}(r) =& \omega_{d-1} \,\mathrm{B}\left( \frac{k+1}{2}, \frac{d+1}{2} \right) = \frac{\pi^{(d-1)/2} \Gamma(\frac{k+1}{2})}{\Gamma(\frac{k+d}{2} + 1)} \cdot r^k && \text{, if } k \text{ is even} \\ \bar{F}_{\kappa,d} =& \frac{d}{d+k} \omega_d \omega_{d-1} \,\mathrm{B}\left( \frac{k+1}{2}, \frac{d+1}{2} \right) = \frac{2^d \pi^{d-1}}{(d-1)!} \cdot \frac{1}{d+k} \,\mathrm{B}\left( \frac{k+1}{2}, \frac{d+1}{2} \right) && \text{, if } k \text{ is even} \end{aligned}$$

*and both expressions are zero if $k$ is odd.*

*Proof.* We directly evaluate:

$$\begin{aligned} F_{\kappa,d}(r) =& \int_{\|y\| \leq 1} \kappa(r \cdot e_1, y)^k \,\mathrm{d}\, y \\ =& \int_{-1}^{+1} \int_{\|z\| \leq \sqrt{1-s^2}, z \in \mathbb{R}^{d-1}} (rs)^k \,\mathrm{d}\, z \,\mathrm{d}\, s \\ =& \omega_{d-1} r^k \int_{-1}^{+1} s^k (1 - s^2)^{(d-1)/2} \,\mathrm{d}\, s \end{aligned}$$

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

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

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

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

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

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

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:

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

□

**Proposition 3.2.12.** *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:*

$$\begin{aligned}
F_{\kappa,d}(r) =& \omega_{d-1} \sum_{k=0}^{\infty} \mathrm{B}\left(k+\frac{1}{2}, \frac{d+1}{2}\right) a_{2k} r^{2k} \\
=& \pi^{(d-1)/2} \sum_{k=0}^{\infty} \frac{\Gamma(k+\frac{1}{2})}{\Gamma(k+\frac{d}{2}+1)} \cdot a_{2k} r^{2k} \\
\bar{F}_{\kappa,d} =& d\omega_d \omega_{d-1} \sum_{k=0}^{\infty} \mathrm{B}\left(k+\frac{1}{2}, \frac{d+1}{2}\right) \frac{a_{2k}}{d+2k} \\
=& \frac{2^d \pi^{d-1}}{(d-1)!} \sum_{k=0}^{\infty} \mathrm{B}\left(k+\frac{1}{2}, \frac{d+1}{2}\right) \frac{a_{2k}}{d+2k}
\end{aligned}$$

*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 $a_k \geq 0$. □

We now return to finishing the proof of the main result of this section.

*Proof of Theorem 3.2.7.* The claim can be restated as $\Delta_\kappa(\hat{\mu}_n, U_d) = A + B - C$, where

$$\begin{aligned}
A =& \frac{1}{n^2} \sum_{i,j} \kappa(x_i, x_j) \\
B =& \frac{\Gamma(\frac{d}{2}+1)}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{\Gamma(k+\frac{1}{2}) a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \frac{d}{d+2k} \\
C =& \frac{2}{n} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\sqrt{\pi}} \sum_{k=1}^{n} \frac{\Gamma(k+\frac{1}{2}) a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \sum_{i=1}^{n} \|x_i\|^{2k}
\end{aligned}$$

We firstly know that:

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

$$\begin{aligned}
A =& \frac{1}{n^2} \sum_{i,j} \kappa(x_i, x_j) \\
B =& \frac{\omega_{d-1}}{\omega_d} \sum_{k=0}^{\infty} \frac{d}{d+2k} \mathrm{B}\left(k+\frac{1}{2}, \frac{d+1}{2}\right) a_{2k} \\
C =& \frac{2}{n} \frac{\omega_{d-1}}{\omega_d} \sum_{k=1}^{n} \mathrm{B}\left(k+\frac{1}{2}, \frac{d+1}{2}\right) a_{2k} \sum_{i=1}^{n} \|x_i\|^{2k}
\end{aligned}$$

We compute:

$$\begin{aligned}
B =& \frac{\omega_{d-1}}{\omega_d} \sum_{k=0}^{\infty} \frac{d}{d+2k} \mathrm{B}\left(k+\frac{1}{2}, \frac{d+1}{2}\right) a_{2k} \\
=& \pi^{-1/2} \frac{\Gamma(\frac{d+2}{2})}{\Gamma(\frac{d+1}{2})} \sum_{k=0}^{\infty} \frac{d}{d+2k} \frac{\Gamma(k+\frac{1}{2})\Gamma(\frac{d+1}{2})}{\Gamma(k+\frac{d}{2}+1)} a_{2k} \\
=& \pi^{-1/2} \Gamma\left(\frac{d}{2}+1\right) \sum_{k=0}^{\infty} \frac{d}{d+2k} \frac{\Gamma(k+\frac{1}{2}) a_{2k}}{\Gamma(k+\frac{d}{2}+1)}
\end{aligned}$$

and

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

We now estimate the error rate of the expression in Theorem 3.2.7, upon taking only finite sum.

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

$$\begin{aligned}
B =& \frac{\Gamma(\frac{d}{2}+1)}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{\Gamma(k+\frac{1}{2}) a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \frac{d}{d+2k} \\
C =& \frac{2}{n} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\sqrt{\pi}} \sum_{k=1}^{n} \frac{\Gamma(k+\frac{1}{2}) a_{2k}}{\Gamma(k+\frac{d}{2}+1)} \sum_{i=1}^{n} \|x_i\|^{2k}
\end{aligned}$$

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 $\leq 0.03$ and $\gamma \leq 0.5$ gives relative error $\leq 10^{-6}$.

# Chapter 4

# Concentration inequality

Concentration inequalities allow us to control a probabilistic quantity. It generally takes the following form: For each $n$, let $f_n(X_1, \dots X_n)$ be a function of random variables. Then for every $\epsilon > 0$, the following holds:

$$\mathbb{P}\left[ \left| f_n(X_1, \dots X_n) - \mathbb{E}f_n \right| \geq \epsilon \right] \leq g(n, \epsilon)$$

where $\lim_{n\to\infty} g(n, \epsilon) = 0$. This allows a precise quantification of error. Specifically for us, we are interested in estimating the covariance matrix and the Wasserstein distance.

Most of the probability theory appearing in this thesis is contained in this chapter. The nontrivial task for us is to mold the standard results in controlling *global* estimation into *simultaneous local* estimation. These are Propositions 4.1.6 and 4.2.7, the main results of this chapter.

## 4.1 Covariance matrix

(★) The main result of this section is Proposition 4.1.6, where we establish bounds for local covariance estimation. It is used to prove the main technical theorem of Chapter 5, which is Theorem 5.5.3.

Our main tool is the *matrix Hoeffding inequality* [92, Theorem 1.3][1]. Here onwards, we will use $\|A\|$ to denote the operator norm of a given matrix $A$: $\|A\| := \sup_{\|x\|=1} \|Ax\|$.

[1]Our version of the matrix Hoeffding inequality follows from the one in [92] by noting that for any matrix $A$, the operator norm $\|A\|$ equals $\max(\lambda_{\max}(A), \lambda_{\max}(-A))$ where $\lambda_{\max}$ denotes the largest eigenvalue. And moreover, $\|A\| \leq \alpha$ implies that $\alpha^2 \cdot \mathrm{Id} - A^2$ is positive definite.

**Theorem 4.1.1** (Matrix Hoeffding)**.** *Let $Y_1, \dots Y_m$ be independent Hermitian random $D \times D$ matrices so that for each $i$ we have both $\mathbb{E}Y_i = 0$ and $\|Y_i\| \leq \alpha_i$ for some real number $\alpha_i \geq 0$. Write $\sigma^2 = \sum_{i=1}^m \alpha_k^2$. Then for every $\epsilon \geq 0$,*

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

This inequality can be used to establish concentration of vectors.[2]

**Corollary 4.1.2.** *Let $X_1, \dots X_m$ be independent random vectors in $\mathbb{R}^D$ satisfying $\mathbb{E}X_i = 0$, and $\|X_i\| \leq \alpha_i$ for some real number $\alpha_i$. Write $\sigma^2 = \sum_1^m \alpha_i^2$. Then for every $\epsilon \geq 0$,*

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

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

**Definition 4.1.3.** Given $X \sim \mu$, the *covariance matrix* of $\mu$ is the following $D \times D$ matrix:

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

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

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

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

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

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

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

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

**Proposition 4.1.4** (Concentration inequalities for covariance)**.** *Let $\mu$ be a Borel probability measure on $\mathbb{R}^D$ and let $\boldsymbol{X} = (X_1, \dots X_m)$ be an i.i.d. sample drawn from $\mu$. Suppose that the support of $\mu$ is contained in a ball of radius $r$. Then for each $\epsilon \geq 0$,*

$$\mathbb{P}\big(\|\hat{\Sigma}_0 - \Sigma[\mu]\| \geq \epsilon\big) \leq 2D \cdot \exp\left(-\frac{m\epsilon^2}{512r^4}\right)$$
$$\mathbb{P}\big(\|\hat{\Sigma} - \Sigma[\mu]\| \geq \epsilon\big) \leq (4D+2) \cdot \exp\left(-\frac{m\epsilon^2}{1152r^4}\right)$$

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

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

*Proof.* We may assume that $r = 1$ without loss of generality, since for general $r$ we know that $r^2\Sigma$ is the covariance of $r \cdot X$ for all $X \sim \mu$. Thus, we have $\|X - \mathbb{E}X\| \leq 2$ by the triangle inequality and the constraint on the support of $\mu$. The bound for $\hat{\Sigma}_0$ is obtained directly by applying the matrix Hoeffding inequality from Theorem 4.1 as follows. Writing $\Sigma[\mu] = \Sigma$, set $Y_i = \frac{1}{m}((X_i - \mathbb{E}X)(X_i - \mathbb{E}X)^\top - \Sigma)$. Then $\|Y_i\| \leq (4+4)/m$ and $\sigma^2 = m \cdot (8/m)^2 = 64/m$. Since $\hat{\Sigma}_0 = \hat{\Sigma} + (\bar{X} - \mathbb{E}X)(\bar{X} - \mathbb{E}X)^\top$, we have

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

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

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

In the last inequality, we used the bound for $\hat{\Sigma}_0$ as well as Corollary 4.1.2, with $\sigma^2 = 4$. Choosing $\alpha = 2/3$ to make the exponents equal, we obtain the second bound. □

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

**Proposition 4.1.5.** *Let $\mathbf{X} = (X_1, \dots X_m)$ be an i.i.d. sample drawn from $\mu$ and let $U \subseteq \mathbb{R}^D$ be a Borel set which is contained in a ball of radius $r$. Denote by $\hat{\Sigma}_U$ the covariance $\Sigma[\delta_{\mathbf{X}}|_U]$, and similarly write $\Sigma_U = \Sigma[\mu|_U]$. Then for any error level $\epsilon > 0$, we have that $\hat{\Sigma}_U$ estimates $\Sigma_U$:*

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

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

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

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

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

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

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

Here Proposition 4.1.4 was applied in the only inequality above. Note that the possibility $\mathcal{S}_\emptyset$ is correctly accounted for since we included $k=0$ when indexing the sum in the second line above. □

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

**Proposition 4.1.6.** *Let $\mu$ be a Borel measure supported on a compact subset $K \subset \mathbb{R}^D$, and let $\mathbf{X} = (X_1, \dots X_m)$ be a $\mu$-i.i.d. sample. Given a radius $r > 0$, consider for $1 \leq i \leq m$ the covariances $\hat{\Sigma}_i := \Sigma[\delta_{\mathbf{X}_i}|_{U_i}]$ and $\Sigma_i = \Sigma[\mu|_{U_i}]$, where $\mathbf{X}_i = \{X_j | j \neq i\}$ and $U_i = \mathcal{B}_r(X_i)$. Let $\epsilon, \delta, \varrho > 0$ where we assume[3] that $\epsilon \leq 2r^2$. Then the following holds:*

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

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

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

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

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

where $\mathbf{x}_i = \{x_j | j \neq i\}$. By the union bound, symmetry, and Proposition 4.1.5, we then have:

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

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

$$\int (4D+2)k(1 - u_x(1-\xi))^{m-1} \,\mathrm{d}\,\mu(x) \leq (4D+2)k(1 - u_0(1-\xi))^{m-1} \tag{4.1.1}$$

Letting right hand side of (4.1.1) to be $\leq \delta$, we get the condition:

$$\begin{aligned}
&(4D+2)k(1 - u_0(1-\xi))^{m-1} \leq \delta \\
\iff &\frac{-1}{\log(1 - u_0(1-\xi))} \cdot \log\left(\frac{(4D+2)k}{\delta}\right) \leq m - 1
\end{aligned} \tag{4.1.2}$$

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

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

where the first inequality is due to Lemma 2.2.1, and the second inequality follows from the assumption that $\epsilon^2 \leq 4r^4$.[4] Using the fact that $\log((4D+2)/\delta) \geq 2$ and Lemma 2.2.4, we obtain the claimed sufficient condition for (4.1.2):

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

To establish that $u_0 > 0$, consider the covering of $K$ by balls of radius $r/2$. Since $K$ is compact, it admits a subcover $\{\mathcal{B}_{r/2}(x) \mid x \in J\}$, with $J$ a finite set. Thus, every $x \in K$

[4] By similar reasoning, the left hand side of (4.1.2) is at least $\frac{1}{u_0}(1150r^4/\epsilon^2)$, so that this sufficient condition doesn't weaken the bound much.

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

## 4.2 Wasserstein distance

(★) 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 4.2.7, which is derived by simplifying Corollary 1.2 from [25]. This is used to prove the main theorem of Chapter 6, which is Theorem 6.1.1. To see it in action, refer to the last section of Chapter 6.

We use the notion of *covering number* for this:

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

**Theorem 4.2.1** (Boissard-Le Gouic)**.** *Let $(M, d, \mu)$ 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 $0 < r < R/4$:*

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

*Then the following holds:*

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

To apply this to compact subsets of a Euclidean space, we use a lemma from [14]:

**Lemma 4.2.2.** *The $D$-dimensional unit ball $\mathcal{B}_D$ satisfies:*

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

*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 $r$[5], and so we have $N_{\mathrm{cover}}(\mathcal{B}_D, r) \leq N'$. Now consider a maximal packing

[5]Suppose that balls of radii $r/2$ centered at $x_1, \dots x_{N'}$ 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 $x_1, \dots x_{N'}$, which means that balls of radii $r/2$ centered at $N' + 1$ points $\{y, x_1, \dots x_{N'}\}$ is also a packing. This contradicts maximality.

by balls of radii $r/2$ centered at $x_1, \dots x_{N'}$. Then,

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

□

**Corollary 4.2.3.** *Let $\mu$ 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:*

$$\mathbb{E}\Big[\, \mathrm{W}(\hat{\mu}_m, \mu)\Big] \leq \frac{c_1}{m^{1/D}} \leq \frac{c_2}{m^{1/D}}$$
$$\textit{where } c_1 = 32R \cdot \left(\frac{2}{D-2}\right)^{2/D},\ c_2 = 51R$$

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

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

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

(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:

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

The derivative of $f(t)$ has the same sign as $(t-1)\log(t-1) - t$[6], 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 $\leq 1$; 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. □

To obtain a concentration inequality, we use the Proposition A2 from [25]:

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

**Proposition 4.2.4.** *Let $(E, d, \mu)$ be a measured Polish metric space of a finite diameter $R$ and suppose that $\mu$ has a finite $p$-th moment. Then we have:*

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

Combining the above with Corollary 4.2.3, we obtain that:

**Proposition 4.2.5** (Global concentration)**.** *Let $\mu$ 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 $m \geq f(t)$:*

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

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

*Proof.* Follows directly by using Proposition 4.2.4 and letting the expected value of the Wasserstein distance be $\leq t/2$ in Corollary 4.2.3. □

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

**Proposition 4.2.6** (Local 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 $\mathbf{X}$. 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 $m \geq \max(N, 2u^{-1})$:*

$$\mathbb{P}\bigg( \mathrm{W}(\hat{\mu}_m|_U, \mu|_U) \geq t \bigg) \leq c \cdot m^N \gamma^m$$

*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

$$\begin{aligned}\mathbb{P}\bigg(\mathrm{W}(\hat{\mu}_m|_U, \mu|_U) \geq t\bigg) &= \sum_{I\subseteq\{1,\dots m\}} \mathbb{P}(\mathcal{S}_I)\cdot\mathbb{P}\bigg(\mathrm{W}(\hat{\mu}_m|_U, \mu|_U) \geq t \,\bigg|\, \mathcal{S}_I\bigg) \\ &= \sum_{I\subseteq\{1,\dots m\}} u^{|I|}(1-u)^{m-|I|}\mathbb{P}\bigg(\mathrm{W}(\hat{\mu}_m|_U, \mu|_U) \geq t \,\bigg|\, \mathcal{S}_I\bigg) \\ &= \sum_{k=0}^{m}\binom{m}{k}u^k(1-u)^{m-k}\mathbb{P}\bigg(\mathrm{W}(\hat{\mu}_m|_U, \mu|_U) \geq t \,\bigg|\, \mathcal{S}_{\{1,\dots k\}}\bigg)\end{aligned} \tag{4.2.1}$$

Now we apply Proposition 4.2.5 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 (4.2.1) is thus bounded by:

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

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

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

**Proposition 4.2.7** (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 $m \geq \max(N, 2/u_-)$:*

$$\mathbb{P}\bigg(\max_i \mathrm{W}(\hat{\mu}_m|_{U_i}, \mu|_{U_i}) \leq t\bigg) \geq 1 - \delta_m$$

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

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

*where*

$$c = \left(\frac{u_+}{1-u_+}\right)^N, \quad N = \left\lceil \left(\frac{204r}{t}\right)^D \right\rceil, \quad \gamma = 1 - u_-(1-\xi), \quad \xi = \exp\left(\frac{-t^2}{8r^2}\right)$$
$$u_- = \inf_{x\in\operatorname{supp}\mu} \mu(\mathcal{B}(x,r)), \quad u_+ = \sup_{x\in\operatorname{supp}\mu} \mu(\mathcal{B}(x,r))$$

*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 $U_i$:

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

Then we have:

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

We then apply the union bound:

$$\begin{aligned}
\mu^m(E_1 \cup \cdots \cup E_m) \leq & \mu^m(E_1) + \cdots + \mu^m(E_m) \\
= & m \cdot \int \mu^{m-1}\left( (x_2, \dots x_m) \middle| (x_1, x_2, \dots x_m) \in E_1 \right) \mathrm{d}\, x_1 \\
\leq & m \cdot \int \left( \frac{u_x}{1-u_x} \right)^N (m-1)^N \gamma_x^{m-1} \, \mathrm{d}\, x
\end{aligned}$$

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

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

and the claim is shown. □

# Chapter 5

# Tangent space and dimension estimation

## 5.1 Introduction

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

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

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

$$\hat{d}_\eta(\mu) = \min\left\{k \;\middle|\; \frac{\lambda_{k+1} + \cdots + \lambda_D}{\lambda_1 + \cdots + \lambda_D} \leq \eta\right\}$$
$$\mathcal{L}_k(\mu) = \operatorname{span}\Big(\mathcal{E}(\mu, \lambda_1), \ldots, \mathcal{E}(\mu, \lambda_k)\Big)$$

where $(\lambda_1, \ldots \lambda_D)$ are eigenvalues of $\Sigma[\mu]$, $\mathcal{E}(\mu, \lambda)$ is the $\lambda$-eigenspace of $\Sigma[\mu]$, and $\Sigma[\mu]$ is the covariance matrix of $\mu$.

Given $m$ points $\mathbf{x} = \{x_1, \ldots x_m\} \subset \mathbb{R}^D$, local PCA at an open set $W \subseteq \mathbb{R}^D$ performs PCA on points of $\mathbf{x}$ that lie in $W$. We are interested in $W$ given by an open ball. Given a radius parameter $r > 0$, let $\mathbf{x}_i := \{x_j \mid j \neq i\} \cap \{y \mid \|y - x_i\| < r\}$ and let $\hat{\mu}_i := \frac{1}{\#\mathbf{x}_i} \sum_{y \in \mathbf{x}_i} \delta_y$ be the local empirical measure at $x_i$. Define the *k-dimensional tangent space estimator* and the *intrinsic dimension estimator* with threshold $\eta$:

$$\hat{\Pi}(\mathbf{x}, r, i, k) := \mathcal{L}_k(\hat{\mu}_i)$$
$$\hat{d}(\mathbf{x}, r, i, \eta) := \hat{d}_\eta(\hat{\mu}_i) \tag{5.1.1}$$

When we calculate $\hat{\Pi}$ and $\hat{d}$ for a sample drawn near a $d$-dimensional manifold[1], we will get accurate estimations of tangent spaces and the intrinsic dimension $d$. Intuitively, this is because when a manifold is zoomed in closely enough at each point, its curvature flattens out and we essentially get a $d$-dimensional disk. Let's translate this intuition to precise mathematics. To do this, we precisely describe how we draw a random sample near a manifold.

[1]Note that in explanations like this, dependence on hyperparameters is implicit; $\hat{\Pi} = \hat{\Pi}(\mathbf{x}, r, i, k)$, $\hat{d} = \hat{d}(\mathbf{x}, r, i, \eta)$

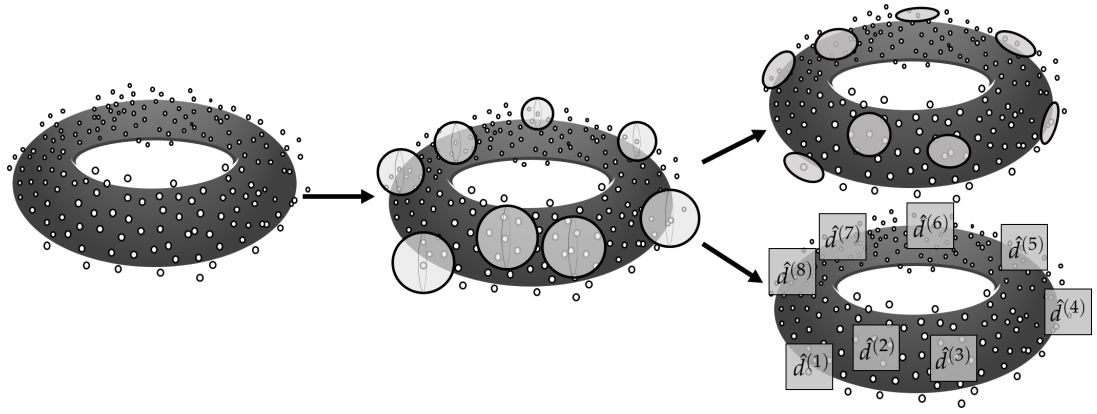


Figure 5.1: An illustration of Local PCA. Left: Dataset concentrated near a torus. Middle: Local neighborhood selection. Top Right: Tangent space estimation. Top bottom: Dimension estimation.

**Setup.** Let $M \subset \mathbb{R}^D$ be a smoothly embedded $d$-dimensional compact manifold. Let $\mu_0$ be a Borel probability measure on $\mathbb{R}^D$ with a probability density function $\varphi : M \to \mathbb{R}_{\geq 0}$: for each open $U \subseteq \mathbb{R}^D$, define

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

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

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

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

Additionally, denote by $\omega_d = \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ the volume of the unit $d$-dimensional ball. Denote by $\measuredangle(\Pi_1, \Pi_2)$ the principal angle between subspaces $\Pi_1, \Pi_2$ (Definition 5.4.1). Denote by $\mathbb{P}(E)$ the probability of event $E$. Denote by $\varphi_{\max}, \varphi_{\min}$ the maximum and the minimum of the function $\varphi$. Our main results ensure accurate estimations if:

1. $r$ is small enough to ignore curvature

2. $r$ is big enough to ignore noise

3. $mr^d$ is big enough to ensure dense sampling

Here, recall that $r$ is the radius parameter used to isolate a local neighborhood, as in Equation (5.1.1), and $d$ is the dimension of the manifold $M$.

**Main Results.**

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

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

Here $T_i$ is the tangent space of $M$ at $X_i^{\perp}$, the orthogonal projection of $X_i$ to $M$. $\widehat{T}_i = \hat{\Pi}(\mathbf{X}, r, i, d)$ is the tangent space estimator defined in (5.1.1). $S_1, S_2$ are:

$$\begin{aligned} S_1 &= \frac{\sin\theta}{(d+2)^{3/2}} \frac{\varphi_{\min}}{c_1 d \varphi_{\max} + c_2 \alpha \tau} \\ S_2 &= \frac{c_3 (d+2)^3}{\omega_d \varphi_{\min} \sin^2\theta} \log\left( \frac{c_4 D \varrho}{\delta} \right) \end{aligned}$$

where $(c_1, c_2, c_3, c_4) = (928, 192, 18574, 14)$.

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

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

where $\hat{d}_i = \hat{d}(\mathbf{X}, r, i, \eta)$ is the dimension estimator defined in (5.1.1). $S_1, S_2$ are:

$$\begin{aligned} S_1 &= \frac{1}{(d+2)D(1+\eta^{-1})} \frac{\varphi_{\min}}{c_1 d \varphi_{\max} + c_2 \alpha \tau} \\ S_2 &= \frac{c_3 (d+2)^2 D^2 (1+\eta^{-1})^2}{\omega_d \varphi_{\min}} \log\left( \frac{c_4 D \varrho}{\delta} \right) \end{aligned}$$

where $(c_1, c_2, c_3, c_4) = (1392, 288, 41791, 14)$.

**Remarks.**

- We explain the presence of $r^d$ and $\omega_d$ in the condition for sample size $m$. For simplicity, suppose a uniform distribution, so that $\varphi_{\min} = 1/V_M$, with $V_M = \mathcal{H}^d(M)$ being the intrinsic volume of the manifold $M$. Then the condition $r^d(n/\log n) \geq S_2$ of Theorem A can be rephrased as:

$$\left(\frac{m}{\log m}\right) \cdot \left(\frac{\omega_d r^d}{V_M}\right) \geq \frac{c_3(d+2)^3}{\sin^2\theta} \cdot \log\left(\frac{c_4 D\varrho}{\delta}\right)$$

  Here, note that $(\omega_d r^d)/V_M$ is the proportion of area taken up by a small disk of radius $r$ lying on the manifold $M$. Thus, the left hand side approximately measures the number of points lying on the small disk. In our derivation, the $\omega_d r^d$ term can be traced to $u_0$ in Theorem 5.5.3.

- Due to the $r^d$ term in the condition $r^d(m/\log m) \geq S_2$, we see that setting $r = (S\log m/m)^{1/d}$ for some $S \geq S_2$ ensures that this condition is always met. Indeed, this is the prescription of $r$ appearing in the Theorem 2 of [2]. The constant $S_2$ is fully calculated in our main theorems, improving Theorem 2 of [2]. As such, a condition of the form $r^\alpha(m/\log m) \geq S_2$ is likely the sharpest at $\alpha = d$. This also matches with the interpretation of "number of points lying in each local neighborhood" in the previous item.

- If $\varphi$ vanishes in a small region, we may avoid division by zero by replacing $\varphi_{\min}$ by $\Phi(r_-)$. Here $\Phi$ quantifies local concentration of the measure $\mu_0$. It is defined as $\Phi(r) = \inf_{x\in M} \mu_0\big(U_{x,r}\big)/(\omega_d r^d)$ and $U_{x,r} = \{y \in M \mid \mathrm{d}(x, \Pi_x(y)) \leq r\}$, where $\Pi_x$ is the projection map to $T_xM$. Also $r_-$ is defined as $r_- = r(1 - r^2/4\tau^2) - 2s$. This stronger result is stated in Theorem 5.5.3.

- Conditions for $r$ given by two inequalities can be collectively replaced by one upper bound on a function $Q$, defined in Proposition 5.3.4.

**Structure of the chapter.** Theorems A and B follow easily from Theorem 5.5.3 in Section 5, which is about estimating covariance matrices locally. Theorem 5.5.3 is proven by combining:

- Local concentration of covariance
- Lipschitz continuity of covariance vs. Wasserstein
- Flattening a manifold locally

The first ingredient, the local concentration inequality, is developed in an earlier chapter: Proposition 4.1.6. The second ingredient shows that given two compactly supported probability measures $\mu, \nu$ valued in $\mathbb{R}^D$, there is a Lipschitz relation of the form $\|\Sigma[\mu]-\Sigma[\nu]\| \leq C \cdot \mathrm{W}_1(\mu, \nu)$ where $\Sigma[\mu]$ is the covariance matrix of $\mu$ (Proposition 5.2.3). Flattening a manifold locally refers to precisely quantifying the Wasserstein distance between a small local section of a noisy non-uniform measure on manifold versus the uniform distribution over the tangential disk (Proposition 5.3.4). The Lipschitz relation then translates the Wasserstein bound to the bound on matrix norms.

**Related works.** The task of estimating geometric and topological quantities of manifolds from finitely many sample points lies at the crux of statistical inference, and as such the literature surrounding these topics is vast. Below we have described some of the techniques of which we are aware, and direct the reader to [101, 58, 34] for a more comprehensive survey.

**Tangent space estimation.** Probabilistic bounds on tangent space estimation using Local PCA have been studied in considerable detail, for example in [2, 93, 53, 84]. To the best of our knowledge, our work is the first in which the tangent space estimation applies to:

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

The dimensional dependence, for example, reflects the fact that covariance of the uniform distribution over the $d$-dimensional unit disk have $O(1/d)$ terms (see Lemma 2.2.5).

In [53] and [93], the underlying probability measure is assumed to be uniform, and only estimation at a single point is considered. In [84], various constants have not been

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

**Dimension estimation.** The idea to use local principal component analysis for estimating intrinsic dimension is ancient, dating back at least to [43]. As such, there is a plethora of literature on the problem of estimating intrinsic dimensions. The work of [60] provides a practical and widely-used maximum likelihood estimator, but there are no known theoretical guarantees of its correctness even for synthetic data. The minimax-based estimator of [54] does come with such guarantees, but in order to compute it one is compelled to solve minimisation problems over the symmetric group on $m$ elements (with $m$ being the total size of the input dataset); thus, this estimator becomes intractable in practice. The recent work of [21] introduces a far more efficient Wasserstein-based estimator with guarantees[2], but does not adapt to noise. Our efforts in this chapter were motivated by the desire to find a suitable balance between practical efficiency, theoretical soundness and compatibility with noise.

**Concentration inequality.** Our concentration inequality for covariance matrices, Proposition 4.1.4, is directly derived from the matrix Hoeffding inequality in [92]. A more sophisticated approach, such as the one from [55], may be used to improve our concentration inequality. For instance, the constants appearing in Proposition 4.1.4 may be improved. Similar methods for analyzing (non-local, non-manifold) PCA are also studied in [56, 76].

**Other Techniques.** We also list related techniques that appear in other papers. A cubic bound of the form $\|\Sigma[\mu] - \Sigma[\nu]\| \leq Cr^3$, where $\mu, \nu$ are probability measures supported on a ball of radius $r$ in $\mathbb{R}^D$, is derived for uniform measures in [12]. We also obtain a similar inequality (Proposition 5.2.3 and Corollary 5.3.5). The key difference in the two derivations is that our approach uses the Wasserstein distance rather than the total variation distance from [12] to quantify similarity of measures. Our inequality has

[2]We note in passing that the number of points we require to ensure a $1 - \delta$ probability of correct dimension estimation in our result is $m \sim \log(1/\delta)$, which improves on the rate $m \sim \log(1/\delta)^3$ of [21].

the advantage of allowing non-uniformity and of having explicit constants.

We use a transportation plan in Proposition 5.3.4 to quantify how much a measure supported near a manifold locally deviates from the uniform measure on a tangential disk. This transportation plan is executed with a similar idea as the proof of Proposition 3.1 in [91]. However, their transportation plan does not involve noise and applies to different types of local covariance matrices.

In [10], local polynomial regression were used to estimate manifolds and their tangent spaces from uniform point samples lying on tubular neighbourhoods. Compared to this work, our results have the advantage of not requiring the noise to be uniformly distributed. Our result only estimates tangent spaces and not higher-order information like curvature. However, the Wasserstein bound could potentially be leveraged to produce bounds on polynomial approximations.

There is an extensive body of research dedicated to understanding the effects of additive noise on the inference of principal components and singular values. In particular, we point out that there are principled methods to choose the threshold for the singular values, such as [38]. Meanwhile, PCA projection is known to be sensitive to distortions caused by noise, and algorithms like Randomised SVD have been proposed to reduce this effect [77].

Local PCA has been extensively used in contexts independent of the manifold hypothesis [43, 52, 94, 69], although the theoretical analysis is either heuristic or makes strong assumptions on the underlying distribution (e.g. Gaussian). Theoretical analysis in manifold learning is a flourishing field, with many significant examples including [45, 44, 1, 2, 42, 41, 54, 10, 91] and many others.

## 5.2 Lipschitz property of covariance matrix

Our goal in this section is to outline sufficient conditions under which the assignment $\mu \mapsto \Sigma[\mu]$ becomes a Lipschitz function with respect to the Wasserstein distance [97] on its domain. Throughout this section, we use the notation $X \sim \mu$ and $Y \sim \nu$, whenever probability distributions $\mu, \nu$ are defined.

**Lemma 5.2.1.** *Given Borel probability measures $\mu, \nu$ valued in $\mathbb{R}^D$, define $\tilde{\mu} = \mathrm{Law}(X - \mathbb{E}X)$ and similarly $\tilde{\nu}$. Then for each $p \geq 1$,*

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

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

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

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

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

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

**Lemma 5.2.2.** *For probability measures $\mu, \nu$ defined on $\mathbb{R}$ and supports contained the interval $[-R, +R]$, we have the $2R$-Lipschitz relation for all $p \geq 1$:*

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

*Proof.* Since $\mathrm{W}_p$ is increasing in $p$, it suffices to prove the assertion for $p = 1$.

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

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

**Proposition 5.2.3.** *Suppose $\mu, \nu$ are probability measures on $\mathbb{R}^D$ such that each measure comes with a ball of radius $r$ that contains the support of the measure. Then for $p \geq 1$, we have the following Lipschitz property:*

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

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

*Proof.* We assume that $r = 1$, since the case for general $r$ follows by scaling: $r$ affects the covariance matrix on the order of $r^2$ and the Wasserstein distance on the order of $r$. Also, the second inequality follows from the first by Lemma 5.2.1, so it suffices to show the first inequality. Since we are then working with $\tilde{\mu}$ and $\tilde{\nu}$ and since covariance matrix is invariant under translation, we may rewrite $\mu = \tilde{\mu}$ and $\nu = \tilde{\nu}$ and assume that $\mu, \nu$ have zero means. We may also assume that both $\operatorname{supp}\mu$ and $\operatorname{supp}\nu$ are contained within $\mathcal{B}_2(0)$ by the triangle inequality; there is a ball $\mathcal{B}_1(x)$ of radius 1 containing $\operatorname{supp}\mu$, so that by triangle inequality, $\operatorname{supp}\mu \subseteq \mathcal{B}_1(x) \subseteq \mathcal{B}_2(0)$.

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

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

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

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

where $U^\top\mu = \mathrm{Law}(U^\top X)$ and $(U^\top\mu)_i$ denotes the marginal of $U^\top\mu$ at its $i$th coordinate. The first inequality is Lemma 5.2.2 with $2R = 4$. The second inequality is a general fact that applies to the Wasserstein distances between marginals. The last equality follows from the fact that the Wasserstein distance is invariant with respect to isometry applied simultaneously to the two measures. Finally, multiplying by the Lipschitz constant 2 for the non-centered measures, we get the Lipschitz constant 8. The inequality for other $p$ follows since $\mathrm{W}_p$ is increasing in $p$. □

## 5.3 Flattening a measure on manifold

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

**Definition 5.3.1.** Given a metric space and a positive integer $d$, denote by $\mathcal{H}^d$ the $d$-dimensional Hausdorff measure [83] on the metric space:

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

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

**Definition 5.3.2.** Suppose $M$ is a $d$-dimensional smooth compact manifold with a smooth embedding into $\mathbb{R}^D$ and $\varphi : M \to \mathbb{R}^+$ is a continuous function satisfying $\int_M \varphi\,\mathrm{d}\,\mathcal{H}^d = 1$. Let $\mu_0$ be the Borel probability measure given by defining for each open $U \subseteq \mathbb{R}^D$ the following:

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

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

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

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

The following are notions from differential geometry relevant to us.

**Definition 5.3.3.** For each compact Riemannian manifold $M \subset \mathbb{R}^D$,

1. For each $x, y \in M$, let $\mathrm{d}_M(x, y)$ be the length of the shortest geodesic connecting $x$ and $y$.[3]

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

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

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

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

**Proposition 5.3.4.** *Let $(\mu_0, \mu) \in \mathcal{P}(M, s)$ where $M \subseteq \mathbb{R}^D$ is a compact smoothly embedded $d$-dimensional manifold with reach $\tau$ and $s \geq 0$. Let $x \in \operatorname{supp} \mu$, let $x_\perp$ be any point in $\mathcal{B}_s(x) \cap M$, and let $r$ be a number satisfying the conditions $2s \leq r \leq (\sqrt{2} - 1)\tau - 2s$ and $r \leq \tau/(2\sqrt{2}d)$. Then the following holds for any $p \geq 1$:*

$$\mathrm{W}_p(\nu, \tilde{\nu}) \leq \tau \cdot Q\left(\frac{r}{\tau}, \frac{s}{\tau}\right)$$

$$\text{where } \nu := \mu|_{\mathcal{B}_r(x)}, \text{ and } \tilde{\nu} := \mathcal{H}^d|_{\mathcal{B}_r(x_\perp) \cap T_{x_\perp} M}$$

[3]Equivalently, $\mathrm{d}_M(x, y)$ be the infimum of lengths of all piecewise regular curves that connect $x$ and $y$. This follows from the Hopf-Rinow Theorem; see Corollary 6.21 and 6.22 in [59].

*where $Q$ is given by:*

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

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

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

*$\mathcal{B}^\circ_-$ is the tangential disk of radius $r_-$ centred at $x_\perp$, and $\Pi$ is the projection map to the tangent space $T_{x_\perp}M$, restricted to $\mathcal{B}_r(x)\cap M$.*

*Proof.* We use the following multi-step transportation plan (see Figure 5.3), from $\nu_0 := \nu$, going through $\nu_1, \nu_2, \nu_3, \nu_4$ which we define below and finally reaching $\nu_5 := \tilde{\nu}$. Informally, these steps can be summarized as

1. Perform a naive denoising on $\nu_0$ to get $\nu_1$
2. Apply projection to get $\nu_2$
3. Fold in the portion of $\nu_2$ on the outer rim to the inside to get $\nu_3$
4. Flatten out the nonuniformity and get $\nu_4$.
5. Rescale radius uniformly to get $\nu_5$.

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

$$\begin{aligned} \nu_1(U) =& \mathbb{P}(X\in U \mid X+Y\in\mathcal{B}_r(x)) \\ =& \frac{1}{u}\mathbb{P}(X\in U \text{ and } X+Y\in\mathcal{B}_r(x)) \\ \text{where } u =& \mu(\mathcal{B}_r(x)) \end{aligned} \tag{5.3.1}$$

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

$$\mathrm{W}_p(\nu_0,\nu_1) \le \mathbb{E}_{(X+Y,X)\sim\nu_{01}}\|(X+Y)-X\| \le s$$

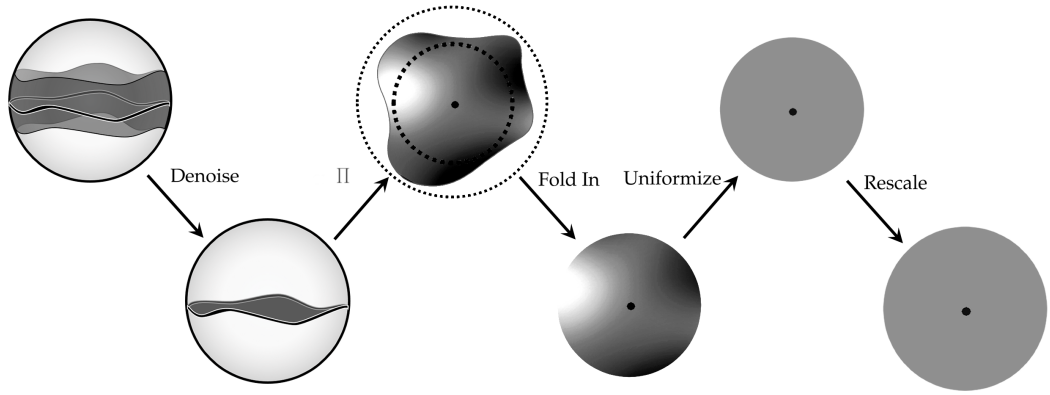


Figure 5.2: An overview of the transportation plan in the proof of Proposition 5.3.4. The last four sub-diagrams take place on the tangent space. Nonuniform shadings in the 3rd, 4th sub-diagrams indicate nonuniform probability distribution.

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

By Equation (5.3.1), $\nu_1$ is well understood in regions where the condition $X + Y \in \mathcal{B}_r(x)$ either always or never holds. If $X \in \mathcal{B}_{r-s}(x)$, then since $\|Y\| \leq s$, the triangle inequality implies $X + Y \in \mathcal{B}_r(x)$. Similarly if $X \notin \mathcal{B}_{r+s}(x)$, then $X + Y \notin \mathcal{B}_r(x)$. By also noting that $\|x - x_\perp\| \leq s$, the triangle inequality once again implies $\mathcal{B}_{r-2s}(x_\perp) \subseteq \mathcal{B}_{r-s}(x)$ and $\mathcal{B}_{r+s}(x) \subseteq \mathcal{B}_{r+2s}(x_\perp)$. Applying Equation (5.3.1), we get the following:

$$\begin{aligned} \nu_1(U) &\leq \frac{\mu_0(U)}{u} && \text{for any } U \\ \nu_1(U) &= \frac{\mu_0(U)}{u} && \text{for } U \subseteq \mathcal{B}_{r-2s}(x_\perp) \\ \nu_1(U) &= 0 && \text{for } U \subseteq \mathcal{B}_{r+2s}(x_\perp)^{\mathrm{c}} \end{aligned} \tag{5.3.2}$$

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

**Step 2.** We define $\nu_2$ by pushing forward $\nu_1$ along the projection map to the tangent space, and we must do it where the map is invertible. In this proof, we define $\Pi$ as the projection map to the tangent space $T_{x_\perp} M$, restricted to $\mathcal{B}_r(x) \cap M$. By Lemma 2.1.7, we know that the projection map is a diffeomorphism. Furthermore,

$$\mathcal{B}_-^\circ \subseteq \Pi(\mathcal{B}_- \cap M), \quad \Pi(\mathcal{B}_+ \cap M) \subseteq \mathcal{B}_+^\circ \tag{5.3.3}$$

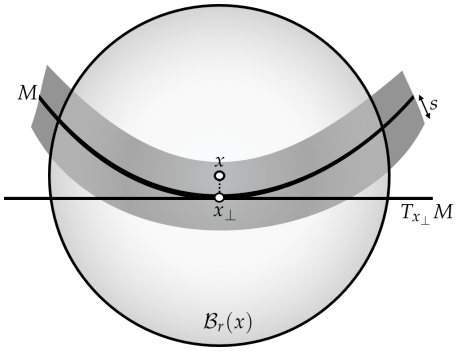


Figure 5.3: Measure $\mu$ and its restriction $\mu|_{\mathcal{B}_r(x)}$, where $x \in \mathbb{R}^D$ and $x_\perp \in M$.

where, denoting $\mathring{\mathcal{B}}_r$ by the open ball of radius $r$ in $T_{x_\perp}M$ centered at 0, [4]

$$\mathcal{B}_- = \mathcal{B}_{r-2s}(x_\perp), \quad \mathcal{B}_+ = \mathcal{B}_{r+2s}(x_\perp)$$
$$\mathcal{B}_-^\circ = \mathring{\mathcal{B}}_{r_-}, \quad \mathcal{B}_+^\circ = \mathring{\mathcal{B}}_{r_+}$$
$$r_- = r\left(1 - \frac{r^2}{4\tau^2}\right) - 2s, \quad r_+ = r + 2s$$

The transportation plan is the application of Lemma 3.1.1 to the pushforward along $\Pi$. In performing the transportation, we regard the tangent space as embedded: $T_{x_\perp}M \subseteq \mathbb{R}^D$ so that the transportation happens in the ambient space $\mathbb{R}^D$. By the last result mentioned in Lemma 2.1.4 (Proposition 6.3, [72]), the transportation cost then is bounded as:

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

Thus by Equations (5.3.2) and (5.3.3),

$$\begin{aligned}
&\nu_2(U) \le \frac{\mu_0(\Pi^{-1}U)}{u} && \text{for } U \subseteq \mathring{\mathcal{B}}_+ \\
&\nu_2(U) = \frac{\mu_0(\Pi^{-1}U)}{u} && \text{for } U \subseteq \mathring{\mathcal{B}}_- \\
&\nu_2(U) = 0 && \text{for } U \subseteq (\mathring{\mathcal{B}}_+)^{\mathrm{c}}
\end{aligned} \tag{5.3.4}$$

Meanwhile, we can evaluate $\mu_0(U)$ when $U \subseteq \mathring{\mathcal{B}}_+$ explicitly using the area formula from geometric measure theory[5], which is a generalization of chain rule:

$$\mu_0(\Pi^{-1}U) = \int_{\Pi^{-1}U} \varphi \,\mathrm{d}\,\mathcal{H}^d = \int_U \varphi(\Pi^{-1}y)\,\mathrm{J}\,\Pi^{-1}(y)\,\mathrm{d}\,y \tag{5.3.5}$$

[4]Note that $(r-2s)(1-(r-2s)^2/4\tau^2) \le (r-2s)(1-r^2/4\tau^2) = r(1-r^2/4\tau^2) - 2s(1-r^2/4\tau^2) \le r(1-r^2/4\tau^2) - 2s$

[5]See for example [40] for a standard reference in geometric measure theory

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

$$\begin{aligned}
\nu_2(U) &\leq \frac{1}{u}\int_U \varphi(\Pi^{-1}y)\,\mathrm{J}\,\Pi^{-1}(y)\,\mathrm{d}\, y && \text{for } U \subseteq \mathring{\mathcal{B}}_+ \\
\nu_2(U) &= \frac{1}{u}\int_U \varphi(\Pi^{-1}y)\,\mathrm{J}\,\Pi^{-1}(y)\,\mathrm{d}\, y && \text{for } U \subseteq \mathring{\mathcal{B}}_- \\
\nu_2(U) &= 0 && \text{for } U \subseteq (\mathring{\mathcal{B}}_+)^{\mathrm{c}}
\end{aligned} \tag{5.3.6}$$

**Step 3.** We saw that $\nu_2$ can be written in terms of $\mu_0$ inside radius $r_-$ and vanishes outside radius $r_+$. The annular region between the two radii is harder to understand since it is where curvature and noise interact, as indicated by Equation (5.3.1). In Step 3 we remove this annular region, so that we only need to deal with $\nu_2$ restricted to $\mathring{\mathcal{B}}_-$. We decompose $\nu_2$ as $\nu_2 = \nu_2^- + \nu_2^+$, where we define for each Borel set $U \subseteq T_{x_\perp}M$ the following:

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

Define

$$\begin{aligned}
\nu_3 &:= \mathfrak{m}^{-1}\nu_2^- \\
\mathfrak{m} &:= \nu_2^-(T_{x_\perp}M)
\end{aligned}$$

The transportation plan is to: (a) transport $\nu_2^+$ to the Dirac delta distribution centered at $0 \in T_xM$ and (b) transport this Dirac delta distribution back to $\frac{1-\mathfrak{m}}{\mathfrak{m}}\nu_2^-$. By Lemma 3.1.2, we have the bound:

$$\mathrm{W}_p(\nu_2, \nu_3) \leq (r_+ + r_-)(1-\mathfrak{m}) \leq 2r(1-\mathfrak{m})$$

since the first part of this transportation moves by distance at most $r_+$, the second part moves by at most $r_-$, and the total mass to move is $(1-\mathfrak{m})$. Equation (5.3.6) carries over since $\nu_3$ and $\nu_2^-$ are proportional; for each open $U \subseteq T_{x_\perp}M$,

$$\nu_3(U) = \frac{1}{u\mathfrak{m}}\int_{U\cap\mathring{\mathcal{B}}_-} \varphi(\Pi^{-1}y)\,\mathrm{J}\,\Pi^{-1}(y)\,\mathrm{d}\, y \tag{5.3.7}$$

**Step 4.** We flatten out the non-uniformity in $\nu_3$. As in Equation (5.3.7) above, $\nu_3$ is given by the probability density function $\psi(y) := \varphi(\Pi^{-1}y)\,\mathrm{J}\,\Pi^{-1}(y)$ times a constant. Defining $\nu_4 = \mathcal{H}^d|_{\mathring{\mathcal{B}}_-}$, we can directly apply Lemma 3.1.3:

$$\mathrm{W}_p(\nu_3, \nu_4) \leq \frac{\omega_d r_-^d}{u\mathfrak{m}} \cdot (\psi_{\max} - \psi_{\min}) \cdot 2r_-$$

where the factor $\omega_d r_-^d$ is needed to rescale the Lebesgue measure $\mathrm{d}\,y$ in Equation (5.3.7) into $\widetilde{\mathrm{d}\,y} = \mathrm{d}\,y/(\omega_d r_-^d)$ so that $\int_{\mathring{\mathcal{B}}_-} \widetilde{\mathrm{d}\,y} = 1$, so that Lemma 3.1.3 can be applied. In the above, extrema of $\psi$ are taken over $\mathring{\mathcal{B}}_-$. Since $\psi$ is the product of $\varphi$ and the Jacobian, the variation $\psi_{\max} - \psi_{\min}$ can be controlled with the triangle inequality as follows:

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

Here the extrema of $\varphi$ are taken over the geodesic ball $\Pi^{-1}\mathring{\mathcal{B}}_-$. By Proposition 2.1.7, we see that:

$$\begin{gathered} J_- \leq \mathrm{J}\,\Pi^- \leq J_+ \\ \text{where } J_- = 1,\ J_+ = \left(1 - \frac{\sqrt{2}r}{\tau}\right)^{-d} \end{gathered} \tag{5.3.8}$$

We furthermore note that, by Equation 5.3.6,

$$\begin{aligned} u\mathfrak{m} &= \int_{\mathring{\mathcal{B}}_-} \varphi(\Pi^{-1}y)\,\mathrm{J}\,\Pi^{-1}(y)\,\mathrm{d}\,y \geq \omega_d r_-^d J_- \varphi_{\min} \\ \implies \varphi_{\min} &\leq \frac{u\mathfrak{m}}{\omega_d r_-^d} \cdot \frac{1}{J_-} \end{aligned}$$

Thus the transportation cost is bounded as:[6]

$$\mathrm{W}_p(\nu_3, \nu_4) \leq \left(\frac{\omega_d r_-^d}{u\mathfrak{m}}\,(\varphi_{\max} - \varphi_{\min})\, J_+ + \frac{J_+ - J_-}{J_-}\right) \cdot 2r_-$$

**Step 5.** Here we simply rescale $\mathring{\mathcal{B}}_-$ from radius $r_-$ to $r$ radially, which multiplies the associated probability density function by a constant factor (Lemma 2.1.10), so that we get another uniform distribution. By Lemma 3.1.1, the transportation cost is bounded by

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

[6]We note at this point that the extrema of $\varphi$ may be taken over $\mathcal{B}_{r+2s}(x_\perp)$ instead, since $\mathcal{B}_{r+2s}(x_\perp) \supseteq \Pi^{-1}(\mathring{\mathcal{B}}_-)$. This relaxation is done for a compatibility with another extrema of $\varphi$ taken later.

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

$$\begin{aligned}
&\mathrm{W}_p(\nu_0,\nu_5)\\
\leq&\,\mathrm{W}_p(\nu_0,\nu_1)+\mathrm{W}_p(\nu_1,\nu_2)+\mathrm{W}_p(\nu_2,\nu_3)+\mathrm{W}_p(\nu_3,\nu_4)+\mathrm{W}_p(\nu_4,\nu_5)\\
\leq&\,s+\frac{(r+2s)^2}{\tau}+2r(1-\mathfrak{m})+\\
&+\left(\frac{\omega_d r_-^d}{u\mathfrak{m}}\left(\varphi_{\max}-\varphi_{\min}\right)J_+ +(J_+-1)\right)\cdot 2r+\left(\frac{r^3}{4\tau^2}+2s\right)
\end{aligned}\tag{5.3.9}$$

Using Equations (5.3.4), (5.3.6) and (5.3.8), we obtain the following bounds:

$$\begin{aligned}
u\mathfrak{m}&=\mu_0(\Pi^{-1}\mathring{\mathcal{B}}_-)\leq\varphi_{\max}J_+\omega_d r_-^d\\
u(1-\mathfrak{m})&\leq\mu_0(\Pi^{-1}(\mathring{\mathcal{B}}_+-\mathring{\mathcal{B}}_-))\leq\varphi_{\max}J_+\omega_d(r_+^d-r_-^d)
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1-\mathfrak{m}}{\mathfrak{m}}=\frac{u(1-\mathfrak{m})}{u\mathfrak{m}}\leq\frac{\varphi_{\max}J_+\omega_d(r_+^d-r_-^d)}{u\mathfrak{m}}=\Phi'(\Omega^{-d}-1)\\
&\text{with }\Omega=\frac{r_-}{r_+},\Phi'=\frac{\varphi_{\max}J_+\omega_d r_-^d}{u\mathfrak{m}}\geq 1
\end{aligned}$$

We can bound $\int\nu_2^{\mathrm{out}}$ using the above, as follows:

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

where the first inequality holds by plugging in the upper bound for $(1-\mathfrak{m})/\mathfrak{m}$, and the second inequality holds since $\Phi'\geq 1$. Plugging these into Equation (5.3.9), we get that

$$\begin{aligned}
\mathrm{W}_p(\nu_0,\nu_5)\leq&\,s+\frac{(r+2s)^2}{\tau}+2r(1-\Omega^d)\varphi_{\max}J_+\frac{\omega_d r_-^d}{u\mathfrak{m}}\\
&+\left(\frac{\omega_d r_-^d}{u\mathfrak{m}}\left(\varphi_{\max}-\varphi_{\min}\right)J_+ +(J_+-1)\right)\cdot 2r+\left(\frac{r^3}{4\tau^2}+2s\right)
\end{aligned}$$

We bound $J_+$ using Lemma 2.2.2. By applying the assumption $r\leq\tau/(2\sqrt{2}d)$, we see that the lemma applies with $c=1/2$:

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

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

[7] We plug in the definition $J_-=1$, and we also use a slight abuse of notation and identify $\nu_k$ with $\iota_*\nu_k$ for $k=2,\dots 5$, where $\iota:T_{x_\perp}M\hookrightarrow\mathbb{R}^D$ is the inclusion of tangent space. This is not a problem, since generally $\mathrm{W}_p(\iota_*\mu_1,\iota_*\mu_2)\leq\mathrm{W}_p(\mu_1,\mu_2)$ holds for any measures $\mu_1,\mu_2$ on $T_{x_\perp}M$.

[8] See Equation 5.3.3.

**Corollary 5.3.5.** *In Proposition 5.3.4, suppose that we additionally assume that there exists $\alpha$ such that the following Lipschitz continuity holds for every $x, y \in M$:*

$$\|\varphi(x) - \varphi(y)\| \leq \alpha \cdot \mathrm{d}_M(x, y)$$

*Suppose we also assume $s \leq r^2/(2\tau)$. Then we have the following quadratic bound:*

$$\mathrm{W}(\nu, \tilde{\nu}) \leq Q_2 \cdot \tau\rho^2$$

*where $Q_2$ is defined as:*

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

*Proof.* We use the notation $\rho = r/\tau, \sigma = s/\tau$. Firstly by the assumption $\sigma \leq \rho^2/2$,

$$\Omega = \frac{\rho - \rho^3/4 - 2\sigma}{\rho + 2\sigma} \geq \frac{1 - \rho^2/4 - \rho}{1 + \rho} \geq \frac{1 - c\rho}{1 + c\rho}, \text{ where } c = \frac{9}{8}$$

Then, assuming $\rho \in [0, 8/9]$, Lemma 2.2.3 says:

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

By the Lipschitz condition and the noise bound,

$$\begin{aligned}
Q(\rho, \sigma) =& 3\sigma + (\rho + 2\sigma)^2 + 2\rho(1 - \Omega^d)\frac{1}{\Phi}\varphi_{\max}\left(1 + 4\sqrt{2}d\rho\right) \\
&+ \left(\frac{1}{\Phi}\left(\varphi_{\max} - \varphi_{\min}\right)\left(1 + 4\sqrt{2}d\rho\right) + 4\sqrt{2}d\rho\right) \cdot 2\rho + \frac{1}{4}\rho^3 \\
\leq& \frac{3}{2}\rho^2 + (1 + \rho)^2\rho^2 + \frac{1}{\Phi}\varphi_{\max}4dc\left(1 + 4\sqrt{2}d\rho\right)\rho^2 \\
&+ \left(\frac{1}{\Phi}(2(r + 2s)\alpha)(1 + 4\sqrt{2}d\rho) + 4\sqrt{2}d\right) \cdot 2\rho^2 + \frac{1}{4}\rho^3
\end{aligned}$$

where the Lipschitz relation is applied to bound $\varphi_{\max} - \varphi_{\min} \leq 2(r + 2s)\alpha$ by using *two* radial geodesics of length $\leq r_+ = r + 2s$ in the unit ball of radius $r_+$ in the tangent space $T_{x_\perp}M$. Factoring out $\rho^2$ and plugging back in the definition $c = \frac{9}{8}$, we get:

$$\frac{1}{\rho^2}Q(\rho, \sigma) \leq \left(8\sqrt{2}d + \frac{5}{2} + \frac{9}{4}\rho + \rho^2\right) + \frac{\varphi_{\max}}{\Phi}\frac{9}{2}d(1 + 4\sqrt{2}d\rho) + \frac{4(\rho + \rho^2)\alpha\tau}{\Phi}(1 + 4\sqrt{2}d\rho)$$

Using the assumption $\rho \leq \frac{1}{2\sqrt{2}d}$, we get the bounds $1 + 4\sqrt{2}d\rho \leq 3$, and $9\rho/4 + \rho^2 \leq 1$, and $\rho + \rho^2 \leq \frac{1}{2}$. We obtain the claimed bound by plugging them in. □

## 5.4 Principal angles

To work with a general notion of angles, we define and study principal angles in this section. Indeed, every pair of linear subspaces of the same dimension can be characterised by *principal angles*, up to (simultaneous) rigid motion.

**Definition 5.4.1.** Given $\pi_1, \pi_2 \in Gr(d, D)$, let $A_i \in \mathbb{R}^{D\times d}$ be a matrix with orthonormal columns that span $\pi_i$. 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 $(\pi_1, \pi_2)$ are defined as the angles $(\theta_1, \dots \theta_d) \in [0, \pi/2]^d$ such that $(\cos\theta_1, \dots \cos\theta_d) = \measuredangle(\pi_1, \pi_2)$, which satisfy $\theta_1 \le \dots \le \theta_d$.

We note in particular that $\theta_1 = \dots = \theta_{d_0} = 0 < \theta_{d_0+1}$, where $d_0 = \dim(\pi_1) = \dim(\pi_2)$.

The largest principal angle has a simple interpretation:

**Lemma 5.4.2.** *If $\measuredangle(\pi_1, \pi_2) = (\cos\theta_1, \dots \cos\theta_d)$ 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 $\pi_i$. 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 $\cos\theta_D = \min_{\|y\|=1, y\in\pi_2}\max_{\|x\|=1, x\in\pi_1}\langle x, y\rangle$. □

Principal angles characterise pairs of subspaces up to rotation.

**Proposition 5.4.3.** *$\measuredangle$ 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 $S(k, j) = \{(t_1, \dots t_k) \mid 1 \ge t_1 \ge \dots \ge t_k \ge 0,\ t_1 = \dots = t_j = 1\}$, 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* $\underline{\measuredangle}(\pi_1, \pi_2) = \underline{\measuredangle}(\pi_1', \pi_2')$*, then there exists an element* $A \in O(D)$ *such that* $(A\pi_1, A\pi_2) = (\pi_1', \pi_2')$*. Furthermore, if* $(t_1, \dots t_d) \in [0,1]^d$ *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* $\underline{\measuredangle}(\pi_1, \pi_2) = (t_1, \dots t_d)$*.*

*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 $\underline{\measuredangle}(\pi_1, \pi_2) = \underline{\measuredangle}(\pi_1', \pi_2')$. Let $A_i \in \mathbb{R}^{D \times d}$ be a matrix with orthonormal columns spanning $\pi_i$, 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 $\underline{\measuredangle}(\pi_1, \pi_2) = \underline{\measuredangle}(\pi_1', \pi_2')$, 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:

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

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 $Z = [[U, 0], [0, W]]$, 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. □

**Corollary 5.4.4.** *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* $\pi_i$*. 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 $\Theta = \mathrm{diag}(\theta_{d_0+1}, \dots \theta_d) \in \mathbb{R}^{(d-d_0)\times(d-d_0)}$ is the diagonal matrix of nonzero principal angles $\measuredangle(\pi_1, \pi_2) = (\theta_1, \dots \theta_d)$.*

*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. □

To control angles between tangent spaces, we will use the following variant of the Davis-Kahan theorem [36, 102] (recall the eigenvalue notations given in the Introduction):

**Theorem 5.4.5** (Davis-Kahan-Wang-Samworth)**.** *Let $A, B \in \mathbb{R}^{D\times D}$ be real symmetric matrices. Let $1 \le d_1 \le d_2 \le D$ and assume that $\min(\lambda^{\mathrm{gap}}_{d_1-1}A, \lambda^{\mathrm{gap}}_{d_2}A) > 0$. Let $\pi_A$ be the span of the eigenspaces corresponding to $\lambda_{d_1}A, \lambda_{d_1+1}A, \dots \lambda_{d_2}A$, and let $\theta_1 \le \dots \le \theta_d$ be the principal angles between $(\pi_A, \pi_B)$. Then we have:*

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

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

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

## 5.5 Tangent space and dimension estimation

In this section, we combine the Propositions 4.1.6, 5.2.3, and 5.3.4 to prove Theorem 5.5.3. This in turn implies both Theorem A and B.[9]

**Definition 5.5.1.** Given a $d$-dimensional subspace $\Pi \subseteq \mathbb{R}^D$, denote the $D\times D$ orthogonal projection matrix to $\Pi$ by $\mathrm{P}_\Pi$, which is a real symmetric matrix, given concretely as:

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

[9]Technical note: In Theorems A and B, use Lemma 2.2.4, and use $\log(14D) > 1+\log(4D+2)$ assuming $D \ge 2$.

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

**Definition 5.5.2.** Let $\mathbf{X} = (X_1, \dots X_m)$ be an i.i.d. sample drawn from $\mu$, a Borel probability measure on $\mathbb{R}^D$. Given $x \in \mathbb{R}^D$ and $r > 0$, define:

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

If $\Pi \subseteq \mathbb{R}^D$ is a $d$-dimensional subspace, then Lemma 2.2.5 says that:

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

Thus an approximation to this covariance matrix in Proposition 5.3.4 amounts to the approximation of a projection matrix, and justifies the definition of $\hat{\mathrm{P}}_i$.

**Theorem 5.5.3.** *Let $(\mu, \mu_0) \in \mathcal{P}(M, s)$[10] where $M$ is a smoothly embedded compact $d$-dimensional manifold $M \subseteq \mathbb{R}^D$ with reach $\tau$ and $s \geq 0$ is a real number. Let $\varphi$ be the probability density function of $\mu_0$ which satisfies $\|\varphi(x) - \varphi(y)\| \leq \alpha \cdot \mathrm{d}_M(x, y)$. Let $X_1, \dots X_m$ be an i.i.d. sample drawn from $\mu$ and let $X_1^\perp, \dots X_m^\perp$ be their orthogonal projections to $M$. Given $\delta, \epsilon, \alpha > 0$ and assuming*[11] *$\epsilon < 2$, suppose $r, m$ satisfy the following:*

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

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

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

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

$$Q_2 = \left(\frac{7}{2} + 8\sqrt{2}d\right) + \frac{(27/2)d\varphi_{\max} + 6\alpha\tau}{\Phi}, \text{ where } \Phi = \frac{\mu_0(\Pi^{-1}\mathcal{B}_-^\circ)}{\omega_d r_-^d}$$

*Proof.* Out of total allowed error $\epsilon$, we will allocate one half $\epsilon/2$ to the concentration inequality (Proposition 4.1.6) and the other half $\epsilon/2$ to the curvature (Proposition 5.3.4). Throughout the proof, we use the shorthand $U_i = \mathcal{B}_r(X_i^\perp)$.

[10] See Definition 5.3.2.

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

**Concentration inequality:** By Proposition 4.1.6, we may use $k = \lfloor \varrho m \rfloor$ points for local covariance estimation by error level $r^2\epsilon/2(d+2)$:

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

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

**Curvature:** By combining Corollary 5.3.5 and Proposition 5.2.3, the following holds[12] for every $x \in \operatorname{supp}\mu$:

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

Note that $\frac{r^2}{d+2}\,\mathrm{P}_{X_i^\perp}$ is the covariance of the uniform measure over the tangential disk of radius $r$, by Lemma 2.2.5.

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

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

as desired. We note that the assumptions $2s \leq r$ and $r + 2s \leq (\sqrt{2}-1)\tau$ of Proposition 5.3.4 follow from the assumption on $r$ and $\epsilon < 2$. □

**Remark.** Note that the $r^d$ term in Theorems A and B appear from $u_0$; see the end of the upcoming subsection, where we use the inequality $u_0 \geq \omega_d r^d \varphi_{\min}$.

### 5.5.1 Proof of Theorem A

To use Theorem 5.5.3, we relate the projection matrices to angular deviation between subspaces using the Davis-Kahan theorem.

*Proof of Theorem A.* This is a direct corollary of plugging in $\epsilon = (\sin\theta)/(2\sqrt{d+2})$ in Theorem 5.5.3. Assuming that, the following holds for each $i \leq \lfloor \varrho m \rfloor$:

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

[12]Applying Corollay 5.3.5 requires assuming $\rho + \rho^2 \leq \sqrt{2}-1$ and $\rho \leq 1/(\sqrt{8}d)$. But this assumption is automatically satisfied by the $r$ in the assumption of the theorem, where we already assume $\rho \leq \epsilon/(16(d+2)Q_2) \leq 1/(192(d+2)^2)$. Thus these assumptions become redundant.

Since both $\mathrm{P}_i$ and $\hat{\mathrm{P}}_i$ are real symmetric matrices and since eigenvalues of $\mathrm{P}_i$ are $(1, \dots 1, 0, \dots 0)$, its $d$-th spectral gap is 1 and therefore letting $A = \mathrm{P}_i$, $B = \hat{\mathrm{P}}_i$ in the Davis-Kahan theorem gives the following[13]:

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

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

In Theorem A, the conditions for $(r, m)$ used in Theorem 5.5.3 are made stricter for the sake of easy interpretability. We explain how this is done.

**Condition on $r$.** The following is the required upper bound on $\rho = r/\tau$:

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

Using $\Phi \geq \varphi_{\min}$ (follows from Equation (5.3.5) and the Jacobian of inverse-projection being $\geq 1$), we get the following upper bound on $Q_2$:

$$\begin{aligned} Q_2 =& \left(\frac{7}{2} + 8\sqrt{2}d\right) + \frac{1}{\Phi}\left(\frac{27d}{2}\varphi_{\max} + 6\alpha\tau\right) \\ \leq& \left(\frac{7}{2} + 8\sqrt{2}d\right) + \frac{1}{\varphi_{\min}}\left(\frac{27d}{2}\varphi_{\max} + 6\alpha\tau\right) \\ \leq& \left(\frac{7}{2} + 8\sqrt{2} + \frac{27}{2}\right) d \cdot \frac{\varphi_{\max}}{\varphi_{\min}} + \frac{6\alpha\tau}{\varphi_{\min}} \\ \leq& \frac{29d\varphi_{\max} + 6\alpha\tau}{\varphi_{\min}} \end{aligned} \tag{5.5.1}$$

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

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

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

**Condition on $m$.** The required lower bound for $m/\log m$ is obtained by also plugging in $\epsilon = \sin\theta/(2\sqrt{d+2})$ in Theorem 5.5.3, and noting that $u_0 \geq \omega_d r_-^d \varphi_{\min}$, by Equations (5.3.5) and (5.3.3). Furthermore, we use the following:

$$r_- = r\left(1 - \frac{r^2}{4\tau^2}\right) - 2s \geq r \cdot \left(1 - \frac{r}{\tau} - \frac{r^2}{4\tau^2}\right)$$

[13] In the equation, note that we could choose $\epsilon = \epsilon/(2\sqrt{d})$ for a slightly tighter bound. Our choice of $\epsilon$ is for cleanliness of the final expression produced.

Assuming that $\rho = r/\tau$ satisfies $\rho + \rho^2/4 \leq c/d$ for some constant $c > 0$ and applying Lemma 2.2.2 with $t = \rho + \rho^2/4 \leq c/d$, we get:

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

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

### 5.5.2 Proof of Theorem B

To relate a perturbation of eigenvalues to a perturbation of covariance matrices, we use the Hoffman-Wielandt theorem [49]. Before stating it, we prove a simple lemma that gives a simple, clean version of the Hoffman-Wielandt theorem for real symmetric matrices.

**Lemma 5.5.4.** *For a metric space $M$ and its $n$-fold product space $M^n$, the following function is a metric on $M^n$:*

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

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

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

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

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

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

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

**Theorem 5.5.5** (Hoffman-Wielandt)**.** *For normal matrices $A, A'$ of dimension $D \times D$, there is an enumeration of eigenvalues $(\lambda_1, \ldots \lambda_D)$ of $A$ and $(\lambda'_1, \ldots \lambda'_D)$ of $A'$ such that*

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

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

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

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

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

**Proposition 5.5.6.** *Let $\vec{\lambda} = (\lambda_1, \ldots \lambda_D) \in \mathbb{R}^D$ be such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$. Let $\vec{\lambda}(d, D) = \frac{1}{d+2}(1, \ldots 1, 0 \ldots 0) \in \mathbb{R}^D$ where there are $D - d$ zeros. Let $\eta$ be a tolerance parameter such that $0 < \eta < 1/(2d)$.*

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

*where $\hat{d}_\eta$ is defined in the Introduction.*

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

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

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

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

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

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

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

*Proof of Theorem B.*

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

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

The proof is then completed by applying Proposition 5.5.6. We note how the expression $Q_2$ is weakened by using Equation (5.5.1), which is also used in deriving Theorem A:

$$\frac{\epsilon}{16(d+2)Q_2} \geq \frac{1}{48(d+2)D(1+\eta^{-1})} \frac{\varphi_{\min}}{29d\varphi_{\max} + 6\alpha\tau} = \frac{1}{(d+2)D(1+\eta^{-1})} \frac{\varphi_{\min}}{c_1 d\varphi_{\max} + c_2\alpha\tau}$$

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

### 5.5.3 Concluding remarks

In this chapter we used Wasserstein distance to quantify the deviation of a random sample from a manifold, and used simultaneous-local concentration inequalities to give

probabilistic bounds on how well Local PCA works assuming the manifold hypothesis. In measuring Wasserstein distance, a key challenge was to construct a precise transportation plan and track all of the distances involved. The main theorems thus rely on many careful calculations.

We use analogous techniques to those used in this chapter to prove the main theorem in the upcoming chapter, on the theoretical guarantee for a singularity detection algorithm (Theorem 6.1.1). In fact, this chapter was developed as an important component of the singularity detection theorem. In the next chapter, we will again use Wasserstein distance to quantify deviation of measures, for the setting of a union of two manifolds. Instead of using a concentration inequality of covariance matrices, a concentration inequality of Wasserstein distance will be used.

Separately to the singularity detection algorithm, Local PCA itself plays a foundational role to understanding a manifold from a finite sample. This is because its philosophy of local linearisation, which is the very same principle underlying calculus. The approach taken in this chapter is modular, clear, and explicit, so that one could further sharpen the main theorems of this chapter by improving each module of the proof.

# Chapter 6

# Singularity detection - Theory

## 6.1 Introduction

In this chapter we define *singularity score*, an estimator that detects singularities in stratified spaces. The algorithmic implementation of the singularity score is done in the next chapter (Chapter 7), and in this chapter we purely study the theoretical properties of the singularity score.

The singularity score is calculated by the following simple heuristic observation: when a manifold is zoomed in close, it resembles a flat disk. More precisely, the singularity score is calculated in the following steps:

1. Given a point cloud, isolate the local neighborhoods near each data point.

2. At each local neighborhood, use PCA projection to reduce data dimensionality.

3. Construct an empirical measure with the projected local data sample, and calculate the statistical distance between the empirical measure and $\mathfrak{u}_d$. Here $\mathfrak{u}_d$ is the uniform distribution over the unit $d$-dimensional disk.

Our choice of statistical distance is the kernel MMD distance, which we found to be particularly effective in detecting singularities in the programming implementation. The proof presented in this Chapter also works for the Wasserstein distance.

We start by recalling the PCA dimension estimator $\hat{d}_\eta$ and linear regression $\mathcal{L}_\eta$, defined

in the beginning of the previous Chapter:

$$\hat{d}_\eta(\mu) = \min\left\{k \;\middle|\; \frac{\lambda_{k+1}+\cdots+\lambda_D}{\lambda_1+\cdots+\lambda_D} \leq \eta\right\}$$
$$\mathcal{L}_\eta(\mu) = \text{span}\left(\mathcal{E}(\mu,\lambda_1),\ldots,\mathcal{E}(\mu,\lambda_{\hat{d}_\eta(\mu)})\right)$$

Here we used a slight abuse of notation and defined $\mathcal{L}_\eta = \mathcal{L}_{\hat{d}_\eta(\mu)}$ as it appeared in the previous Chapter.

We now 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 $\kappa(x,y) = \exp(-\frac{1}{2}\cdot\|x-y\|^2)$. 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*.

**Singularity score.** 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 $\mu$ along the projection to $\mathcal{L}\mu$. Let $\mathbf{x} = \{x_1,\ldots x_n\} \subset \mathbb{R}^D$ and let $r > 0$. Denote $\mathbf{x}(z) = \mathbf{x} \cap \mathcal{B}(z,r)\backslash\{z\}$, where $\mathcal{B}(z,r) \subseteq \mathbb{R}^D$ is the open ball of radius $r$, centred at $z$[1]. 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 $\nu$ 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 $\eta$. We now state the setup and the main theorem.

**Main theorem**

**Theorem 6.1.1.** *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 $T_xM_1$ and $T_xM_2$ intersect at a $d_0$-dimensional subspace, and all principal angles of the pair are $\geq \phi$. Let $\mu$ be the uniform measure over $M$, and let $\mathbf{X}_n = (X_1, \dots X_n)$ be an iid sample of size $n$ drawn from $\mu$.*

*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:*

- *When the distance of $X_i$ from $M_1 \cap M_2$ is less than $c_A r$, then $\hat{\sigma}_i > 2\xi$.*
- *When the distance of $X_i$ from $M_1 \cap M_2$ is greater than $c_B r$, then $\hat{\sigma}_i < \xi$.*

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

A prominent tool for the theorem is the Wasserstein distance, instead of the kernel MMD, which is possible since $\Delta(\mu, \nu) \leq \sqrt{2\gamma} \cdot \mathrm{W}(\mu, \nu)$ for the Gaussian kernel $\kappa(x, y) = e^{-\gamma\|x-y\|^2}$ (Lemma 3.2.6). The advantage of the Wasserstein distance is that it is intuitively easy to prove geometric claims. The proof of the main theorem consist of the following 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 $T_zM^\circ := T_zM \cap \mathcal{B}(0, 1)$, where $\mathcal{B}(0, 1) \subseteq \mathbb{R}^D$ is the unit ball of radius 1. Convergence is quantified using the Wasserstein distance. (Proposition 4.2.7)

2. The singularity score function $\mu \mapsto \sigma(\mu)$ is a Lipschitz continuous function in $\mu$, where Lipschitz continuity is quantified using the Wasserstein distance. (Proposition 6.4.1)

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 6.4.4, 6.4.5).

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 2.1.14).

To understand the proof, the reader is advised to start from the last part, Subsection 6.5, 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 algorithm. 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 $c \cdot r$ 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.

## 6.2 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 6.2.1.** 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:

$$\begin{aligned} \lambda_k^{\text{gap}} A =& \lambda_k A - \lambda_{k+1} A \\ \text{Tail}_k \, A =& \lambda_{k+1} A + \cdots + \lambda_D A \\ \text{TQ}_k \, A =& \frac{\text{Tail}_k \, A}{\text{Tail}_0 \, A} \end{aligned}$$

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

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

Also using Proposition 5.2.3 proven in the previous chapter, we then get the following bound on the variation of the spectral gap:

**Lemma 6.2.2.** *Let $\mu, \nu \in \mathcal{P}$ be such that the support of each measure is contained in a ball of radius* 1*. Then,*

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

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

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

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

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

Now the claim follows by applying Proposition 5.2.3 and the fact that Frobenius norm satisfies $\|B\|_{\mathrm{F}} \leq \sqrt{D} \cdot \|B\|$ generally for any $B \in \mathbb{R}^{D \times D}$.

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

□

The variation of tail quotient can be controlled as follows:

**Lemma 6.2.3.** *Let $\mu, \nu \in \mathcal{P}$ be such that the support of each measure is contained in a ball of radius* 1*. Assume that $\mathrm{W}(\mu, \nu) \leq \beta/(16D)$, where $\beta = \|\vec{\lambda}\Sigma\mu\|_1$. Then the following holds for all $k$:*

$$|\operatorname{TQ}_k(\mu) - \operatorname{TQ}_k(\nu)| \leq 32D\beta^{-1} \cdot \mathrm{W}(\mu, \nu)$$

*Proof.* Denote $A = \Sigma\mu, A' = \Sigma\nu$. The Hoffman-Wielandt theorem and Proposition 5.2.3 imply:

$$\begin{aligned} |\operatorname{Tail}_k A - \operatorname{Tail}_k A'| =& \left| \sum_{i>k} \lambda_k A - \lambda_k A' \right| \leq \sum_{i>k} |\lambda_k A - \lambda_k A'| \leq \|\vec{\lambda}A - \vec{\lambda}A'\|_1 \\ \leq& \sqrt{D} \cdot \|\vec{\lambda}A - \vec{\lambda}A'\|_2 \leq \sqrt{D} \cdot \|A - A'\|_{\mathrm{F}} \leq 8D \cdot \mathrm{W}(\mu, \nu) \end{aligned}$$

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 $\|B\|_{\mathrm{F}} \leq \sqrt{D} \cdot \|B\|$ generally for any $B \in \mathbb{R}^{D\times D}$. Define the following notations:

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

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:

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

□

## 6.3 Covariance of two disks

The following proposition is required to understand spectral gap of the localised measure of a union of two manifolds.

**Proposition 6.3.1.** *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 $\mu_i$ is the uniform measure over $\pi_i \cap \mathcal{B}(0, 1)$. Then eigenvalues of the covariance of $\mu$ 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 $(\pi_1, \pi_2)$ with $\theta_1 = \ldots = \theta_{d_0} = 0 < \theta_{d_0+1}$.*

*Proof.* By starting from the matrix form in Corollary 5.4.4 and then by applying multiple 2-dimensional rotations to the standard matrix form of $(\pi_1, \pi_2)$, 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 $\pi_i$:

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

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 $\pi_i$.

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. □

## 6.4 Singularity score

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.

**Proposition 6.4.1.** *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:*

$$\begin{aligned} \mathfrak{S}(\mathcal{L}_\eta\mu, \mathcal{L}_\eta\nu) \leq & 16\sqrt{d}s^{-1} \cdot \mathrm{W}(\mu,\nu) \\ |\sigma_\eta(\mu) - \sigma_\eta(\nu)| \leq & \sqrt{2\gamma}(1 + 16\sqrt{d}s^{-1})\,\mathrm{W}(\mu,\nu) \end{aligned}$$

*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 $(\pi_1, \pi_2)$.*

*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 5.2.3:

$$\mathfrak{S}(\mathcal{L}\mu, \mathcal{L}\nu) \leq \frac{2\sqrt{d}}{s}\|\Sigma\mu - \Sigma\nu\| \leq \frac{16\sqrt{d}}{s}\,\mathrm{W}(\mu,\nu)$$

For the singularity score, we get:

$$|\sigma(\mu) - \sigma(\nu)| = |\Delta(\mu_\perp, \mathfrak{u}_k) - \Delta(\nu_\perp, \mathfrak{u}_k)| \leq \Delta(\mu_\perp, \nu_\perp) \leq \sqrt{2\gamma}\,\mathrm{W}(\mu_\perp, \nu_\perp)$$

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

$$\begin{aligned}
\mathrm{W}(\mu^\perp, \nu^\perp) &= \mathrm{W}\left(\Pi(\mu, \mathcal{L}\mu), \Pi(\nu, \mathcal{L}\nu)\right) \\
&\leq \mathrm{W}\left(\Pi(\mu, \mathcal{L}\mu), \Pi(\nu, \mathcal{L}\mu)\right) + \mathrm{W}\left(\Pi(\nu, \mathcal{L}\mu), \Pi(\nu, \mathcal{L}\nu)\right) \\
&\leq \mathrm{W}(\mu,\nu) + \mathfrak{S}(\mathcal{L}\mu, \mathcal{L}\nu) \\
&\leq \left(1 + \frac{16\sqrt{d}}{s}\right)\mathrm{W}(\mu,\nu)
\end{aligned}$$

where in the second to last inequality we applied Lemmas 3.1.4 and 3.1.5. □

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

**Proposition 6.4.2.** *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 $\mathrm{W}(\mu,\nu)$ is sufficiently small; explicitly, assume that:*

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

*where $\beta = \|\vec{\lambda}\Sigma\mu\|_1$, $G_k = \mathrm{TQ}_{k-1}(\mu) - \mathrm{TQ}_k(\mu)$, and $s = \lambda_k^{\mathrm{gap}}(\mu)$. Then the following hold:*

*(1) There is a bound for spectral gaps:*

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

*(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* $J_{k,a}(\mu) = [\mathrm{TQ}_k(\mu) + \frac{a}{2}G_k,\ \mathrm{TQ}_{k-1}(\mu) - \frac{a}{2}G_k]$.

*(3) For a choice of* $\eta \in J_{k,a}(\mu)$*, the following hold:*

$$\begin{aligned}
&\hat{d}_\eta(\mu) = \hat{d}_\eta(\nu) = k \\
&\mathfrak{S}(\pi_1, \pi_2) \leq 16\sqrt{d}s^{-1} \cdot \mathrm{W}(\mu, \nu) \\
&|\sigma_\eta\mu - \sigma_\eta\nu| \leq \sqrt{2\gamma}(1 + 16\sqrt{d}s^{-1}) \cdot \mathrm{W}(\mu, \nu)
\end{aligned}$$

*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* $(\pi_1, \pi_2)$*.*

*Proof.* (1) By Lemma 6.2.2 and the assumption on $\mathrm{W}(\mu, \nu)$, we have:

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

(2) By Lemma 6.2.3 and the assumption on $\mathrm{W}(\mu, \nu)$, for each $j$ we have:

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

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 6.4.1. □

We recall the following definition given in Chapter 3:

**Definition 6.4.3.** Let $\mu$ be a Borel probability measure on $\mathbb{R}^D$. For $x \in \mathrm{supp}(\mu)$ and $r > 0$, define the following:

$$\mu_{x,r} := g_{x,r}(\mu|_{\mathcal{B}(x,r)})$$

where $g_{x,r}(\nu)$ is the pushforward of the measure $\nu$ along the affine linear map $y \mapsto r^{-1}(y - x)$. Furthermore, define the following limit, if it exists:

$$\mu_{x,0} := \lim_{r \to 0} \mu_{x,r}$$

where convergence is measured using the Wasserstein distance. In other words, $\mu_{x,0}$ is the unique measure satisfying $\lim_{r\to 0} \mathrm{W}(\mu_{x,r}, \mu_{x,0}) = 0$.

We establish some limit behaviour of the singularity score.

**Proposition 6.4.4.** *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 3.1.8, 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. □

**Proposition 6.4.5.** *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 6.3.1, 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:

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

Therefore we can apply Proposition 6.4.1, 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:

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

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. □

## 6.5 Proof of the main theorem

In this section we prove Theorem 6.1.1, the main theorem of this chapter.

**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.

$\tau, \psi, \zeta, s_0$ are defined as:

$$\tau = \min(\tau_1, \tau_2), \quad \psi = \frac{\sin^2(\phi/2)}{d+2}, \quad \zeta = \frac{\psi \cdot d}{128(d+2)}, \quad s_0 = \sqrt{2\gamma}\left(1 + \frac{16\sqrt{d}}{\psi}\right)$$

$\eta_-, \eta_+$ are defined as:

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

$\xi, \xi_0$ 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:

$$\begin{aligned} r_0 &= \min\left(c_5, c_7, c_9, \frac{\xi_0}{2c_8}, \frac{\xi_0}{8c_6}, \frac{\xi_0}{4c_{10}}\right) \cdot \tau \\ c_A &= \min\left(c_5, \frac{\xi_0}{8c_6}\right) \\ c_B &= \max\left(h(M_1, M_2)^{-1}, h(M_2, M_1)^{-1}\right) \\ \text{where } h(M_1, M_2) &= \inf_{x \in M_1} \frac{\mathrm{d}(x, M_2)}{\mathrm{d}(x, M_1 \cap M_2)} \end{aligned}$$

where the constants $c_5, \dots c_{10}$, which depend only on $d$, are defined in Propositions 3.1.8, 3.1.7, and 3.1.9.

Finally, we *fix a choice* of $\eta, r$ 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 $\eta, \eta'$, then we have $\sigma_\eta(\nu) = \sigma_{\eta'}(\nu)$. In particular, due to Propositions 6.3.1, 6.4.5, $x \in M_1 \cap M_2$ 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 $\mathbf{x} = (x_1, \dots x_n) \subset M$. 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:

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

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

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 $\xi$ 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 6.4.5 and 6.4.4, 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:

$$\sigma_i \geq \sigma(\mu_{y,0}) - |\sigma(\hat{\mu}_i) - \sigma(\mu_{y_i,0})| = 3\xi - |\sigma(\hat{\mu}_i) - \sigma(\mu_{y_i,0})| \tag{6.5.1}$$

where $y_i \in M_1 \cap M_2$ is a point satisfying $\mathrm{d}(x_i, y_i) = \mathrm{d}(x_i, M_1 \cap M_2)$[3].

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

$$\sigma_i \leq \sigma(\mu_{x_i,0}) + |\sigma(\hat{\mu}_i) - \sigma(\mu_{x_i,0})| = 0 + |\sigma(\hat{\mu}_i) - \sigma(\mu_{x_i,0})| \tag{6.5.2}$$

Even though Equations (6.5.1) and (6.5.2) 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.**

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

Differences of singularity scores are controlled using Proposition 6.4.2, which is a Lipschitz continuity relation with respect to the Wasserstein distance. Our definition of $\zeta$ is obtained by setting $a = 1/2$, $\beta = d/(d+2)$, $G = s = \psi$ in the condition in Proposition 6.4.2. This then implies that for all $x \in M$ and $\nu \in \mathcal{P}$,

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

where we recall our definition $s_0 = \sqrt{2\gamma}(1 + 16\sqrt{d}\psi^{-1})$. The definition of $\xi_0$ allows us to make a more straightforward inference:

$$\mathrm{W}(\mu_{x,0}, \nu) \leq \xi_0 \implies |\sigma(\mu_{x,0}) - \sigma(\nu)| \leq \xi \tag{6.5.3}$$

Therefore we can control error terms in Equation (6.5.1), (6.5.2) 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 3.1.8 and 3.1.7 imply that:

$$\rho,\ s \leq c_5 \implies \mathrm{W}(\mu_{x,r}, \mu_{y,r}) \leq c_6(\rho + s)$$
$$\rho \leq c_9 \implies \mathrm{W}(\mu_{y,r}, \mu_{y,0}) \leq c_{10}\rho$$

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

$$\mathrm{W}(\mu_{x,r}, \mu_{y,0}) \leq \mathrm{W}(\mu_{x,r}, \mu_{y,r}) + \mathrm{W}(\mu_{y,r}, \mu_{y,0}) \leq \frac{\xi_0}{2} \tag{6.5.4}$$

*Smooth part.* Suppose that $x$ satisfies $\mathrm{d}(x, M_1 \cap M_2) \geq c_B r$. Here our choice of $c_B$ allows us to apply Proposition 2.1.14, 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 3.1.9 implies:

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

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

$$\mathrm{W}(\mu_{x,r}, \mu_{x,0}) \leq \frac{\xi_0}{2} \tag{6.5.5}$$

**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 $\mathbf{X}_n = (X_1, \dots X_n)$ 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 4.2.7[4], and obtain the following. Whenever $n \geq \max(N, 2/u_-)$, we have:

$$\mathbb{P}\left( \max_i \mathrm{W}(\hat{\mu}_i, \mu_{X_i,r}) \leq \frac{\xi_0}{2} \right) \geq 1 - \delta_m \tag{6.5.6}$$

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

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

where

$$c = \left( \frac{u_+}{1-u_+} \right)^N, \quad N = \left\lceil \left( \frac{408}{\xi_0} \right)^D \right\rceil, \quad \gamma = 1 - u_-(1 - e^{-\xi_0^2/32})$$
$$u_- = \inf_{x \in \operatorname{supp}\mu} \mu(\mathcal{B}(x,r)), \quad u_+ = \sup_{x \in \operatorname{supp}\mu} \mu(\mathcal{B}(x,r))$$

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 $\delta, \mu, r, \xi_0$, 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 $1 - \delta$:

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

Equations (6.5.4) and (6.5.5) apply verbatim for the random setting:

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

[4]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 $r^{-1}$.

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

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

This precisely fits the condition in Equation (6.5.3), from which we obtain that:

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

Plugging them into Equations (6.5.1) and (6.5.2), we obtain the conclusion of the theorem:

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

### 6.5.1 Concluding remarks

In this chapter we used the tools developed in the previous sections to prove Theorem 6.1.1, which guarantees that our singularity detection algorithm works correctly in the case of two transversally intersecting manifolds of equal dimensions. This theorem marks the main theoretical achievement of the DPhil thesis, for it uses all of the mathematical techniques developed so far. Calculations leading to the conclusion of the theorem require bounding Wasserstein distances, which often entail many subtle details that require a lot of careful attention.

An extension of the singularity detection theorem would expand its correctness to more general stratified spaces than a union of two manifolds. This requires adapting some key techniques to a more general setting, but ingredients such as Proposition 2.1.14 and Proposition 3.1.7 are nontrivial to prove even for the case of two manifolds. As such, one would need analogous results for general stratified spaces in order to generalise Theorem 6.1.1.

In the upcoming chapter, we will explore the singularity detection algorithm in action. We will see that, due to the simple design of the algorithm using fast, well-understood statistical techniques, the algorithm performs well in datasets, both synthetic and real.

In particular, we will see that it is faster and more principled than other singularity detection algorithms based on topological techniques.

# Chapter 7

# Singularity detection - Experiments

## 7.1 Introduction

In this chapter, we present algorithmic aspects of the singularity score defined in the previous chapter. In fact, we develop further methodologies that aren't covered by the theoretical analyses done in the previous chapter - we go one step further to compute the *singularity p-values* from the *singularity scores*, thereby directly performing a local goodness-of-fit test. Furthermore, a *global* test of whether the geometric space underlying a dataset is a manifold is presented.

This chapter and the previous chapter attempt to overcome a restrictive assumption in data science known as the *Manifold Hypothesis*. The Manifold Hypothesis states that, generally, the data points we encounter are sampled from a manifold, potentially with additional perturbations by noise. It is an appealing hypothesis since manifolds are a general class of geometric spaces that encapsulate non-linear distribution of data.

However, already for many low-dimensional data sets that can be visualized it becomes obvious that the Manifold Hypothesis can be too strong. Instead, real data can have *singularities* – points at which the local geometry of data distribution is non-manifold. Some datasets even have intrinsically singular geometry, such as branching points seen on road networks, neurons, and cosmic filaments. Furthermore, high-dimensional data with non-constant intrinsic dimension is expected to possess singularities, since a connected manifold must possess the same intrinsic dimension everywhere.

We propose Hades (*Hypothesis-testing Algorithm for Detection and Exploration of*

*Singularities*), an algorithm that uses local goodness-of-fit tests to detect singularities. This expands the *singularity score* defined in the previous chapter. The intution of our algorithm is the same as before, and uses two elementary observations: firstly, locally a manifold resembles flat disk; secondly, this resemblance can be quantified with the distance between an empirical measure and a uniform measure. Indeed, given a point in a point cloud, suppose we identify its neighboring points as an empirical measure. If the Manifold Hypothesis holds this measure should resemble the uniform measure on a disk, if not this is evidence for a singularity at this point.

Our algorithm quantifies this resemblance with a goodness-of-fit test, from which we directly obtain a p-value for rejecting the null hypothesis that a data point has a smooth, non-singular neighborhood. We use an explicit formula for kernel MMD (maximum mean discrepancy) to perform the goodness-of-fit test, and this has a time complexity that is linear in the data dimension. This significantly improves the time complexity of previous topological methods, whose time complexity is exponential in the data dimension.

We demonstrate our algorithm Hades on synthetic and real world datasets. On synthetic datasets we observe through scatterplots and AUC score (area-under-curve) that singularities are correctly detected (Figure 7.5, 7.6, 7.6). We use three real world datasets - road networks, cyclo-octane conformations, and images of hand-written digits and clothings. In the road network dataset, consisting of points on the 2D plane that trace aerial photographs of roads, the algorithm is able to cleanly extract intersections and sharp corners on roads (Figure 7.8). In the cyclo-octane conformations dataset, consisting of 24-dimensional points that represent all possible positions of carbon atoms in a cyclo-octane, the algorithm recovers the known singular structure in the dataset, given by two circles formed by intersection of a Klein bottle and a sphere (Figure 7.9). In the image datasets, the algorithm identifies anomalous images that deviate heavily from the same type of images in the class (Figure 7.10, 7.11).

**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 [46], 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 [86, 100, 85, 13, 11, 12]. Evidence for real world data containing multiple manifolds of mixed dimension have been recently studied [29, 30, 67]. 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 information at multiple scales of data. In [17, 18, 19], persistent intersection homology was used to discover stratified structure of data. In [99, 26, 27], 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 [71] and a complex projective variety [48] has also been studied. In [88, 98], 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 [60, 33, 32, 39, 50, 20, 98]. Dimension reduction methods in the literature include [16, 103, 89, 68, 95]. For a survey of dimension estimation and dimension reduction algorithms, see [31, 96].

## 7.2 Algorithm

The main intuition for Hades is that a manifold locally resembles a flat disk, and this resemblance can be measured with a goodness-of-fit test. This property fails at singular points; for example, a branching point of a planar graph locally spans 2 dimensions, but it doesn't span the full 2-dimensional disk. The flat disk that a manifold resembles locally is a subset of the tangent space to the manifold. Therefore, it is necessary to first project the

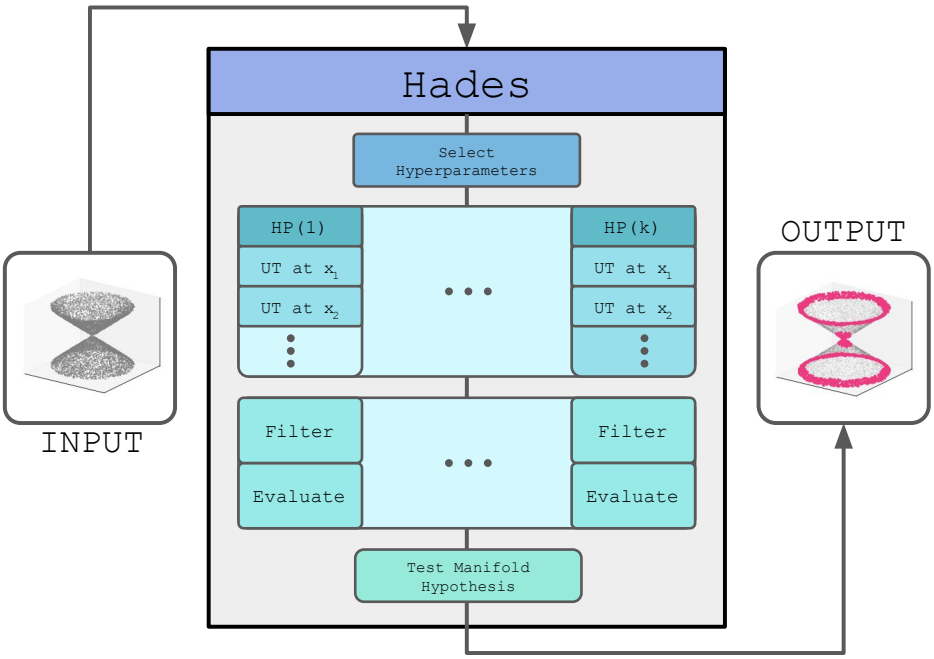


Figure 7.1: Schematic for a fully automated run of Hades. UT stands for Uniformity Test.

local sample to the tangent space. After this, we assess deviation of the local empirical measure from the uniform distribution over the flat disk. Our method is a goodness-of-fit test using the kernel maximum mean discrepancy (MMD). Using the asymptotic distribution of the kernel MMD, we can compute the p-value of the goodness-of-fit test.

### 7.2.1 Main loop

We first explain the main loop of the algorithm, which performs local goodness-of-fit tests over query points to obtain p-values and then chooses a cutoff for determining singular points. The main loop requires fixing the hyperparameters $(r, \eta)$, where $r$ is the radius parameter and $\eta$ is the PCA threshold parameter. We explain later in Subsection 7.2.3 how the optimal set of hyperparameters is chosen. Note that Algorithm 1 is an algorithmic implementation of the singularity score defined in the previous chapter, plus a step for calculating p-values.

**Algorithm 1** Hades (Fixed hyperparameters)

**Input**: Points $x_1, \dots x_n \in \mathbb{R}^D$ and hyperparameters $(r, \eta)$.

**Output**: Labels $y_1, \dots y_n \in \{0, 1\}$, singularity p-values $\sigma_1, \dots \sigma_n \in \mathbb{R}$, and a cutoff $\xi$.

**for** $z \in \{x_1, \dots x_n\}$ **do**

1. Get a local neighborhood $\mathbf{z} \subseteq \mathbf{x}$ near $z$, using $r$.

2. Estimate $\hat{d}$, the intrinsic dimension of $\mathbf{z}$, using $\eta$.

3. Compute $\tilde{\mathbf{z}}$, a $\hat{d}$-dimensional projection of $\mathbf{z}$.

4. Perform a goodness-of-fit test of the empirical distribution defined by $\tilde{\mathbf{z}}$, against the uniform distribution over the unit $\hat{d}$-dimensional disk. The test is performed using a null distribution of kernel MMD. The $p$-value of the goodness-of-fit test is the singularity score $\sigma(z)$.

**end for**

Select a global cutoff value $\xi$ so that the points $x_i$ satisfying $\sigma(x_i) < \xi$ are labelled $y_i = 1$, and declared to be singular points.

**Remark.** In our methods, we use a formula for directly measuring the kernel MMD between an empirical measure and the uniform measure over a disk (Theorem 7.2.1). As such, we don't need to produce a random sample from the disk when performing the goodness-of-fit tests, and makes our method conceptually clean.

We now provide details of computations for each $z \in \{x_1, \dots x_n\}$,

**Step 1** : *Neighbor determination.* We use the radius parameter $r$ to isolate local neighborhoods. The neighborhood of $z$ is defined as $\mathbf{z} = \{r^{-1}(x_j - z) \mid \|x_j - z\| \leq r\}$. In $\mathbf{z}$, each point $x_j$ is translated and rescaled so that $\mathbf{z}$ fits into a unit ball. [1]

**Step 2** : *Dimension estimation.* We use PCA with variance threshold $\eta$. Given the local neighborhood $\mathbf{z}$, let the empirical covariance matrix be $S = n_z^{-1} \sum_{w \in \mathbf{z}} (w - \bar{z})(w - \bar{z})^\top$ where $n_z$ is the number of points in $\mathbf{z}$ and $\bar{z}$ is the mean of $\mathbf{z}$. Let $\lambda_1 \geq \dots \geq \lambda_D$ be the eigenvalues of $S$. Our local estimated dimension $\hat{d}$ is defined as the largest $k$ such that $\lambda_{k+1} + \dots + \lambda_D \leq \eta \cdot (\lambda_1 + \dots + \lambda_D)$.

[1] We also implemented isolating neighborhoods consisting of $k$ nearest-neighbors, where $k$ is fixed in advance.

**Step 3** : *Dimension reduction.* We use PCA projection. Suppose $S = U\Lambda U^\top$ is a diagonalisation of $S$. Equipped with the estimated dimension $\hat{d}$, let $\tilde{U}$ be the matrix of leftmost $\hat{d}$ columns of $U$. The projected local sample is defined as $\tilde{\mathbf{z}} = \tilde{U}\tilde{U}^\top \mathbf{z}$.

**Step 4** : *Uniformity test.* We use a goodness-of-fit test using kernel methods. Given $\tilde{\mathbf{z}}$ obtained from projection, define the empirical measure $\hat{\mu}_z = n_z^{-1} \sum_{x \in \tilde{\mathbf{z}}} \delta_x$, where $\delta_x$ is the Dirac delta measure centered at $x$. Let $\mathfrak{u}_{\hat{d}}$ be the uniform measure over the unit $\hat{d}$-dimensional disk. We define the *singularity score* as follows::

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

where $\Delta$ is the MMD associated to a kernel $\kappa$ that is fixed ahead in time. Finally our *singularity p-value* $\sigma$ is defined as the p-value obtained from comparing $\tilde{\sigma}$ against a null distribution. Let $\hat{\nu} = n_z^{-1} \sum_{x \in \mathbf{Z}} \delta_x$ be the empirical measure constructed from $\mathbf{Z}$, where $\mathbf{Z}$ is an i.i.d. sample of size $n_z$ drawn from $\mathfrak{u}_{\hat{d}}$. Let $\Phi$ be the cumulative density function associated to the random variable $\Delta(\hat{\nu}, \mathfrak{u}_{\hat{d}})$. Then $\sigma$ is defined as:

$$\sigma(z) = 1 - \Phi(\tilde{\sigma}(z))$$

**Cutoff selection.** After performing the Steps 1-4 above for $i = 1, \ldots n$, singularity scores $\sigma(x_1), \ldots \sigma(x_n)$ are obtained. We now describe how to obtain a cutoff value $\xi$ so that we give a label $y_i = 1$ whenever $\sigma_i \leq \xi$, and conversely give the label $y_i = 0$ whenever $\sigma_i > \xi$.

If $x_i$ is sampled near a singularity, we expect $\sigma(x_i)$ to be very small, since it is the p-value obtained from the uniformity test. Equivalently, this means that $-\log \sigma(x_i)$ is large. Under the assumption that each type of singular geometry contributes to a concentrated mass in the distribution of $\sigma_i$, we thus seek a dip in density of the distribution of $-\log(\sigma_i)$.

Let $\varphi$ be the density function of $-\log(\sigma_i)$, obtained through methods like histogram or kernel density estimation. Since singular strata has measure zero in a stratified space[2], only a small proportion of a random sample sampled from a stratified space is sampled near the singular strata. Thus in order to detect the small dip in density, we examine

[2]If $M$ is a $d$-dimensional stratified space with dense top strata, then the $d$-dimensional Hausdorff measure of the singular points of $M$ is zero.

$-\log\varphi$. Finally, this dip is detected by applying the Kneed algorithm [79] to the function $-\log\varphi$.

**Remark.** There are many choices for each of the 4 steps above that work independently of other steps. These may replace the current choices to improve the algorithms in the future. For example, a more sophisticated application might use UMAP [68] to perform local dimensionality reduction, and statistical distances such as the Wasserstein distance could be used in place of kernel MMD in the last step of Algorithm 1. 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 [35].

### 7.2.2 Uniformity Test

We describe details in evaluating MMD and the null distribution in the goodness-of-fit test at Step 4 of Algorithm 1. The MMD is evaluated using the following *explicit* expression for a power series kernel which computes in $O(m^2)$ where $m$ is the size of the empirical distribution. This was proven earlier in Chapter 3, stated as Theorem 3.2.7.

**Theorem 7.2.1.** *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 $\kappa$ be a kernel given by $\kappa(x,y) = \sum_{k=0}^{\infty} a_k \langle x, y\rangle^k$, and let $\Delta$ be the MMD associated to $\kappa$. 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 $\Gamma$ 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. The MMD is a V-statistic, for which asymptotic convergence under scaling by sample size holds true (Section 5, [82]):

**Theorem 7.2.2.** *Let $\mu$ 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 $\mu$. Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function satisfying $\kappa(x,y) = \kappa(y,x)$,*

*and let $\Delta$ be the MMD associated to $\kappa$. Then there is a convergence 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 $\lambda_k$ 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. Informed by [78], we use exponential decay to estimate probabilities of very low p-value, lying outside the domain of simulation of Monte Carlo.

### 7.2.3 Hyperparameter selection

Algorithm 1 describes Hades for one set of hyperparameters. To choose the optimal set of hyperparameters, we define the *dispersion score*, which is made to be minimised among outputs of all hyperparameters. The *dispersion score* measures the total degree of *local dispersion* defined by a binary label on a point set. The full run of Hades performs Algorithm 1 over a (2-dimensional) grid of hyperparameters $(r, \eta)$ and chooses the output that produced the smallest dispersion score. The dispersion score is defined using *purity score* and *separation score*:

**Definition 7.2.3.** Let $\mathbf{x} = (x_1, \dots x_n)$, $\mathbf{y} = (y_1, \dots y_n)$ be points and their binary labels, $x_i \in \mathbb{R}^D$, $y_i \in \{0, 1\}$. For each $i = 1, \dots n$, let $\mathcal{N}(i) \subseteq \{1, \dots n\}$ be a set satisfying $i \in \mathcal{N}(i)$. Define a partition $\mathcal{I}_0 \sqcup \mathcal{I}_1 = \{1, \dots n\}$, 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:

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

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, $\alpha$ 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 $\mathcal{D}_i(x) \leq x$. [3]

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 $\mathfrak{D}$.

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 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 7.5, the singular points marked in blue are not point-like clusters sought by the classical clustering quality measures.)

[3] 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 $F_{a,b}(t) = (\frac{t-a}{1-a})^b$.

### 7.2.4 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 $X_1, \dots X_n$ drawn from a geometric space $M$ and singularity p-values $\sigma_1, \dots \sigma_n$ calculated from them, the following should hold:

- If $M$ is a manifold, then $\sigma_1, \dots \sigma_n$ should distribute uniformly over $[0, 1]$.
- If $M$ has a singularity, then $\sigma_1, \dots \sigma_n$ should be concentrated near 0.

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

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

Now for a fixed $i$, consider the singularity p-value $\sigma_i$ calculated at the neighborhood of $X_i$ by using the random sample $X_1, \dots X_n$ 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 $X_1, \dots X_n$ landing in the local neighborhood of $X_i$, we see that this marginal distribution of $\sigma_i$ approximates $\tilde{Z}_k$. Here, $Z_k = \Delta(\hat{\nu}_k, \mathfrak{u}_d)$ where $\hat{\nu}_k$ 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 $\sigma_i$ 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 $X_1, \dots X_n$ do not overlap, and we may expect the singularity scores $\sigma_1, \dots \sigma_n$ 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 $\{q_1, \dots q_k\} \subset [0, 1]$ and for each $q \in \{q_1, \dots q_k\}$, calculate

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

2. UPUP (Uniformity p-value Uniformity p-value) Construct an empirical distribution $\hat{\nu}$ from $\sigma_1, \dots \sigma_n$ and perform the uniformity test using the kernel MMD method developed in this chapter (Theorem 7.2.1).

3. KS (Kolmogorov-Smirnov) Again construct $\hat{\nu}$ from $\sigma_1, \dots \sigma_n$ and perform the one-sample Kolmogorov-Smirnov test against the uniform distribution over $[0, 1]$.

### 7.2.5 Computational Complexity

We now compute the computational complexity of Hades. Suppose $\mathbf{x} \subset \mathbb{R}^D$ is a $D$-dimensional dataset consisting of $n$ points. Suppose that neighborhoods of the query points, defined by points within distance $r$, contain $k$ points in average.[4] Suppose that $\mathbf{x}$ lies on a $d$-dimensional stratified space. Suppose further that tangent cones at all points of the stratified space are at mots $d_0$-dimensional. Given this information, we now account for computational complexity of each step in Hades.

**Constructing kd-tree.** In order to isolate neighborhoods in Step 1, we use the kd-tree algorithm. Construction of the kd-tree on $n$ points has the time complexity of:

$$T_0 = O(n \log n)$$

**Step 1.** On a single query point, retrieving $k$ nearest neighbors has the time complexity of:

$$T_1 = O(k \log n)$$

[4] Alternatively, we can assume that $k$ was fixed in advance and we isolated neighborhoods consisting of $k$ nearest neighbors of each query point.

**Step 2, 3.** In this step, SVD is performed on a rectangular matrix of dimension $k \times D$, which has the time complexity of $O(\min(D, k) \cdot Dk)$. 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 $O(dk)$. This step thus amounts to the time complexity of:

$$T_{23} = O(Dk^2 + dk)$$

**Step 4.** In this step we compute the MMD of a $\hat{d}$-dimensional point set of size $k$. Following the expression computed in Theorem 7.2.1, the time complexity for this step is:

$$T_4 = O(k^2 d + k + kd) = O(k^2 d)$$

Summing up all of the time complexities, we get the total time complexity of:

$$\begin{aligned} T_{\text{total}} =& T_0 + q(T_1 + T_{23} + T_4) \\ =& O\bigg( n \log n + qk \log n + q(d + D)k^2 \bigg) \end{aligned}$$

Taking $q = n$, and using $d = O(D)$, we get a simpler expression:

$$T'_{\text{total}} = O\bigg( (n \log n) Dk^2 \bigg)$$

We remark that Algorithm 1 can be trivially parallelized since the outputs $\sigma_i$ for $i = 1, \ldots n$ can be computed separately, so that $q$ may be replaced by $q/m_{\text{core}}$, where $m_{\text{core}}$ is the number of computational cores used for parallel computation.

## 7.3 Comparison with other methods

### 7.3.1 Topological algorithms for singularity detection

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* [98, 88, 19, 17, 18, 99, 26, 27]. Persistent homology computes topological features at varying scales of data, and the main idea behind

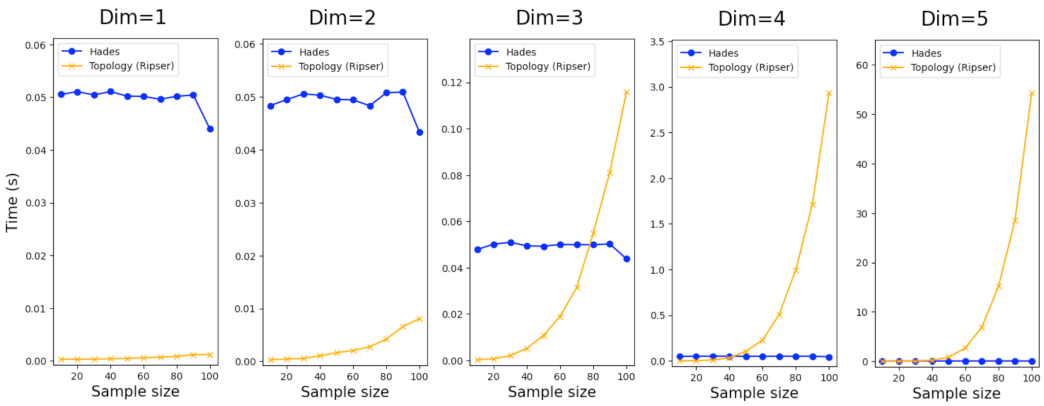


Figure 7.2: Comparison of computation time of local shape analysis in Hades (blue) versus Ripser (orange), a highly optimised library for computing persistent homology.

topological methods for singularity detection is to compute persistent homology on local neighborhoods of data. In particular, the recent algorithms in [98, 88] 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. In comparison, the time complexity of persistent homology increases exponentially in the intrinsic dimension of data. The computational complexity of Ripser [15], 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 Steps 2-4 in Algorithm 1, where in our algorithm we instead use PCA and kernel MMD. Using the computational complexity of Algorithm 1 given above, we have the following comparison of computational complexity incurred by local shape analysis:

$$\begin{aligned} \text{PCA + Kernel goodness-of-fit (Hades):} \quad & O(k^2 D) \\ \text{Persistent homology:} \quad & O(k^{3d+3}) \end{aligned}$$

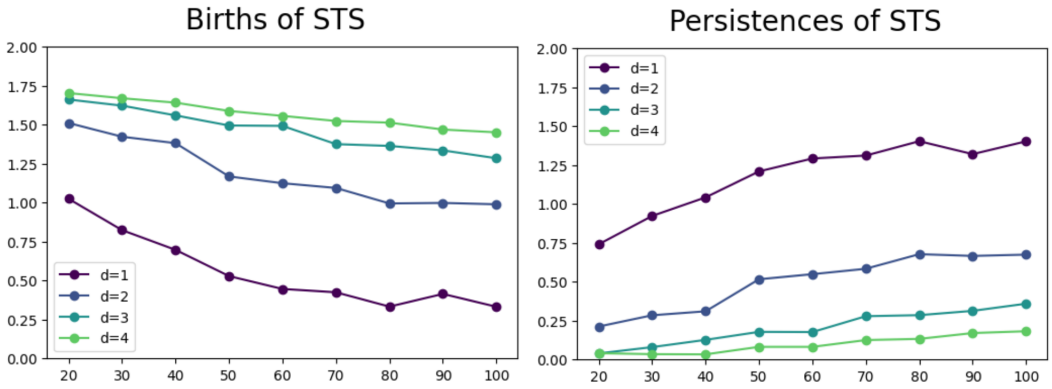


Figure 7.3: Birth and persistence of the STS (significant topological signature) at $(d, n)$.

Thus we observe an exponential dependence of persistent homology computation on the intrinsic dimension $d$ of data, whereas Algorithm 1 has a linear dependence on the ambient dimension $D$.

In Figure 7.2, 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[5], 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 7.3 tabulates birth times and persistences (y-axis)

[5]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.

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 7.3 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 [63, 3], and high-dimensional spheres exhibit systematic low-dimensional topological signals[6].

### 7.3.2 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 7.4 (One-Class SVM [80], Isolation Forest [66], Local Outlier Factor [28]). The first row shows the dataset of two circles and the second row shows the dataset of two disks, and points with high anomaly score are marked in yellow (viridis colormap).

[6]In computational experiments that are yet to be released, the authors observed that a Random Forest classifier can be used to distinguish dimensions of high-dimensional spheres just from using their low-dimensional topological signals.

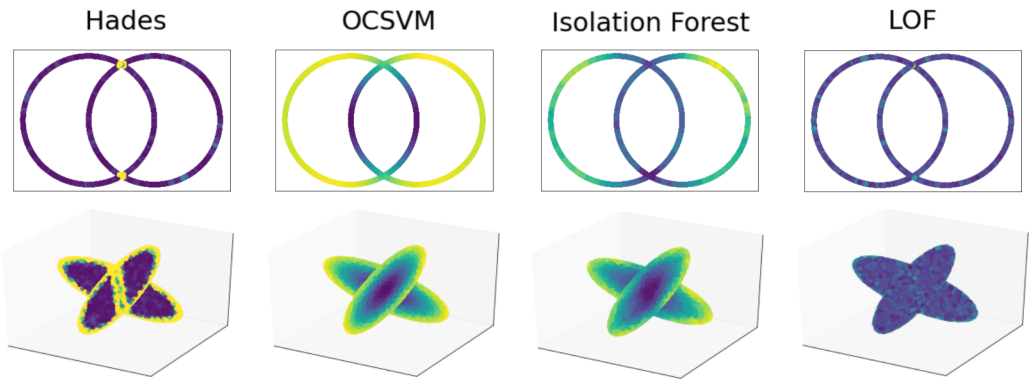


Figure 7.4: Hades is different from the *anomaly detection* algorithms; points marked as highly anomalous are marked in yellow.

## 7.4 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 Section 7.5.

### 7.4.1 Synthetic data: Visualisation and ROC Curves

We first apply Hades to the 2- and 3-dimensional point clouds in Figure 7.5, 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 7.2.4. The sizes of datasets range from 5,000 to 15,000. The time taken to extract singularities from each dataset ranges from 3 minutes to 40 minutes. The time taken per data point ranges from 0.03 seconds to 0.15 seconds.

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

1. One solid $d$-dimensional ball. (Singularity at boundary sphere)
2. Union of two unit $d$-dimensional spheres in $\mathbb{R}^{d+1}$ with centres separated by distance 1, such that they intersect at a $(d-1)$-dimensional sphere. (Singularity at intersection)
3. Union of two unit $2d$-dimensional disks in $\mathbb{R}^{3d}$ that intersect orthogonally at a $d$-dimensional 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 are all $\geq 0.89$. More precisely, the AUC are the following for $d = 1, \ldots 5$:

$$\text{One } d\text{-dimensional solid ball: AUC} = (1.00, 1.00, 1.00, 1.00, 1.00)$$
$$\text{Two } d\text{-dimensional spheres: AUC} = (0.99, 0.93, 0.89, 0.89, 0.89)$$
$$\text{Two } 2d\text{-dimensional disks: AUC} = (0.95, 0.93, 0.94, 0.95, 0.96)$$

The ROC curves and the AUC values are shown in Figures 7.6.

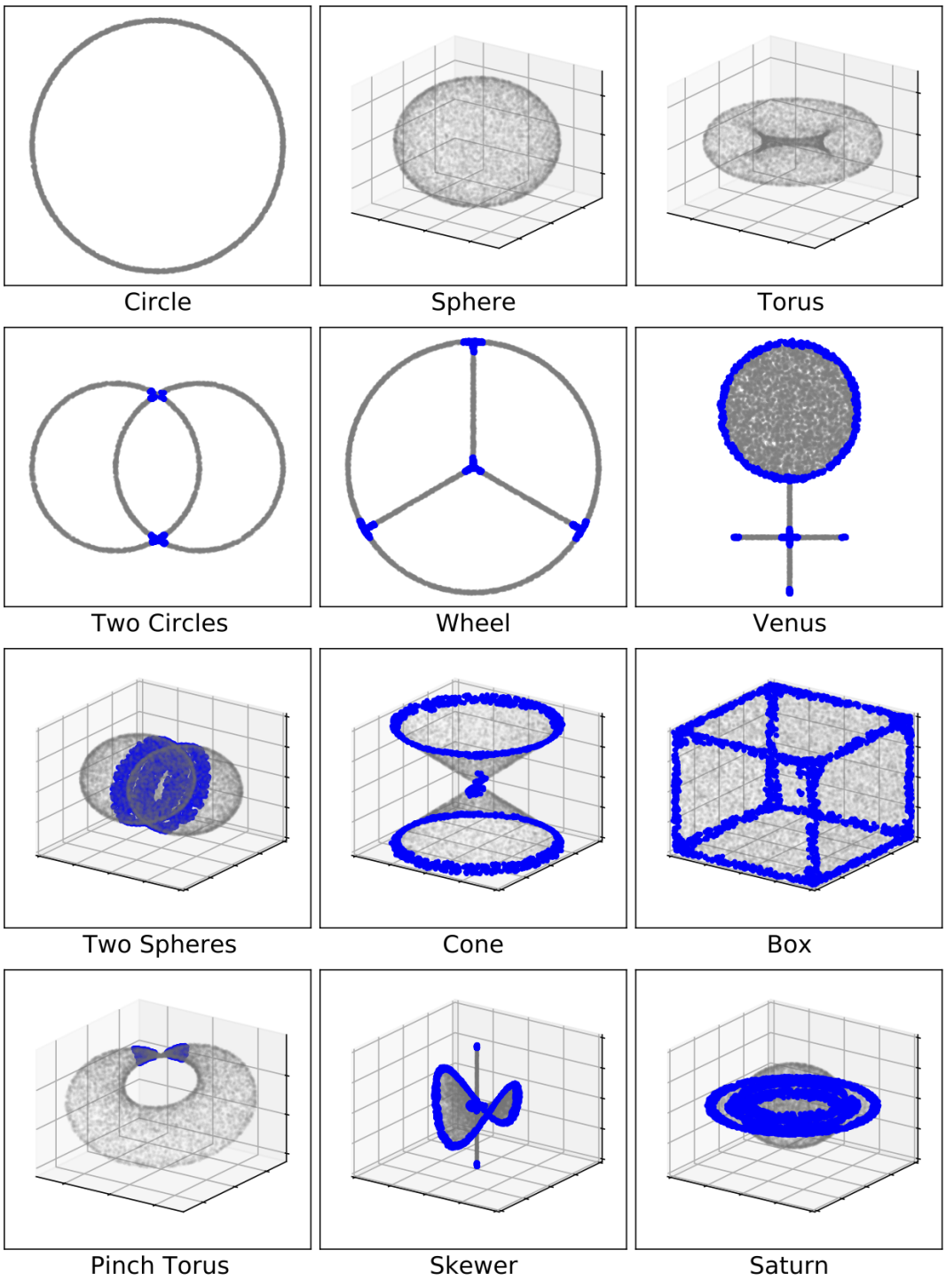


Figure 7.5: Singularities discovered by Hades marked blue in synthetic datasets.

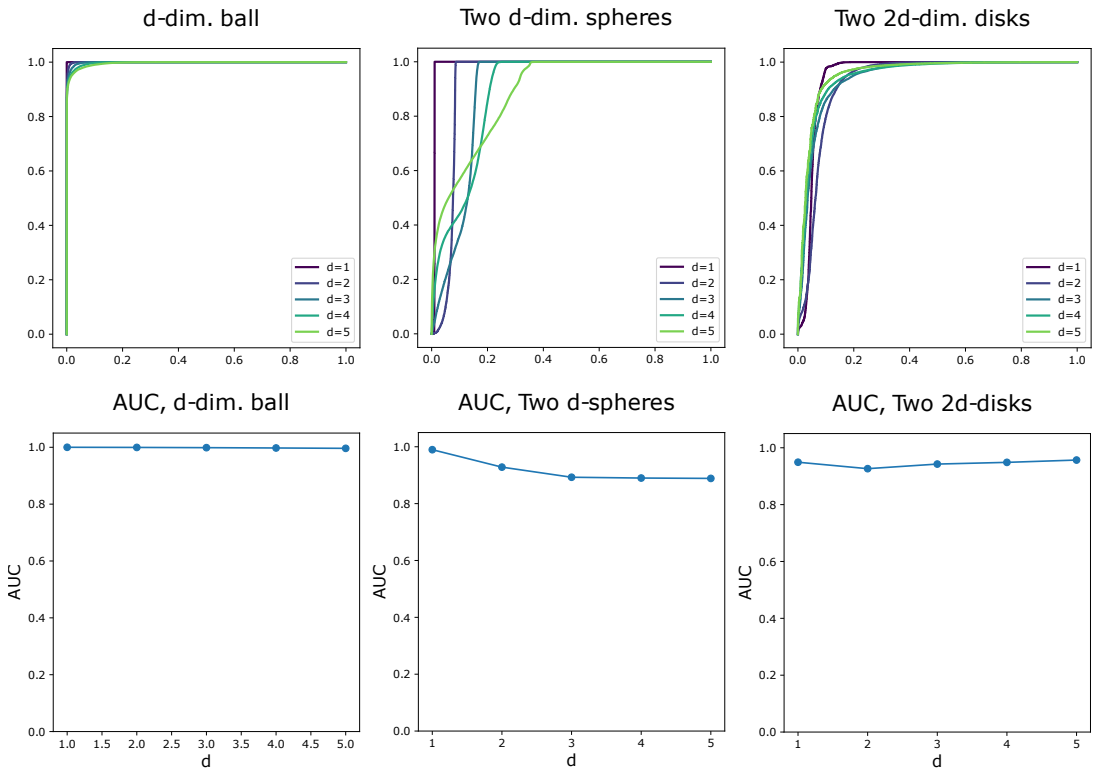


Figure 7.6: ROC curve and AUC scores of singularities discovered by Hades in synthetic datasets.

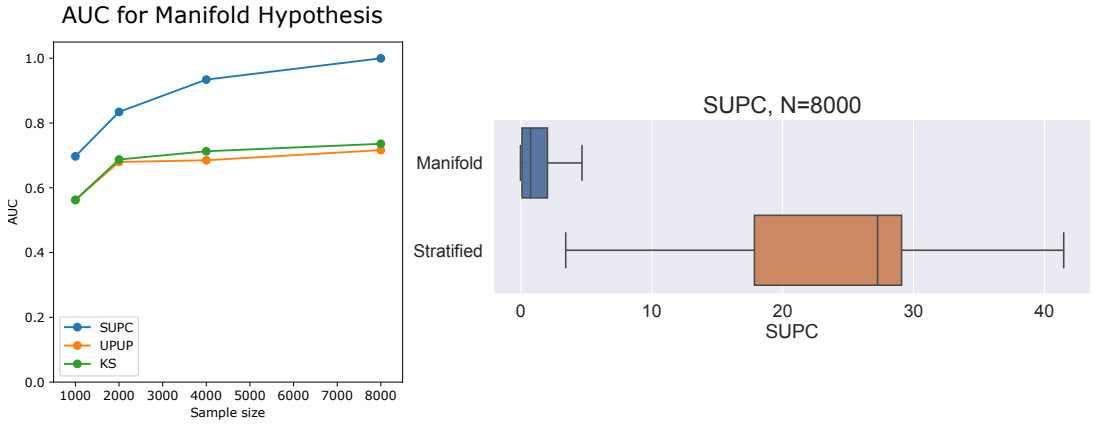


Figure 7.7: Testing the manifold hypothesis with Hades. Left: AUC of manifold hypothesis performed with three different scores: SUPC, UPUP, KS with synthetic datasets. Right: Boxplots of SUPC for manifolds vs. stratified spaces at sample size 8000.

### 7.4.2 Synthetic data: Manifold hypothesis

We test the manifold hypothesis with Hades using the methods described in Subsection 7.2.4. 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). 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 $N \in \{1000, 2000, 4000, 8000\}$ (see the $x$-axis on the left plot of Figure 7.7), 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.

Figure 7.7 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, demonstrat-

ing a clean separation between the two types of data. This indicates that the manifold hypothesis can be effectively tested with SUPC.

### 7.4.3 Real data: Road network

We apply Hades to the Massachusetts Roads Dataset [70], 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 7.8, visual inspection reveals that singularities are accurately detected. Each image had $1500 \times 1500$ 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.

### 7.4.4 Real data: Cyclo-octane conformation

We apply Hades to the dataset of cyclo-octane conformations. This dataset, introduced in [67], 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 [67]. These two circles are singularities of the space of conformations, and indeed they are correctly detected by Hades, as seen in Figure 7.9. The 3D projections of the conformation dataset, obtained using the dimensionality reduction algorithm Isomap [89], is displayed in Figure 7.9; 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 4 seconds on a standard laptop. This shows great improvement from the previous benchmark for this dataset, in [88], in which their singularity detection algorithm *Geometric Anomaly Detection* took at least several hours on parallel processing, as informed by the main author on private communication.

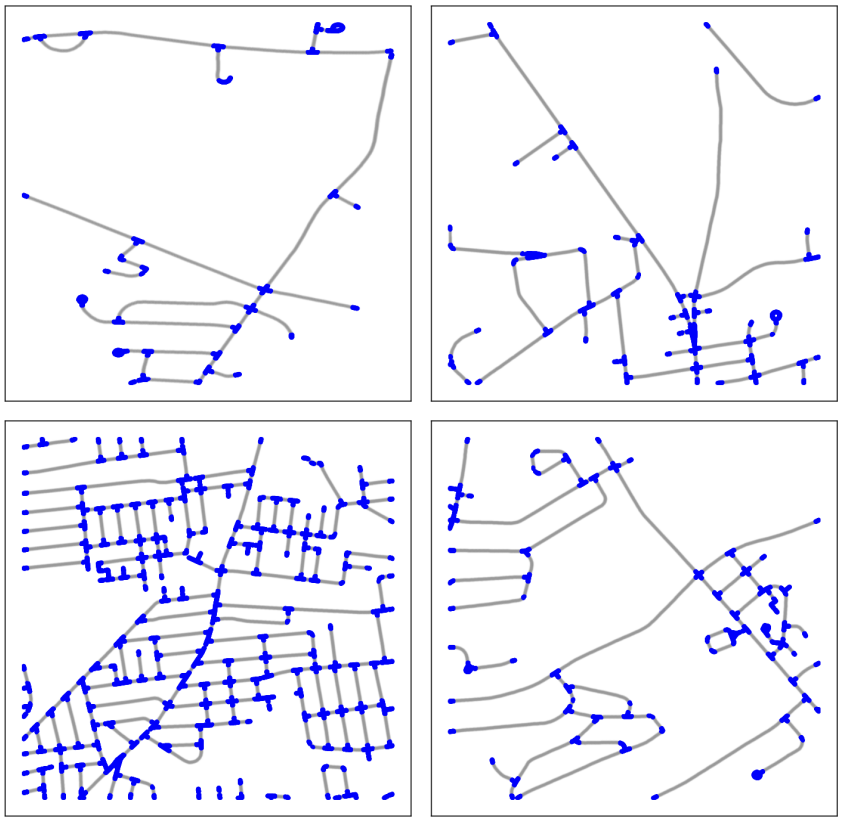

Figure 7.8: Singularities discovered by Hades marked blue in the Massachusetts Roads Dataset.

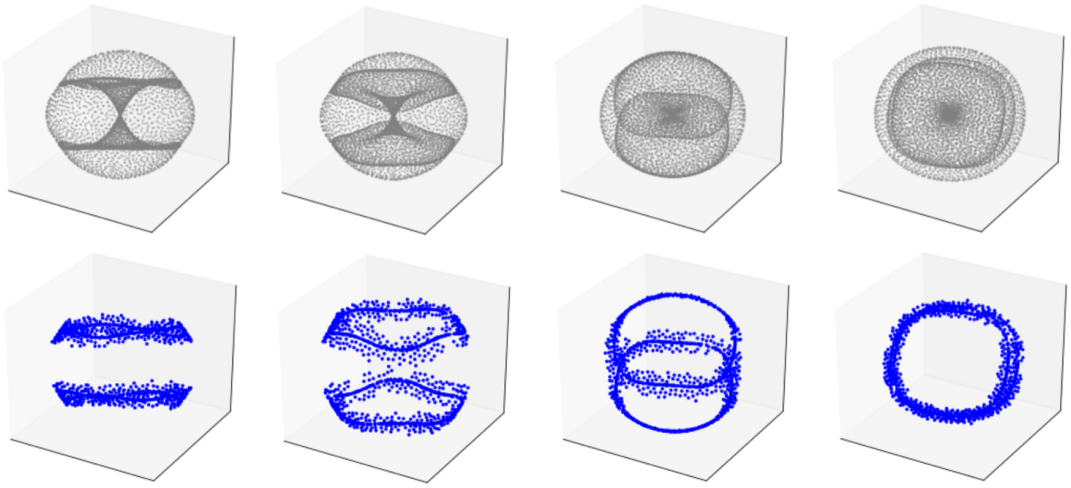

Figure 7.9: Singularities discovered by Hades marked blue in a cyclo-octane conformation dataset, which are union of two circles. Each row shows rotations of the same 3D Isomap projection of the 24-dimensional dataset. The first row shows the whole dataset and the second row shows singularities.

### 7.4.5 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 [57] 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 $28 \times 28$ images of 10 classes of clothing items[7], 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[8], 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 7.10 (MNIST) and Figure 7.11 (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.

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

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


Figure 7.10: 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.

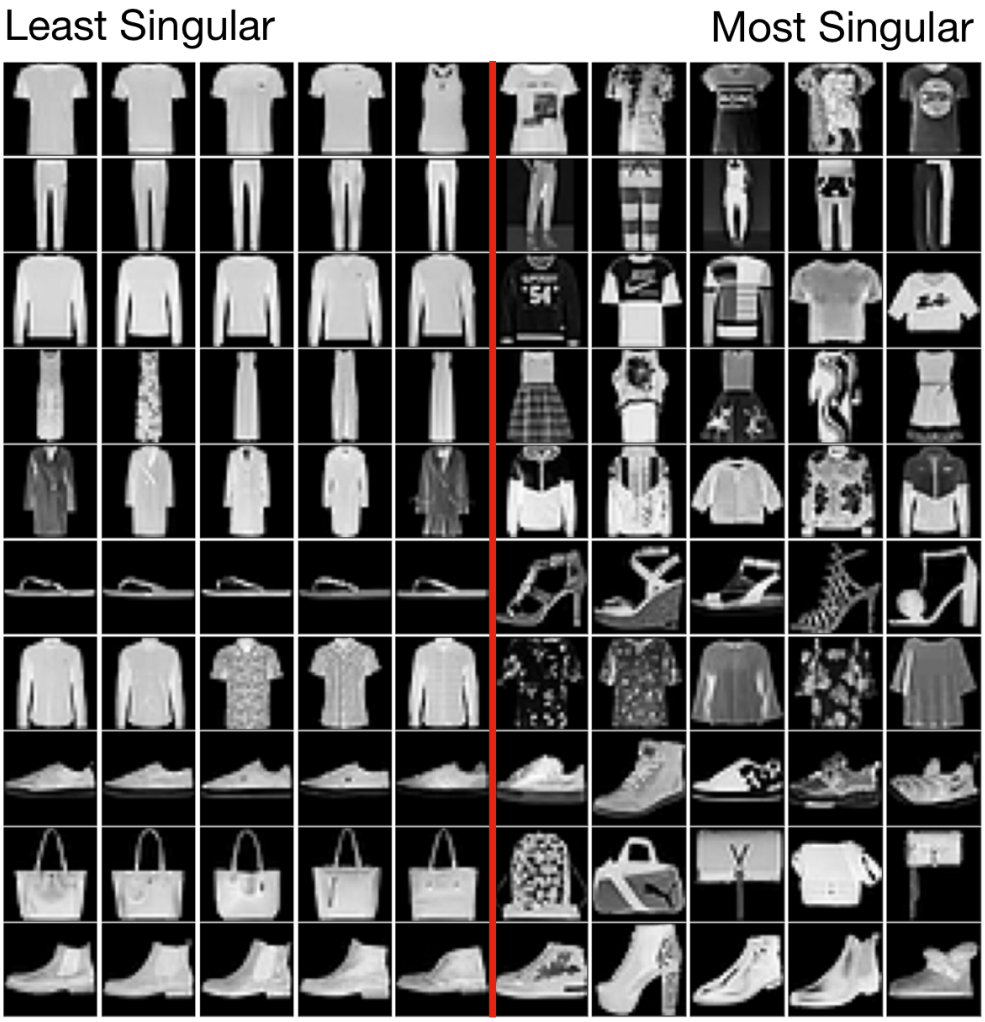


Figure 7.11: Images with the lowest singularity scores (left half) and the highest singularity scores (right half), upon applying Hades to the Fashion-MNIST dataset.

## 7.5 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 7.5.1. 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 7.5.

| Dataset | Description | Size | Time | T/S |
|---|---|---|---|---|
| Circle | One circle | 5000 | 184 | 0.037 |
| Sphere | One 2-dimensional sphere | 10000 | 1153 | 0.115 |
| Torus | One standard embedding of the 2-dimensional torus | 10000 | 932 | 0.093 |
| Two circles | Two intersecting circles | 5000 | 156 | 0.031 |
| Wheel | One circle and three line segments connecting them | 7497 | 305 | 0.041 |
| Venus | One disk and two intersecting line segments attached | 9998 | 805 | 0.081 |
| Two spheres | Two intersecting spheres | 15000 | 2012 | 0.134 |
| Cone | Two joined cones that are cut along top and bottom | 10000 | 1061 | 0.106 |
| Box | One hollow cube | 15000 | 1496 | 0.100 |
| Pinch torus | One torus, pinched along a neck | 10000 | 1326 | 0.133 |
| Skewer | One saddle surface skewered by a line segment | 15000 | 2174 | 0.145 |
| Saturn | One unit sphere and a larger disk cutting the middle | 15000 | 2172 | 0.145 |

Table 7.5.1: 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 $M_{\text{sing}}$ by declaring that $x \in M$ is $s$-close to singularity if the distance from $x$ to $M_{\text{sing}}$ 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 $\omega_d$ 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 $r_0, N_0, \alpha$. 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 $\eta$ was fixed at $\eta = 0.95$.

**Road networks.** Starting from $1500 \times 1500$ 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 [88]. Hades was run directly on the 24-dimensional dataset, with the fixed hyperparameters $(r, \eta) = (0.35, 0.95)$. This took 4 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 $\eta$ 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.

### 7.5.1 Concluding remarks

We introduced and studied Hades, an unsupervised learning algorithm that assigns a singularity score to data points. This is the experimental implementation of the main objective of this DPhil thesis, for which the theory has been developed in the chapters leading to the current one. While the theoretical content of the thesis is nontrivial, a large amount of work has also gone into making Hades readily deployable and statistically principled.

Hades detects singularities 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.

# Chapter 8

# Strange random topology of the circle

## 8.1 Introduction

A conventional wisdom in topological data analysis says the following: if we construct a simplicial complex from a random sample drawn from a manifold, then the topology of the simplicial complex approximates the topology of the manifold. Indeed this is true if we scale down the connectivity radius smaller as the sample size grows larger, but what happens when the connectivity radius stays the same?

We study the strange random topology of the circle, where we find high-dimensional topology arise in a systematic way. We find intervals of filtration radii in which the random Cech complex constructed from the circle is homotopy equivalent to bouquets of spheres, with positive probabilities. Here, a bouquet of spheres is the wedge sum $\vee^a \mathbb{S}^k$. It was known that only $a = 1$ is allowed if $k$ is odd, and all $a \geq 1$ are allowed if $k$ is even [6]. We show that the single odd sphere $\mathbb{S}^{2k+1}$ appears with high probability over long intervals of filtration radii. The bouquet of even sphere $\vee^a \mathbb{S}^{2k}$ appears with a smaller but positive probability over shrinking intervals of filtration radii. In particular, we show that $a$ can get arbitrarily large for $\vee^a \mathbb{S}^{2k}$. To get to our conclusions, we use the expected Euler characteristic and the stability theorem of persistence diagram. Let's describe the setup more precisely.

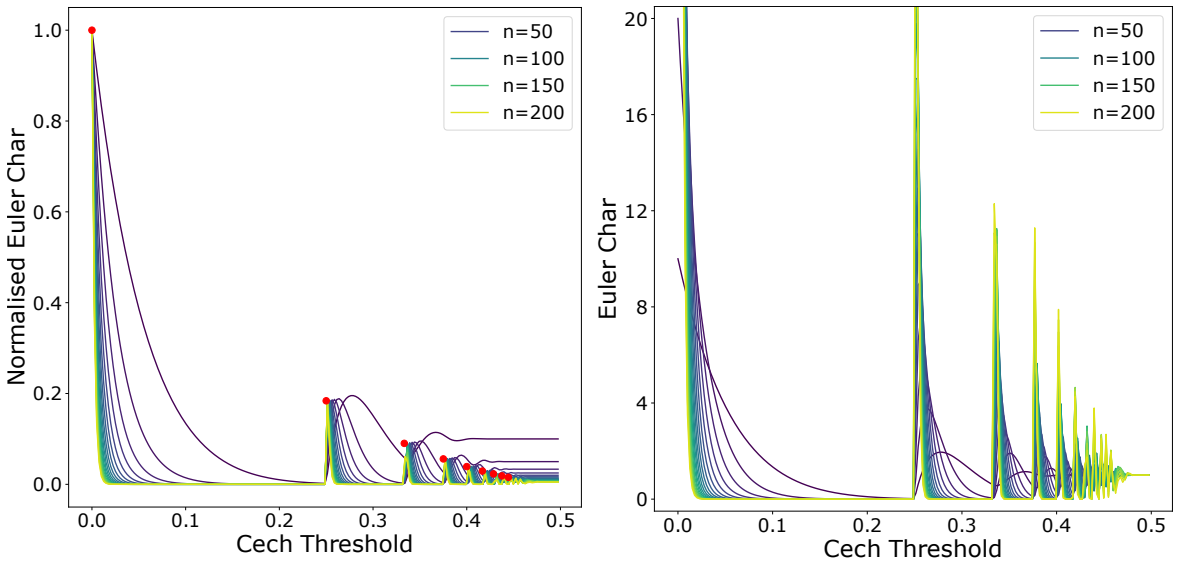


Figure 8.1: Left: Graphs of *normalised* expected Euler characteristics, $y = n^{-1} \cdot \bar{\chi}(n, x)$, for $n \in \{10, 20, \dots 200\}$ and $x \in [0, 1]$. Right: Same as left, but we plot $y = \bar{\chi}(n, x)$, which are (un-normalised) expected Euler characteristics. Yellow curves correspond to larger $n$. Red circles are the limiting spikes, given by $\left(\frac{k}{2k+2}, \frac{(k/e)^k}{(k+1)!}\right)$ for all $k \geq 0$.

**Setup.** We define the circle $\mathbb{S}^1$ as the quotient space $\mathbb{S}^1 = [0, 1]/\sim$, as the interval of length 1, glued along endpoints: $0 \sim 1$. A bouquet of spheres $\vee^a \mathbb{S}^k$ is defined as the wedge sum of $a$ copies of $\mathbb{S}^k$.[1] For a positive integer $n$, let $\mathbf{X}_n$ be the i.i.d.[2] sample of size $n$, drawn uniformly from $\mathbb{S}^1$. The Cech complex of filtration radius $\leq r$ is denoted by $\check{\mathrm{C}}(\mathbf{X}_n, r)$. In doing this construction, we always use the intrinsic topology of the circle, i.e. the Cech complex is a nerve complex constructed from arcs. We use the following notation for the expected Euler characteristic and expected Betti number, for Theorems A1 and A2:

$$\bar{\chi}(n, r) = \mathbb{E}_{\mathbf{X}_n}\left[\chi(\check{\mathrm{C}}(\mathbf{X}_n, r))\right], \quad \bar{b}_k(n, r) = \mathbb{E}_{\mathbf{X}_n}\left[\dim H_k(\check{\mathrm{C}}(\mathbf{X}_n, r))\right]$$

**Theorem A1** (Expected Euler Characteristic)**.** Let $n > 0$ and $r \in (0, 1)$. The following equality holds:

$$\bar{\chi}\left(n, \frac{1-r}{2}\right) = \sum_{k=1}^{\lfloor 1/r \rfloor} \binom{n}{k} (1 - kr)^{k-1} (kr)^{n-k}$$

[1] We take the convention that for a topological space $K$, we define the 0-th wedge sum $\vee^0 K = *$, the singleton point set, and the 1st wedge sum $\vee^1 K = K$ itself.

[2] Independently and identically distributed

In particular, $\bar{\chi}(n,r)$ is a continuous piecewise-polynomial function in $r$.

**Theorem A2** (Expected Betti Number)**.** Let $k \geq 1$. Given $\epsilon > 0$, the following uniform bounds hold for all $r \in \left[\frac{k}{2k+2}, \frac{k+1}{2k+4}\right)$ when $n$ is sufficiently large:

$$\frac{\bar{\chi}(n,r)}{n} - \epsilon \leq \frac{\bar{b}_{2k}(n,r)}{n} \leq \frac{\bar{\chi}(n,r)}{n}$$

By directly plotting the expected Euler characteristic, we observe interesting limiting behaviours. Figure 8.1 shows graphs of $f_n(r) = n^{-1} \cdot \bar{\chi}(n,r)$, which are *normalised* versions of $\bar{\chi}$. As $n$ becomes larger, $f_n(r)$ shows peaks that converge to a sequence of narrow spikes. By definition $\bar{\chi}(n,r) = n \cdot f_n(r)$, and therefore we see that as $n$ becomes larger, $\bar{\chi}(n,r)$ also becomes *arbitrarily large* at those spikes. Theorem A2 then says that the expected Betti number of a certain homological dimension is precisely responsible for each spike. We also compute later (Proposition 8.3.4) that height of the $k$-th limiting spike is precisely given by:

$$\omega_k = \frac{(k-1)^{k-1}}{k! e^{k-1}}$$

Using the Euler characteristic formula, we can now obtain probabilistic bounds on homotopy types. As stated before, all homotopy types arising from nerve complexes of circular arcs were completely classified in [6]: they are either $\mathbb{S}^{2k+1}$ for some $k \geq 0$, or $\vee^a \mathbb{S}^{2k}$ for some $a, k \geq 0$. We observe that $\chi(\mathbb{S}^{2k+1}) = 0$, whereas $\chi(\vee^a \mathbb{S}^{2k}) = a + 1$. Therefore in Figure 8.1, limiting spikes indicate contribution from the even-dimensional sphere bouquets $\vee^a \mathbb{S}^{2k}$ with large $a$, and the plateaus indicate contribution from the odd-dimensional spheres. Recalling that $\mathbf{X}_n$ is an i.i.d. sample of size $n$ drawn from $\mathbb{S}^1$, we introduce Theorems B and C:

**Theorem B** (Odd Spheres)**.** Let $k \geq 0$ be an integer, and also let $\epsilon, \delta > 0$. Suppose that $|r - \nu_k| \leq \tau_k - \epsilon$. Then for sufficiently large $n$, the following homotopy equivalence holds with probability at least $1 - \delta$:

$$\check{\mathrm{C}}(\mathbf{X}_n, r) \simeq \mathbb{S}^{2k+1}$$

where

$$\nu_k = \frac{2k^2 + 4k + 1}{4(k+1)(k+2)}, \quad \tau_k = \frac{1}{4(k+1)(k+2)}$$

We note that $\nu_k = \frac{1}{2}(\frac{k+1}{2k+4} + \frac{k}{2k+2})$ and $\tau_k = \frac{1}{2}(\frac{k+1}{2k+4} - \frac{k}{2k+2})$, so that Theorem B covers most of each interval $r \in [\frac{k}{2k+2}, \frac{k+1}{2k+4}]$.

**Theorem C** (Even Spheres)**.** Let $k \geq 2$, $\eta \in (0,1)$. Suppose that $|r - \rho_{k,n}| \leq \sigma_{k,\eta}/n$. Then for sufficiently large $n$, the following homotopy equivalence holds with probability at least $\eta \cdot k\omega_k$:

$$\check{\mathrm{C}}(\mathbf{X}_n, r) \simeq \vee^a \mathbb{S}^{2k-2}, \quad \text{for some } \frac{(1-\eta)\omega_k \cdot n}{2} \leq a+1 \leq \frac{n}{k}$$

where

$$\rho_{k,n} = \frac{n(k+1)}{2k(n-1)}, \quad \sigma_{k,\eta} = \frac{(1-\eta)^3(k\omega_k)^3}{320\sqrt{k+2}}, \quad \omega_k = \frac{(k-1)^{k-1}}{k!e^{k-1}}$$

To see Theorem C in action, one may simply set $\eta = 1/2$ to obtain results.

**Structure of the chapter.**[3] In Section 2 we prove Theorem A1, i.e. compute the expected Euler characteristic precisely. In Section 3 we analyse the limiting spikes of the expected Euler characteristics. In Section 4 we use the classification of homotopy types arising from a nerve complex of circular arcs, to give constraints on homotopy types and compute probabilistic bounds. Theorem A2 and Theorem C are proven in Section 4. In Section 5 we use the classical method of stability of persistence diagram to prove Theorem B; this section works separately and doesn't use the Euler characteristic method.

Theorem C takes the most work to prove. It is a simplified version of Theorem 8.4.8, which has a few more parameters that can be tweaked to obtain similar variants of Theorem C. Theorem 8.4.8 is obtained by combining three ingredients: Propositions 8.3.4, 8.4.5, and 8.4.7.

**Related works.**

The classical result of Hausmann shows that the Vietoris-Rips complex constructed from the manifold with a small scale parameter recovers the homotopy type of the manifold [47]. Another classical result of Niyogi, Smale, Weinberger shows that if a Cech complex of small filtration radius is constructed from a finite random sample of a Euclidean submanifold, then the homotopy type of the manifold is recovered with high confidence [72].

[3] We remark that the theorems A1, A2, B, C aren't proven in sequential order; they are arranged in that order for a clean exposition of the main results.

Much work has been done for recovering topology of a manifold from its finite sample, when connectivity radius is scaled down with the sample size at a specific rate [23] [37] [51] [24]. A central theme of this body of work is the existence of phase transitions when parameters controlling the scaling of connectivity radius are changed. For a comprehensive survey, see [75] and [22].

In comparison, the setting when connectivity radius is not scaled down with sample size is studied much less. Results on convergence of the topological quantities have been studied [74] [90], but not much attention has been devoted to analysing specific manifolds.

This chapter builds on two important works that characterised the Vietoris-Rips and Cech complexes of subsets of the circle: [6] and [3]. Several variants of these ideas were studied, for ellipse [8], regular polygon [9], and hypercube graph [4]. Randomness in these systems were studied using dynamical systems in [7]. One key tool to further study the topology of Vietoris-Rips and Cech complexes arising from a manifold is metric thickening [5]. Using this tool, the Vietoris-Rips complex of the higher-dimensional sphere has been characterised up to small filtration radii [62].

## 8.2 Expected Euler characteristic

In this section we compute the expected Euler characteristic precisely. We start with a simple calculation that also briefly considers the Vietoris-Rips complex, but soon after we only work with the Cech complex. Let $\mathrm{VR}(\mathbf{X}_n, r)$ denote the Vietoris-Rips complex of threshold $r$. The following proposition reduces computation of expected values to the quantities $T_k$ and $Q_k$, defined below:

**Proposition 8.2.1.** *For each $n > 0$, let $\mathbf{X}_n$ be the iid sample drawn uniformly from $\mathbb{S}^1$. Then we have that:*

$$\mathbb{E}[\chi(\mathrm{VR}(\mathbf{X}_n, r))] = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} T_k(r)$$

$$\mathbb{E}[\chi(\check{C}(\mathbf{X}_n, r))] = 1 + \sum_{k=1}^{n} (-1)^{k} \binom{n}{k} Q_k(1-2r)$$

*where $T_k(r)$ is the probability that every pair of points in $\mathbf{X}_k$ are within distance $r$, and $Q_k(r)$ is the probability that open arcs of length $r$ centered at points of $\mathbf{X}_k$ cover $\mathbb{S}^1$.*

*Expectation is taken over the iid sample* $\mathbf{X}_n$.

*Proof.* Denoting by $s_k(K)$ the number of $k$-simplices in a simplicial complex $K$, we have that:

$$\mathbb{E}[s_k(\mathrm{VR}(\mathbf{X}_n, r))] = \binom{n}{k} T_k(r)$$

and thus

$$\mathbb{E}[\chi(\mathrm{VR}(\mathbf{X}_n, r))] = \sum_{k=0}^{n-1} (-1)^k \mathbb{E}[s_k(\mathrm{VR}(\mathbf{X}_n, r))] = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} T_k(r)$$

The relation for the Cech complex is derived in the same way, except we note the following: the probability that arcs of radius $r$ centered at points of $\mathbf{X}_k$ intersects nontrivially is equal to $1-Q_k(1-2r)$. This is by De Morgan's Law: for any collection of sets $\{U_j \subseteq \mathbb{S}^1\}_{j\in J}$, we have $\cap_{j\in J} U_j = \emptyset$ iff $\cup_{j\in J} U_j^{\mathrm{c}} = \mathbb{S}^1$. In the case of circle (of circumference 1), complement of a closed arc of radius $r$ is an open arc of length $1-2r$. Applying this logic, we obtain:

$$\mathbb{E}[\chi(\check{\mathrm{C}}(\mathbf{X}_n, r))] = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} (1 - Q_k(1-2r))$$

which is easily seen to be the same as the asserted expression (note that $\sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} = 1$.) □

The $Q_k$ were computed by Stevens in 1939 [87]. We reproduce the proof for completeness.

**Theorem 8.2.2** (Stevens)**.** *If $k$ arcs of fixed length $a$ are independently, identically and uniformly sampled from the circle of circumference 1, then the probability that these arcs cover the circle is equal to the following:*

$$Q_k(a) = \sum_{l=0}^{\lfloor 1/a \rfloor} (-1)^l \binom{k}{l} (1-la)^{k-1}$$

*Proof.* The proof is an application of inclusion-exclusion principle. Consider the set $E = \{(x_1, \dots x_k) | 0 \le x_1 < \cdots < x_k < 1\}$. For each collection of indices $J \subseteq \{1, \dots k\}$, define $\bar{E}_J$ and $E_J$ as the following subsets of $E$:

$$\begin{aligned} E_J &= \{(x_1, \dots x_k) \in E | j \in J \iff x_{j+1} - x_j > a\} \\ \bar{E}_J &= \{(x_1, \dots x_k) \in E | j \in J \implies x_{j+1} - x_j > a\} = \bigsqcup_{J' \supseteq J} E_{J'} \end{aligned}$$

By definition, we have $\mathrm{Vol}(E_\emptyset) = Q_k(a)$. To compute it, we apply the inclusion-exclusion principle for the membership of each $E_J$ over $\bar{E}_{J'}$ whenever $J' \supseteq J$. Noting the relation $\sum_{l=1}^{k}(-1)^{l+1}\binom{k}{l} = 1$, we see that:

$$1 = \sum_{J\subseteq\{1,\dots k\}} \mathrm{Vol}(E_J) = \mathrm{Vol}(E_\emptyset) - \sum_{\emptyset\neq J\subseteq\{1,\dots k\}} (-1)^{\#J}\,\mathrm{Vol}(\bar{E}_J)$$

Finally, if $l = \#J$ and $l \leq \lfloor 1/a \rfloor$, then $\mathrm{Vol}(\bar{E}_J) = (1-la)^{n-1}$. This is because demanding gap conditions $x_{i+1} - x_i > a$ at $l$ places is equivalent to sampling $n-1$ points from an interval of length $1 - la$[4]. Meanwhile if $l > \lfloor 1/a \rfloor$, then we always have $\mathrm{Vol}(\bar{E}_J) = 0$. Plugging these into the above equation, we get:

$$\mathrm{Vol}(E_\emptyset) = 1 + \sum_{l=1}^{\lfloor 1/a \rfloor} (-1)^l \binom{k}{l} (1-la)^{n-1}$$

as desired. □

We then get the following:

**Theorem 8.2.3** (Theorem A1)**.** *Expected Euler characteristic of random Cech complex on a circle of unit circumference obtained from $n$ points and filtration radius $(1-r)/2$ is:*

$$\bar{\chi}\left(n, \frac{1-r}{2}\right) = \sum_{k=1}^{\lfloor 1/r \rfloor} \binom{n}{k} (1-kr)^{k-1}(kr)^{n-k}$$

*In particular, $\bar{\chi}(n,r)$ is a continuous piecewise-polynomial function in $r$.*

*Proof.* Substituting the $Q_k$ expression in, we get:

$$\begin{aligned}
\bar{\chi}\left(n, \frac{1-r}{2}\right) =& 1 + \sum_{k=1}^{n} (-1)^k \binom{n}{k} Q_k(r) \\
=& 1 + \sum_{l=0}^{\lfloor 1/r \rfloor} \sum_{k=1}^{n} (-1)^{k+l} \binom{n}{k}\binom{k}{l} (1-rl)^{k-1} \\
=& \sum_{l=1}^{\lfloor 1/r \rfloor} \sum_{k=1}^{n} (-1)^{k+l} \binom{n}{k}\binom{k}{l} (1-rl)^{k-1}
\end{aligned}$$

[4]This can be seen more precisely by considering the collection $E'$ of $(y_1, \dots y_{k-1})$ defined by $y_i = x_{i+1} - x_i > 0$ and $\sum y_i \leq 1$, and then considering the subset $E'_J$ defined by $y_i > a$ for $i \in J$. The quantity of interest is $\mathrm{Vol}(E'_J)/\mathrm{Vol}(E')$. Furthermore, the map $(y_1, \dots y_{k-1}) \mapsto (y_1 - \mathbf{1}_{1\in J}, \dots y_{k-1} - \mathbf{1}_{k-1\in J})$ isometrically maps $E'_J$ to $(1-la)\cdot E'$, so that $\mathrm{Vol}(E'_J) = (1-la)^{k-1}\,\mathrm{Vol}(E')$ due to the $(k-1)$-dimensional volume scaling. This is exactly the original claim.

where we switched the order of summation in the second equality, and isolating the $l = 0$ part cancels out the 1 in the third equality. Noting that $\binom{n}{k}\binom{k}{l} = \binom{n}{l}\binom{n-l}{k-l}$, we further get:

$$\begin{aligned}
\bar{\chi}\left(n, \frac{1-r}{2}\right) &= \sum_{l=1}^{\lfloor 1/r \rfloor} (-1)^l \binom{n}{l} (1-rl)^{-1} \sum_{k=l}^{n} \binom{n-l}{k-l} (rl-1)^k \\
&= \sum_{l=1}^{\lfloor 1/r \rfloor} \binom{n}{l} (1-rl)^{l-1} \sum_{k=0}^{n-l} \binom{n-l}{k} (rl-1)^k \\
&= \sum_{l=1}^{\lfloor 1/r \rfloor} \binom{n}{l} (1-rl)^{l-1} (rl)^{n-l}
\end{aligned}$$

□

## 8.3 Limit behaviour of Euler characteristic

We prove a sequence of lemmas in this section to characterise the limiting spikes in Figure 8.1. The main idea is that only one summand in the expected Euler characteristic contributes mainly to the spike, and this is a polynomial term that can be studied with calculus. The main results of this section are Propositions 8.3.3 and 8.3.4. The two lemmas leading up to it are exercises in calculus that explain the specific situation of our expected Euler characteristic.

**Lemma 8.3.1.** *For $a, b \geq 1$, the function $f(t) = t^a(1-t)^b$ satisfies the following:*

*(a) In the range $0 \leq t \leq 1$, $f(t)$ achieves the unique maximum value at $t = a/(a+b)$:*

$$\max_{0 \leq t \leq 1} f(t) = f\left(\frac{a}{a+b}\right) = \frac{a^a b^b}{(a+b)^{a+b}}$$

*Also, $f(t)$ is increasing on $t \in (0, a/(a+b))$ and decreasing on $t \in (a/(a+b), 1)$.*

*(b) The following linear lower bounds hold:*

$$\begin{aligned}
f(t) \geq & u\Big((a+b)vt - av + 1\Big), \text{ when } 0 < t < \frac{a}{a+b} \\
f(t) \geq & u\Big(-(a+b)vt + av + 1\Big), \text{ when } \frac{a}{a+b} < t < 1
\end{aligned}$$

*where*

$$u = \frac{a^a b^b}{(a+b)^{a+b}}, \quad v = \sqrt{\frac{a+b}{ab}}$$

*(c) For each $\lambda \in [0,1]$, we have that:*

$$\left| t - \frac{a}{a+b} \right| < \frac{(1-\lambda)\sqrt{ab}}{(a+b)^{3/2}} \implies t^a(1-t)^b > \lambda u$$

*Proof.* The first two derivatives are:

$$\begin{aligned} f'(t) =& \Big( a - (a+b)t \Big) t^{a-1}(1-t)^{b-1} \\ f''(t) =& \Big( (a+b)(a+b-1)t^2 + 2a(1-a-b)t + a(a-1) \Big) t^{a-2}(1-t)^{b-2} \end{aligned}$$

The first derivative vanishes at $t \in \{a/(a+b), 0, 1\}$ and the second derivative vanishes at $t \in \{t_0 \pm \eta_0, 0, 1\}$ where

$$t_0 = \frac{a}{a+b}, \quad \eta_0 = \frac{1}{a+b}\sqrt{\frac{ab}{a+b-1}} > \frac{\sqrt{ab}}{(a+b)^{3/2}} = \eta_1$$

The first derivative is positive at $(0, a/(a+b))$ and negative at $(a/(a+b), 1)$. Thus the maximum at $t \in [0,1]$ is given by:

$$f(t_0) = \frac{a^a b^b}{(a+b)^{a+b}}$$

Thus

$$\begin{aligned} f(t) &\geq \frac{f(t_0)}{\eta_1}(t - t_0) + f(t_0), \text{ when } 0 < t < t_0 \\ f(t) &\geq \frac{-f(t_0)}{\eta_1}(t - t_0) + f(t_0), \text{ when } t_0 < t < 1 \end{aligned}$$

and

$$\pm \frac{f(t_0)}{\eta_1}(t - t_0) + f(t_0) = \frac{a^a b^b}{(a+b)^{a+b}} \left( \pm \frac{(a+b)^{3/2}}{\sqrt{ab}} \left( t - \frac{a}{a+b} \right) + 1 \right)$$

(c) follows from the linear bound of (b). □

**Lemma 8.3.2.** *Let $m, n \geq 1$ be integers and define:*

$$f_{m,n}(t) = \binom{n}{m} (mt)^{m-1}(1-mt)^{n-m}$$

*Then $f_{m,n}$ satisfies the following:*

*(a) $f_{m,n}(t)$ is increasing when $0 < t < t_0$ and decreasing when $t_0 < t < 1/m$ where $t_0 = \frac{1}{n-1}(1 - \frac{1}{m})$.*

*(b) The maximum over $0 < t < 1/m$ is given by:*

$$\max_{0<mt<1} f_{m,n}(t) = f_{m,n}(t_0) = \binom{n}{m} \frac{(m-1)^{m-1}(n-m)^{n-m}}{(n-1)^{n-1}}$$

*(c) For each $\lambda \in [0,1]$, we have that:*

$$|t - t_0| < \frac{(1-\lambda)\sqrt{(m-1)(n-m)}}{m(n-1)^{3/2}} \implies f_{m,n}(t) > \lambda f_{m,n}(t_0)$$

*(d) The normalised limit of maximum as $n \to \infty$ is given by:*

$$\lim_{n\to\infty} \frac{\max_{0<t<1/m} f_{m,n}(t)}{n} = \frac{(m-1)^{m-1}}{m! e^{m-1}}$$

*Proof.* (a)-(c) follow from the previous lemma. For (d), we compute:

$$\begin{aligned}\lim_{n\to\infty} \frac{\max_{0<t<1/m} f_{m,n}(t)}{n} =& \frac{(m-1)^{m-1}}{m!} \lim_{n\to\infty} (n-1)(n-2)\cdots(n-m+1)\frac{(n-m)^{n-m}}{(n-1)^{n-1}} \\ =& \frac{(m-1)^{m-1}}{m!} \lim_{n\to\infty} \frac{(n-m)^{n-m}}{(n-1)^{n-m}}\end{aligned}$$

and also

$$\lim_{n\to\infty} \frac{(n-m)^{n-m}}{(n-1)^{n-m}} = \lim_{n\to\infty} \left(1 - \frac{m-1}{n-1}\right)^{n-m} = \lim_{n\to\infty} \left(1 - \frac{m-1}{n-1}\right)^{n-1} = \frac{1}{e^{m-1}}$$

which gives the desired expression. □

**Proposition 8.3.3.** *Suppose that $m, n$ are integers with $2 \leq m < \sqrt{n}$. The following holds for $\bar{\chi}(n, r)$.*

*(a) The following bounds hold:*

$$a_{m,n} \leq \frac{\bar{\chi}(n, s_{m,n})}{n} \leq M \leq a_{m,n} + b_{m,n}$$

*where*

$$\begin{aligned}M =& \max\left\{ \frac{1}{n}\bar{\chi}\left(n, \frac{1-r}{2}\right) \;\middle|\; r \in \left(\frac{1}{m+1}, \frac{1}{m}\right)\right\} \\ s_{m,n} =& \frac{(m-1)n}{2(n-1)m} \\ a_{m,n} =& \binom{n}{m} \frac{(m-1)^{m-1}(n-m)^{n-m}}{n(n-1)^{n-1}} \\ b_{m,n} =& en^{m-1}\left(1 - \frac{1}{m+1}\right)^{n-1}\end{aligned}$$

*(b) We have the following limits:*

$$\lim_{n\to\infty} a_{m,n} = \frac{(m-1)^{m-1}}{m!e^{m-1}}, \quad \lim_{n\to\infty} b_{m,n} = 0$$

*(c) Suppose additionally that $n > 2m^2$. Then for each $\lambda \in [0,1]$, we have that:*

$$\left| r - \frac{n-m}{(n-1)m} \right| < \frac{(1-\lambda)\sqrt{(m-1)(n-m)}}{m(n-1)^{3/2}} \implies \frac{1}{n}\bar{\chi}\left(n, \frac{1-r}{2}\right) > \lambda a_{m,n}$$

*This condition for $r$ in particular satisfies $r \in \left(\frac{1}{m+1}, \frac{1}{m}\right]$.*

*Proof.* Let $r \in \left(\frac{1}{m+1}, \frac{1}{m}\right]$ and also write $r = \frac{1}{m} - t$, with $t \in \left[0, \frac{1}{m(m+1)}\right]$. Then we may rewrite the normalised expected Euler characteristic as follows:

$$\begin{aligned}\bar{\chi}\left(n, \frac{1-r}{2}\right) &= \sum_{k=1}^{m} \binom{n}{k} (1-kr)^{k-1}(kr)^{n-k} \\ &= \sum_{k=1}^{m} \binom{n}{k} \left(1 - \frac{k}{m} + kt\right)^{k-1} \left(\frac{k}{m} - kt\right)^{n-k}\end{aligned}$$

We now claim that the $k = m$ term is the dominant one among the above summands. As such, we split the above sum as:

$$\bar{\chi}\left(n, \frac{1-r}{2}\right) = f_{m,n}(t) + E$$

where

$$\begin{aligned}f_{m,n}(t) &= \binom{n}{m} (mt)^{m-1}(1-mt)^{n-m}, \\ E &= \sum_{k=1}^{m-1} \binom{n}{k} \left(1 - \frac{k}{m} + kt\right)^{k-1} \left(\frac{k}{m} - kt\right)^{n-k}\end{aligned}$$

Since $m < \sqrt{n}$, we have $s_{m,n} = \frac{1}{n-1}(1 - \frac{1}{m}) < \frac{1}{m(m+1)}$. Therefore, the previous Lemma tells us that $f_{m,n}(t)$ achieves (global) maximum at $\tilde{s} \in \left(0, \frac{1}{m(m+1)}\right]$, with the maximum value given by:

$$f_{m,n}(\tilde{s}) = n \cdot a_{m,n}, \text{ where } a_{m,n} = \binom{n}{m} \frac{(m-1)^{m-1}(n-m)^{n-m}}{n(n-1)^{n-1}}$$

We also bound $E$ as follows, using the inequality $\frac{m}{m+1} < 1 - mt \leq 1$:

$$\begin{aligned}
E =& \sum_{k=1}^{m-1} \binom{n}{k} \left(1 - \frac{k}{m}(1-mt)\right)^{k-1} \left(\frac{k}{m}(1-mt)\right)^{n-k} \\
\leq& \sum_{k=1}^{m-1} \binom{n}{k} \left(1 - \frac{1}{m+1}\right)^{k-1} \left(1 - \frac{1}{m}\right)^{n-k} \\
\leq& \sum_{k=1}^{m-1} \frac{n^k}{k!} \left(1 - \frac{1}{m+1}\right)^{n-1} \\
\leq& e n^{m-1} \left(1 - \frac{1}{m+1}\right)^{n-1}
\end{aligned}$$

This shows (a). Now (b) follows from the previous Lemma and the fact that $(1 - \frac{1}{m+1})^n$ term causes exponential decay for $b_{m,n}$.

(c) follows from (c) of the previous Lemma. We additionally impose the condition $n > 2m^2$, so that the endpoints of $t$ satisfying the condition fall in the interval $t \in \left[0, \frac{1}{m(m+1)}\right)$. □

**Proposition 8.3.4.** *Let $m \geq 2$, $\epsilon > 0$. The following holds for sufficiently large $n$:*

$$r \in \left[\alpha^-, \alpha^+\right] \implies \frac{1}{n}\bar{\chi}\left(n, \frac{1-r}{2}\right) \in \left[(1-\epsilon)\omega_m, (1+\epsilon)\omega_m\right]$$

*where*

$$\alpha^{\pm} = \frac{n-m}{(n-1)m}\left(1 \pm \frac{\epsilon\sqrt{m-1}}{n}\right), \quad \omega_m = \frac{(m-1)^{m-1}}{m!e^{m-1}}$$

*Proof.* This follows directly from the previous Proposition. $\alpha^{\pm}$ are slight relaxations of the interval in (c), where we set $\lambda = 1 - \epsilon$:

$$\left[\frac{n-m}{(n-1)m} - \epsilon R_1, \frac{n-m}{(n-1)m} + \epsilon R_1\right] \supseteq \left[\frac{n-m}{(n-1)m}(1-\epsilon R_2), \frac{n-m}{(n-1)m}(1+\epsilon R_2)\right]$$

$$\text{where } R_1 = \frac{\epsilon\sqrt{(m-1)(n-m)}}{m(n-1)^{3/2}}, \quad R_2 = \frac{\sqrt{m-1}}{n}$$

□

## 8.4 Random homotopy types

### 8.4.1 Constraints on homotopy types

Let $U_n = \{i/n \,|\, i = 0, 1, \ldots n-1\} \subset \mathbb{S}^1$ be the set of $n$ equally spaced points. Let $\mathcal{N}(n,k)$ be the nerve complex on $U_n$ defined by the open cover consisting of closed intervals

$[i/n, (i+k)/n]$.

**Lemma 8.4.1.** *We have that:*

$$\check{C}(U_n, r) = \check{C}\left(U_n, \frac{\lfloor 2rn \rfloor}{2n}\right) = \mathcal{N}(n, \lfloor 2rn \rfloor)$$

The following result is from [6]:

**Proposition 8.4.2.**

$$\mathcal{N}(n,k) \simeq \begin{cases} \vee^{n-k-1}\mathbb{S}^{2l} & \text{if } \frac{k}{n} = \frac{l}{l+1} \\ \mathbb{S}^{2l+1} & \text{if } \frac{k}{n} \in \left(\frac{l}{l+1}, \frac{l+1}{l+2}\right) \end{cases}$$

*Note that if* $(k,n) = (jl, j(l+1))$*, then* $n-k-1 = j-1$*, so that* $\vee^{n-k-1}\mathbb{S}^{2l} = \vee^{j-1}\mathbb{S}^{2l}$*.*

Using the above, we easily show that:

**Proposition 8.4.3.** *Given* $r \in (0, 1/2)$*, the following two subsets of* $\mathbb{Z}^3$ *are equal:*

$$\left\{(n,a,b) \;\middle|\; \check{C}(U_n, r) \simeq \vee^a \mathbb{S}^{2b}\right\} = \left\{((a+1)(b+1), a, b) \;\middle|\; b+1 \leq \tilde{r}^{-1},\ a+1 \leq \frac{1}{1-(b+1)\tilde{r}}\right\}$$

*where* $\tilde{r} = 1-2r$*. In particular, if* $\tilde{r}^{-1} \in [k, k+1)$*, then* $b \in \{0, 1, 2, \ldots k-1\}$ *and we have* $a \leq k-1$ *when* $b \leq k-2$*.*

*Proof.* To have $\mathcal{N}(n, \lfloor 2rn \rfloor) = \check{\mathrm{C}}(U_n, r) \simeq \vee^a \mathbb{S}^{2b}$, we see from the previous Proposition that the condition is given by $(\lfloor 2rn \rfloor, n) = ((a+1)b, (a+1)(b+1))$. This determines $n$ from $(a,b)$. The condition on $\lfloor 2rn \rfloor$ is then:

$$\begin{aligned} & (a+1)b \leq 2r(a+1)(b+1) < (a+1)b+1 \\ \iff & \tilde{r}(b+1) \leq 1,\ a < \tilde{r}(a+1)(b+1) \\ \iff & (b+1) \leq \tilde{r}^{-1},\ (a+1) < (1-\tilde{r}(b+1))^{-1} \end{aligned}$$

as desired. □

**Remark.** At fixed $l$, let $k = b + k_0$. Then $\frac{1}{1-(b+1)/k} = 1 + \frac{b+1}{k_0 - 1}$ and changing $k_0$ by a single value can have a heavy effect on the upper bound.

**Proposition 8.4.4.** *Let $r \in (0, 1/2)$ and $n$ be given; define $\tilde{r} = 1 - 2r$ and let $k = \lfloor \tilde{r}^{-1} \rfloor$. We have the following relations between subsets of $\mathbb{Z}^2$:*

$$\begin{aligned}
&\left\{(a,b) \;\middle|\; \check{C}(\mathbf{Y}, r) \simeq \vee^a \mathbb{S}^{2b},\ \mathbf{Y} \subset \mathbb{S}^1,\ \#\mathbf{Y} = n\right\} \\
=&\left\{(a,b) \;\middle|\; \check{C}(U_m, r) \simeq \vee^a \mathbb{S}^{2b},\ m \leq n\right\} \\
\subseteq&\left\{(a,b) \;\middle|\; b+1 \leq k,\ a+1 \leq \min\left(\frac{n}{b+1}, \frac{1}{1-(b+1)\tilde{r}}\right)\right\} \\
\subseteq&\left\{(a,b) \;\middle|\; b+1 \leq k-1,\ a+1 \leq \frac{k}{k-b-1}\right\} \cup \left\{(a,k-1) \;\middle|\; a+1 \leq \frac{n}{k}\right\}
\end{aligned}$$

*where in the final expression, $k/0 = \infty$ by convention.*

*Proof.* The first equality holds because for every $\mathbf{Y} \subset \mathbb{S}^1$, there exists $\mathbf{Y}' \subset \mathbf{Y}$ such that $\check{\mathrm{C}}(\mathbf{Y}, r) \simeq \check{\mathrm{C}}(\mathbf{Y}', r) \simeq \check{\mathrm{C}}(U_m, r)$, where $m = \#\mathbf{Y}'$ [6]. The first inclusion follows from the previous Proposition. The second inclusion follows from separating the two cases $b+1 < k$ and $b + 1 = k$. □

### 8.4.2 Probabilistic bounds

For a topological space $K$, we define the following notation for probability:

$$p(K, n, r) = \mathbb{P}[\check{\mathrm{C}}(\mathbf{X}_n, r) \simeq K]$$

We generally have the following:

$$\bar{\chi}(n, r) = \mathbb{E}[\chi(\check{\mathrm{C}}(\mathbf{X}_n, r))] = \sum_K \chi(K) \cdot p(K, n, r)$$

where the sum is well-defined because there are only finitely many combinatorial structures that $\check{\mathrm{C}}(\mathbf{X}_n, r)$ can take. Furthermore if we let $k = \lfloor (1-2r)^{-1} \rfloor$, then Proposition 8.4.4 tells us that:

$$\{K \,|\, p(K, n, r) > 0\} \subseteq \left\{ \vee^a \mathbb{S}^{2b} \middle| b+1 \leq k-1, a+1 \leq \frac{k}{k-b-1} \right\} \cup \left\{ \vee^a \mathbb{S}^{2k-2} \middle| a+1 \leq \frac{n}{k} \right\}$$

From this we infer that[5]:

$$\begin{aligned}\bar{\chi}(n,r) =& A_{<k} + A_k \\ \text{where } A_{<k} =& \sum_{\substack{0\le b\le k-2 \\ (a+1)(k-b-1)\le k}} (a+1)\cdot p(\vee^a \mathbb{S}^{2b}, n, r) \\ A_k =& \sum_{1<a+1\le n/k} (a+1)\cdot p(\vee^a \mathbb{S}^{2k-2}, n, r)\end{aligned}$$

where we used $\chi(\vee^a \mathbb{S}^{2b}) = a+1$. Since sum of probabilities is 1, applying the constraint $(a+1)(k-b-1) \le k$ implies that $A_{<k} \le k$. This implies the following:

**Proposition 8.4.5.** *The following holds:*

$$A_k \le \bar{\chi}(n,r) \le k + A_k$$

*where*

$$A_k = \sum_{1<a+1\le n/k} (a+1)\cdot p(\vee^a \mathbb{S}^{2k-2}, n, r)$$

**Corollary 8.4.6** (Theorem A2)**.** *Let $k \ge 2$. Given $\epsilon > 0$, the following hold for sufficiently large $n$:*

$$1-2r \in \left(\frac{1}{k+1}, \frac{1}{k}\right] \implies \frac{\bar{\chi}(n,r)}{n} - \epsilon \le \frac{\bar{b}_{2k-2}(n,r)}{n} \le \frac{\bar{\chi}(n,r)}{n}$$

Now we're interested in controlling probabilities that $\vee^a \mathbb{S}^{2k-2}$ appear, with large $n$. For this, we further define following:

$$\begin{aligned}p_a =& p(\vee^a \mathbb{S}^{2k-2}, n, r) \\ l =& \lfloor n/k \rfloor - 1 \\ \tilde{\delta} =& \lceil \delta n/k \rceil - 1 \\ A_{k,\delta} :=& \sum_{\tilde{\delta}\le a\le l} (a+1)\cdot p_a = \sum_{\delta n/k \le a+1 \le n/k} (a+1)\cdot p_a \\ B_{k,\delta} :=& \sum_{\tilde{\delta}\le a\le l} p_a\end{aligned}$$

To produce bounds for $B_{k,\delta}$, we split $A_k$ into two parts:

$$A_k = \left(2p_1 + 3p_2 + \cdots + \tilde{\delta} p_{\tilde{\delta}-1}\right) + \left((\tilde{\delta}+1)p_{\tilde{\delta}} + \cdots + (l+1)p_l\right)$$

[5]By convention, in the summation we only consider $a \le 0$ when $b = 0$ and instead consider $a > 0$ when $b > 0$. This is so that the singleton set $\vee^a \mathbb{S}^{2b} = *$ is counted only once.

from which it directly follows that:

$$(\tilde{\delta}+1)B_{k,\delta} \le A_k \le \tilde{\delta}(1-B_{k,\delta}) + (l+1)B_{k,\delta}$$

and therefore

$$\begin{aligned}
&\implies (\tilde{\delta}+1)B_{k,\delta} \le A_k \le \tilde{\delta} + (l+1-\tilde{\delta})B_{k,\delta} \\
&\implies \frac{A_k - \tilde{\delta}}{l+1-\tilde{\delta}} \le B_{k,\delta} \le \frac{A_k}{\tilde{\delta}+1} \\
&\implies \frac{A_k - \lceil \delta n/k \rceil + 1}{\lfloor n/k \rfloor - \lceil \delta n/k \rceil + 1} \le B_{k,\delta} \le \frac{A_k}{\lceil \delta n/k \rceil} \\
&\implies \frac{A_k - \delta n/k}{(1-\delta)(n/k)+1} \le B_{k,\delta} \le \frac{A_k}{\delta n/k}
\end{aligned}$$

In summary, we have the following:

**Proposition 8.4.7.** *Let $n \in \mathbb{Z}^+$, $\delta \in (0,1)$, $r \in (0,1/2)$ be given, and let $k = \lfloor (1-2r)^{-1} \rfloor$. The following holds:*

$$\frac{kA_k - \delta n}{(1-\delta)n + k} \le B_{k,\delta} \le \frac{kA_k}{\delta n}$$

*where*

$$A_k = \sum_{1<a+1\le n/k} (a+1)p_a, \quad B_{k,\delta} := \sum_{\delta n/k \le a+1 \le n/k} p_a, \quad p_a := p(\vee^a \mathbb{S}^{2k-2}, n, r)$$

Now Propositions 8.3.4, 8.4.5, 8.4.7 imply the following, which is a more general version of Theorem C:

**Theorem 8.4.8.** *Let $r \in [\frac{1}{4}, \frac{1}{2})$ and let $k = \lfloor (1-2r)^{-1} \rfloor$. Given $\epsilon, \delta \in (0,1)$, the following implication holds for large enough $n$:*

$$1 - 2r \in [\alpha^-, \alpha^+] \implies B_{k,\delta} \in [\beta^- - \epsilon, \beta^+ + \epsilon]$$

*where*

$$\begin{aligned}
\alpha^{\pm} &= \frac{1}{k}\frac{n-k}{n-1}\left(1 \pm \frac{\sqrt{k-1}}{n} \cdot \frac{\delta(1-\delta)}{5} \cdot \epsilon\right), \\
\beta^- &= \frac{k\omega_k - \delta}{1-\delta}, \quad \beta^+ = \frac{k\omega_k}{\delta} \\
\omega_k &= \frac{(k-1)^{k-1}}{k!e^{k-1}} \\
B_{k,\delta} &:= \sum_{\delta n/k \le a+1 \le n/k} p(\vee^a \mathbb{S}^{2k-2}, n, r)
\end{aligned}$$

*The bounds $\beta^{\pm}$ satisfy $\beta^- \le k\omega_k \le \beta^+$. Also $\beta^- > 0$ iff $\delta < k\omega_k$ and $\beta^+ < 1$ iff $\delta > k\omega_k$.*

*Proof.* We first describe the heuristic reasoning for the bounds, which is rather simple. Proposition 8.4.7 gives us:

$$\frac{kA_k - \delta n}{(1-\delta)n + k} \leq B \leq \frac{kA_k}{\delta n}$$

By Proposition 8.3.4 and 8.4.5, the upper bound has the following approximations:

$$\frac{kA_k}{\delta n} \approx \frac{k\bar{\chi}}{\delta n} \approx \frac{k\omega_k}{\delta}$$

and similarly the lower bound has the following approximations:

$$\frac{kA_k - \delta n}{(1-\delta)n + k} \approx \frac{kA_k - \delta n}{(1-\delta)n} \approx \frac{k\bar{\chi} - \delta}{1-\delta} \approx \frac{k\omega_k - \delta}{1-\delta}$$

The actual proof becomes more complicated due to using a different choice of $\epsilon$ in applying Proposition 8.3.4.

Let $\epsilon' = \delta(1-\delta) \cdot \epsilon/5$. We apply Proposition 8.3.4 with $\epsilon'$ taking the role of $\epsilon$, and this gives the choice of $\alpha^{\pm}$ in the theorem. Therefore $r \in [\alpha^-, \alpha^+]$ implies the following:

$$(1-\epsilon')\omega_k \leq \frac{\bar{\chi}}{n} \leq (1+\epsilon')\omega_k \tag{8.4.1}$$

Before going further, we note the following inequalities for $\epsilon'$, which we will use later:

$$\begin{aligned} \epsilon' = \frac{\delta(1-\delta)\epsilon}{4+1} &\leq \frac{\delta(1-\delta)\epsilon}{4+\delta(1-\delta)\epsilon} \\ \implies \frac{\epsilon'}{1-\epsilon'} &\leq \frac{\delta(1-\delta)\epsilon}{4} \\ \implies \frac{\epsilon'}{1-\epsilon'} &\leq \min\left(4\delta, \delta^{-1} - 1, 1\right) \cdot \frac{\epsilon}{4} \end{aligned} \tag{8.4.2}$$

**Upper bound.**

By Equation (8.4.1) and Proposition 8.4.5, we have:

$$\frac{k\omega_k}{\delta} \geq \frac{1}{1+\epsilon'}\frac{k\bar{\chi}}{\delta n} \geq \frac{1}{1+\epsilon'}\frac{kA_k}{\delta n}$$

By Equation (8.4.2), we have that:

$$\frac{1}{1+\epsilon'}\frac{kA_k}{\delta n} \geq \frac{kA_k}{\delta n} - \epsilon$$

Then Proposition 8.4.7 applies and we have the upper bound.

**Lower bound.**

By Equation (8.4.1) and Proposition 8.4.5, we have:

$$\frac{k\omega_k - \delta}{1-\delta} \leq \frac{1}{1-\delta}\left(\frac{1}{1-\epsilon'}\frac{k\bar{\chi}}{n} - \delta\right) \leq \frac{1}{1-\delta}\left(\frac{1}{1-\epsilon'}\frac{k^2 + kA_k}{n} - \delta\right)$$

Let $L_0$ be the right hand side. We rewrite it as follows:

$$L_0 = L_1 + E_1 = L_2 + E_1 + E_2$$

where

$$\begin{aligned} L_1 =& \frac{kA_k - \delta n}{(1-\delta)(1-\epsilon')n},\ E_1 = \frac{\delta\epsilon' + k^2/n}{(1-\delta)(1-\epsilon')} \\ L_2 =& \frac{kA_k - \delta n}{(1-\delta)n + k},\ E_2 = \frac{kA_k - \delta n}{(1-\delta)(1-\epsilon')n} \cdot \frac{k + (1-\delta)n\epsilon'}{(1-\delta)n + k} \end{aligned}$$

By Equation (8.4.2), the relation $kA_k \leq n$ and by taking $n$ large enough, we see that

$$E_1, E_2 \leq \epsilon/2$$

This implies that:

$$\frac{k\omega_k - \delta}{1-\delta} - \epsilon \leq L_0 - \epsilon = L_2 + E_1 + E_2 - \epsilon \leq L_2$$

Then again Proposition 8.4.7 applies and we have the lower bound.

□

We remark that Theorem C is obtained by setting $\epsilon = \delta = (1-\alpha)k\omega_k/2$. The gap $\alpha^+ - \alpha^-$ is replaced by a smaller but simpler quantity.

## 8.5 Odd spheres

We prove Theorem B using the stability of persistence diagram. In this case, we will be using the Cech complex constructed from the full set of the circle, and then bound the Gromov-Hausdorff distance between the full circle and a finite sample of it. We use the following result from [3]:

**Theorem 8.5.1.** *The homotopy types of the Rips and Cech complexes on the circle of unit circumference are as follows:*

$$\mathrm{VR}(\mathbb{S}^1, r) \simeq \begin{cases} \mathbb{S}^{2l+1} & ,if\ \frac{l}{2l+1} < r < \frac{l+1}{2l+3} \\ \bigvee^{\mathfrak{c}} \mathbb{S}^{2l} & ,if\ r = \frac{l}{2l+1} \end{cases}$$

$$\check{C}(\mathbb{S}^1, r) \simeq \begin{cases} \mathbb{S}^{2l+1} & ,if\ \frac{l}{2l+2} < r < \frac{l+1}{2l+4} \\ \bigvee^{\mathfrak{c}} \mathbb{S}^{2l} & ,if\ r = \frac{l}{2l+2} \end{cases}$$

*where* $\mathfrak{c}$ *is the cardinality of the continuum.*

We also note the stability of persistence:

**Theorem 8.5.2** (Stability of Persistence)**.** *If* $X, Y$ *are metric spaces and* $\mathcal{D}_k M$ *is the* $k$*-dimensional persistence diagram of persistence module* $M$*, then*

$$\mathrm{d}_B(\mathcal{D}_k \mathbf{VR}(X), \mathcal{D}_k \mathbf{VR}(Y)) \leq \mathrm{d}_{GH}(X, Y)$$
$$\mathrm{d}_B(\mathcal{D}_k \check{C}(X), \mathcal{D}_k \check{C}(Y)) \leq \mathrm{d}_{GH}(X, Y)$$

*where* $\mathrm{d}_{GH}$ *denotes the Gromov-Hausdorff distance.*

The following proposition is a more precise version of Theorem B, which specifies an explicit lower bound for the probabilities of homotopy equivalence:

**Proposition 8.5.3.** *For each* $l \geq 0$ *and* $t \in (\frac{l}{2l+2}, \frac{l+1}{2l+4})$*, the following holds with probability at least* $Q_n(r')$*:*

$$\check{C}(\mathbf{X}_n, t) \simeq \mathbb{S}^{2l+1}$$

*where* $r'$ *is:*

$$r' = \frac{1}{4(l+1)(l+2)} - \left| t - \frac{2l^2 + 4l + 1}{4(l+1)(l+2)} \right|$$

*Proof.* Consider a random sample $\mathbf{X}_n = (X_1, \ldots X_n)$. Then with probability $Q_n(r)$, arcs of radius $r$ centered at $\mathbf{X}_n$ covers $\mathbb{S}^1$, so that $\mathrm{d}_{GH}(\mathbf{X}_n, \mathbb{S}^1) \leq \mathrm{d}_H(\mathbf{X}_n, \mathbb{S}^1) \leq r$. This implies:

$$\mathrm{d}_B(\mathcal{D}_k \check{\mathrm{C}}(\mathbf{X}_n), \mathcal{D}_k \check{\mathrm{C}}(\mathbb{S}^1)) \leq \mathrm{d}_{GH}(\mathbf{X}_n, \mathbb{S}^1) \leq r$$

For each $l \geq 0$, we have that:

$$\mathcal{D}_{2l+1} \check{\mathrm{C}}(\mathbb{S}^1) = \left\{ \left( \frac{l}{2l+2}, \frac{l+1}{2l+4} \right) \right\}$$

so that the definition of the bottleneck distance implies that

$$\begin{gathered}\exists(u,v) \in \mathcal{D}_{2l+1}\check{\mathrm{C}}(\mathbf{X}_n) \\ \text{with } \frac{l}{2l+2} - r \leq u \leq \frac{l}{2l+2} + r \\ \frac{l+1}{2l+4} - r \leq v \leq \frac{l+1}{2l+4} + r\end{gathered}$$

This implies that whenever $\frac{l}{2l+2} + r \leq t \leq \frac{l+1}{2l+4} - r$, we have:

$$1 \leq \dim H_{2l+1}\check{\mathrm{C}}(\mathbf{X}_n, t)$$

and due to the enumeration of possible homotopy types, we have that:

$$\check{\mathrm{C}}(\mathbf{X}_n, t) \simeq \mathbb{S}^{2l+1}$$

The condition translates to $\left|t - \frac{1}{2}\left(\frac{l}{2l+2} + \frac{l+1}{2l+4}\right)\right| < \frac{1}{2}\left(\frac{l+1}{2l+4} - \frac{l}{2l+2}\right) - r$, or equivalently

$$\left|t - \frac{2l^2+4l+1}{4(l+1)(l+2)}\right| < \frac{1}{4(l+1)(l+2)} - r$$

and thus we obtain the proof.

□

### 8.5.1 Concluding remarks

This chapter is the last of this thesis, and studies a rather different problem from the chapters leading up to it: the random topology of a Cech complex constructed on a simple data manifold: circle. In the chapter, it was proven that "unexpected" homotopy types (bouquet of high-dimensional spheres) that disagree with the underlying topology (circle) arise with high probability. The results call us to examine a paradigm in topological data analysis, which implicitly assumes that the objective of a topological inference process is to solely recover the underlying topology, while discarding homotopy types that differ from it by regarding them as "noise". Even though this chapter is about topological inference, its spirit is the same one mentioned in the title: "Geometric and topological inference from random samples". The entire thesis focuses on careful examination of some commonly overlooked themes in geometric and topological inferences.

The topic of this chapter began at the very beginning, not end, of the DPhil degree. Initially, the author was suggested to create a singularity detection algorithm based on

topological methods. But one of the first discoveries was that unexpected homotopy types arise from the Vietoris-Rips complexes over spheres, suggesting that there are deeper secrets about topological inferences that are yet to be understood. After many trial-and-error, the author would discover that there are faster, better understood classical techniques that achieves the same job, and that is how the Hades algorithm of Chapter 7 was born. But separately to Hades, the emergence of unexpected high-dimensional topology was still intriguing, so the current chapter was born after careful calculations. This chapter marks a new beginning for topological data analysis, to treat the "topological noise" with more sincerity and uncover the mysteries in topological inference.

# Bibliography


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

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

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

[4] Michał Adamaszek and Henry Adams. On vietoris–rips complexes of hypercube graphs. *Journal of Applied and Computational Topology*, 6(2):177–192, 2022.

[5] Michał Adamaszek, Henry Adams, and Florian Frick. Metric reconstruction via optimal transport. *SIAM Journal on Applied Algebra and Geometry*, 2(4):597–619, 2018.

[6] Michał Adamaszek, Henry Adams, Florian Frick, Chris Peterson, and Corrine Previte-Johnson. Nerve complexes of circular arcs. *Discrete & Computational Geometry*, 56:251–273, 2016.

[7] Michał Adamaszek, Henry Adams, and Francis Motta. Random cyclic dynamical systems. *Advances in Applied Mathematics*, 83:1–23, 2017.

[8] Michał Adamaszek, Henry Adams, and Samadwara Reddy. On vietoris–rips complexes of ellipses. *Journal of Topology and Analysis*, 11(03):661–690, 2019.

[9] Henry Adams, Samir Chowdhury, Adam Quinn Jaffe, and Bonginkosi Sibanda. Vietoris-rips complexes of regular polygons. *arXiv preprint arXiv:1807.10971*, 2018.

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

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

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

[13] Amir Babaeian, Alireza Bayestehtashk, and Mojtaba Bandarabadi. Multiple manifold clustering using curvature constrained path. *PloS one*, 10(9):e0137986, 2015.

[14] Peter Bartlett. Theoretical statistics; stat 210b. 2013.

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

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

[17] Paul Bendich, David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Dmitriy 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.

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

[19] Paul Bendich, Bei Wang, and Sayan 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.

[20] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021.

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

[22] Omer Bobrowski and Matthew Kahle. Topology of random geometric complexes: a survey. *Journal of applied and Computational Topology*, 1:331–364, 2018.

[23] Omer Bobrowski and Goncalo Oliveira. Random čech complexes on riemannian manifolds. *Random Structures & Algorithms*, 54(3):373–412, 2019.

[24] Omer Bobrowski and Shmuel Weinberger. On the vanishing of homology in random čech complexes. *Random Structures & Algorithms*, 51(1):14–51, 2017.

[25] Emmanuel Boissard and Thibaut 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.

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

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

[28] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[29] Bradley CA Brown, Anthony L Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. The union of manifolds hypothesis. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.

[30] Bradley CA Brown, Anthony L Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2022.

[31] Francesco Camastra and Antonino Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.

[32] Francesco Camastra and Alessandro 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.

[33] Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern recognition*, 47(8):2569–2581, 2014.

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

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

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

[37] Henry-Louis de Kergorlay, Ulrike Tillmann, and Oliver Vipond. Random čech complexes on manifolds with boundary. *Random Structures & Algorithms*, 61(2):309–352, 2022.

[38] David Donoho, Matan Gavish, and Elad Romanov. Screenot: Exact mse-optimal singular value thresholding in correlated noise. *The Annals of Statistics*, 51(1):122–148, 2023.

[39] Amir Massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 265–272, 2007.

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

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

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

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

[44] Christopher R Genovese, Marco Perone Pacifico, Isabella Verdinelli, Larry Wasserman, et al. Minimax manifold estimation. *Journal of machine learning research*, 13:1263–1291, 2012.

[45] Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012.

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

[47] Jean-Claude Hausmann et al. On the vietoris-rips complexes and a cohomology theory for metric spaces. *Annals of Mathematics Studies*, 138:175–188, 1995.

[48] Martin Helmer and Vidit Nanda. Conormal spaces and whitney stratifications. *Foundations of Computational Mathematics*, pages 1–36, 2022.

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

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

[51] Matthew Kahle. Random geometric complexes. *Discrete & Computational Geometry*, 45:553–573, 2011.

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

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

[54] Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax rates for estimating the dimension of a manifold. *arXiv preprint arXiv:1605.01011*, 2016.

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

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

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

[58] John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*, volume 1. Springer.

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

[60] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

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

[62] Sunhyuk Lim, Facundo Memoli, and Osman Berat Okutan. Vietoris-rips persistent homology, injective metric spaces, and the filling radius. *arXiv preprint arXiv:2001.07588*, 2020.

[63] Uzu Lim. Strange random topology of the circle. *Discrete & Computational Geometry*, pages 1–26, 2025.

[64] Uzu Lim, Harald Oberhauser, and Vidit Nanda. Tangent space and dimension estimation with the wasserstein distance. *SIAM Journal on Applied Algebra and Geometry*, 8(3):650–685, 2024.

[65] Uzu Lim, Harald Oberhauser, and Vidit Nanda. Hades: Fast singularity detection with local measure comparison. *SIAM Journal on Mathematics of Data Science*, 7(4):1882–1903, 2025.

[66] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[67] Shawn Martin, Aidan Thompson, Evangelos A Coutsias, and Jean-Paul Watson. Topology of cyclo-octane energy landscape. *The journal of chemical physics*, 132(23):234115, 2010.

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

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

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

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

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

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

[74] Taejin Paik and Otto van Koert. Expected invariants of simplicial complexes obtained from random point samples. *Archiv der Mathematik*, 120(4):417–429, 2023.

[75] Mathew Penrose. *Random geometric graphs*, volume 5. OUP Oxford, 2003.

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

[77] Elad Romanov. On the noise sensitivity of the randomized svd. *arXiv preprint arXiv:2305.17435*, 2023.

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

[79] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath 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.

[80] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[81] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

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

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

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

[85] Konstantinos Slavakis, Shiva Salsabilian, David S Wack, Sarah F Muldoon, Henry E Baidoo-Williams, Jean M Vettel, Matthew Cieslak, and Scott T Grafton. Riemannian multi-manifold modeling and clustering in brain networks. In *Wavelets and Sparsity XVII*, volume 10394, pages 9–24. SPIE, 2017.

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

[87] WL Stevens. Solution to a geometrical problem in probability. *Annals of Eugenics*, 9(4):315–320, 1939.

[88] Bernadette J Stolz, Jared Tanner, Heather A Harrington, and Vidit Nanda. Geometric anomaly detection in data. *Proceedings of the National Academy of Sciences*, 117(33):19664–19669, 2020.

[89] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[90] Andrew M Thomas and Takashi Owada. Functional limit theorems for the euler characteristic process in the critical regime. *Advances in Applied Probability*, 53(1):57–80, 2021.

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

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

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

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

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

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

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

[98] Julius Von Rohrscheidt and Bastian Rieck. Topological singularity detection at multiple scales. 2023.

[99] Lukas Waas and Tim Mäder. From samples to persistent stratified homotopy types. *arXiv preprint arXiv:2206.08926*, 2022.

[100] Xu Wang, Konstantinos Slavakis, and Gilad Lerman. Multi-manifold modeling in non-euclidean spaces. In *Artificial Intelligence and Statistics*, pages 1023–1032. PMLR, 2015.

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

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

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