

Robust Topological Representation Learning

Julius Christopher von Rohrscheidt

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology
der Technischen Universität München zur Erlangung eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation.

Vorsitz: Prof. Dr. Ulrich Bauer

Prüfende der Dissertation:

1. Prof. Dr. Bastian Grossenbacher-Rieck
2. Prof. Dr. Roland Kwitt

Die Dissertation wurde am 16.06.2025 bei der Technischen Universität München eingereicht
und durch die TUM School of Computation, Information and Technology am 28.11.2025
angenommen.

Abstract

Modern machine learning has achieved remarkable success across a wide range of applications, from natural language processing to image recognition and far beyond. These advances are driven by powerful architectures, large-scale data, and increasingly efficient training procedures. However, as machine learning models are applied to complex, high-dimensional, and heterogeneous data, their performance face fundamental limitations. Standard approaches often rely on statistical assumptions that oversimplify the true structure of data and provide limited interpretability into what models actually learn. In response to these challenges, this thesis develops mathematically grounded methods that enhance machine learning by integrating insights from topology and geometry. These fields study the shape and structure of data at both local and global scales, offering tools that go beyond what classical statistics can capture. Our contributions include techniques for detecting singular regions in data, characterizing the local structure of graphs, integrating topology into Bayesian workflows, and analyzing how the internal geometry of large language models evolve during training. By embedding topological and geometric reasoning into machine learning workflows, we improve the interpretability, robustness, and performance of models, particularly in settings where conventional methods fall short. Overall, this work offers a topological and geometric perspective that complements the strengths of modern machine learning while addressing its current limitations.

Zusammenfassung

Moderne Verfahren des maschinellen Lernens haben in den vergangenen Jahren beachtliche Erfolge in einer Vielzahl von Anwendungsgebieten erzielt, von der Verarbeitung natürlicher Sprache über die Bildanalyse bis hin zu zahlreichen weiteren Bereichen. Diese Fortschritte beruhen auf leistungsfähigen Modellarchitekturen, großen Datenmengen und immer effizienteren Trainingsmethoden. Gleichzeitig stoßen viele dieser Ansätze an grundsätzliche Grenzen, insbesondere wenn sie auf komplexe, hochdimensionale und heterogene Daten angewendet werden. Häufig basieren die Verfahren auf stark vereinfachenden statistischen Annahmen, die der tatsächlichen Struktur der Daten nicht gerecht werden und nur begrenzt Einblick in das ermöglichen, was ein Modell tatsächlich lernt. Diese Dissertation entwickelt daher neue, mathematisch fundierte Methoden, die maschinelles Lernen um Konzepte aus der Topologie und Geometrie erweitern. Diese Disziplinen untersuchen die Form und Struktur von Daten sowohl auf lokaler als auch globaler Ebene und bieten Werkzeuge, die über klassische statistische Verfahren hinausgehen. Im Mittelpunkt der Arbeit stehen unter anderem Verfahren zur Erkennung singulärer Regionen in Daten, zur Beschreibung lokaler Strukturen in Graphen, zur Integration topologischer Informationen in Bayessche Modellierungsansätze sowie zur Analyse der internen Geometrie großer Sprachmodelle während des Trainingsprozesses. Durch die gezielte Einbettung topologischer und geometrischer Überlegungen in maschinelle Lernverfahren lässt sich die Interpretierbarkeit, Robustheit und Leistungsfähigkeit der Modelle verbessern – insbesondere in Situationen, in denen klassische Methoden an ihre Grenzen stoßen. Insgesamt bietet diese Arbeit eine neue Perspektive auf das maschinelle Lernen, die dessen Stärken ergänzt und zugleich bestehende Schwächen adressiert.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Dr. Bastian Rieck, for his exceptional guidance, unwavering support, and constant encouragement throughout my doctoral journey. His insights, patience, and mentorship have been instrumental in shaping both my research and my academic growth. I am equally grateful to Prof. Dr. Ulrich Bauer for integrating me into the academic environment at TUM and for his continuous support throughout this journey. His perspectives enriched this work and broadened my understanding of the field. I would also like to thank all my colleagues from the AIDOS Lab for the many inspiring conversations, collaborative spirit, and the intellectually stimulating environment they have helped create. Special thanks go to Dr. Sebastian M. Schmon and Dr. Benjamin M. Ruppik for the fruitful collaborations, thoughtful discussions, and their enthusiasm throughout our joint work. I acknowledge the use of OpenAI [130] for language editing assistance in this work. All content was reviewed and verified by the author. To my family, thank you for your unconditional love, patience, and belief in me, especially during the more difficult stretches of this journey. Your support has been my anchor. Finally, I dedicate this thesis to my godfather Prof. Dr. Christopher P. Kobrak, whose intellectual legacy and personal impact continue to inspire me deeply. May this work reflect, in part, the commitment and curiosity he embodied.

Contents

1. Introduction	1
1.1. Motivation	2
1.2. Outline	3
2. Algebraic Topology	5
2.1. Topological Spaces	5
2.2. Some Homological Algebra	5
2.3. Simplicial Complexes	8
2.4. Simplicial Homology	9
2.5. Relative Homology	14
2.6. Local Homology	15
2.7. Manifolds	15
2.7.1. Local homology of manifolds	16
2.8. Stratified Spaces	17
2.9. Euler Characteristic	18
2.10. Local-Global Principles in Topology	19
3. Algebraic Topology and Data	21
3.1. Persistent Homology	21
3.2. Decomposition of Persistence Modules	22
3.3. Persistence Diagrams	22
3.4. Vietoris-Rips and Čech Complexes	23
3.5. The Nerve Theorem	23
3.6. Wasserstein and Bottleneck Distances	25
3.7. Stability of Persistent Homology	25
3.8. Multiparameter Persistence	26
3.9. Euler Characteristic Transforms	27
I. Topological Representation Learning	29
4. Local Topology and Singularities	31
4.1. Introduction	31
4.2. Related Work	33
4.3. Methods	35
4.3.1. Persistent Local Homology	35
4.3.2. Persistent Intrinsic Dimension (PID)	37
4.3.3. Euclidicity	38
4.4. Experiments	41
4.4.1. Parameter Selection	41

Contents

4.4.2. Persistent Intrinsic Dimension is Expressive	42
4.4.3. Euclidicity Captures Known Singularities	43
4.4.4. Stability of Euclidicity	44
4.4.5. Comparison to single-parameter approach	44
4.4.6. Euclidicity of High-Dimensional Spaces	45
4.4.7. Euclidicity of Cytometry Data	47
4.5. Discussion	48
5. Local Topology and Graph Data	51
5.1. Introduction	51
5.2. Related Work	52
5.3. Methods	54
5.3.1. The ECT and its approximation	54
5.3.2. The local ECT (ℓ -ECT)	55
5.3.3. ℓ -ECT in representation learning	56
5.3.4. Subgraph counting	57
5.3.5. Rotation-invariant metric based on local ECTs	59
5.4. Experiments	60
5.4.1. ℓ -ECTs in Graph Representation Learning	61
5.4.2. Implementation details	62
5.4.3. WebKB datasets	63
5.4.4. Heterophilous datasets	64
5.4.5. Amazon dataset	64
5.4.6. Actor dataset	66
5.4.7. WikipediaNetwork dataset	66
5.4.8. Planetoid datasets	67
5.4.9. Post-hoc evaluation	67
5.4.10. Learning Spatial Alignment of Geometric Graphs	69
5.5. Discussion	71
6. Global Topology and Bayesian Inference	73
6.1. Introduction	73
6.2. Related Work	75
6.3. Methods	76
6.3.1. Comparison-Based Posteriors	76
6.3.2. Inference for Comparison-Based Posteriors	80
6.3.3. Losses Based on Geometry & Topology	81
6.4. Experiments	83
6.4.1. Vicsek Model: Swarm Behaviour	84
6.4.2. Fluid Dynamics: Lattice Boltzmann Simulation	86
6.4.3. Percolation Model: Multi-Scale Structures	88
6.5. Discussion	89
7. Local Geometry and Large Language Models	91
7.1. Introduction	91
7.2. Related Work	93
7.3. Methods	94
7.3.1. Large Language Models	94

7.3.2. Latent Space Modeling	95
7.3.3. Comparing Latent Spaces Across Models	96
7.3.4. Local Dimension Estimates	97
7.4. Experiments	98
7.4.1. Fine-Tuning Induces Dataset-Specific Shifts in Local Dimensions	98
7.4.2. Local Dimensions Predict Grokking	101
7.4.3. Local Dimensions Detect Exhaustion of Training Capabilities	103
7.4.4. Local Dimensions Detect Overfitting	105
7.4.5. Computational Resources	108
7.4.6. Limitations	108
7.4.7. Discussion	109
8. Conclusion and Future Work	111
A. Appendix	113
B. Appendix	125
C. Appendix	133
D. Appendix	141
Bibliography	155

1. Introduction

Understanding the fundamental structures underlying complex data is a cornerstone of modern machine learning and data science. Traditional methods often rely on purely statistical approaches, which, while powerful, may fail to capture the deeper geometric and topological properties of data. The increasing complexity of data modalities, from point clouds and graphs to high-dimensional manifolds, calls for robust mathematical frameworks that go beyond classical representations. Topological methods offer a promising avenue in this regard, providing tools that can encode shape and other structural properties of data in ways that are inherently more stable to noise and deformation than traditional methods. In particular, the field of Topological Data Analysis (TDA) has emerged as a potent bridge between algebraic topology and data-driven representation learning, leading to significant advances in understanding the latent structure of complex data spaces. Its primary goal is to extract robust and interpretable topological features from complex, often high-dimensional datasets. The central tool in TDA is *persistent homology*, which enables the identification and quantification of topological structures such as connected components, loops, and voids across multiple scales. In this sense, it captures the (global) *shape* of the data in a way that accounts for varying scale parameters.

Bridging the gap between TDA and machine learning involves developing methodologies that integrate topological insights into data-driven pipelines in a principled manner. A comprehensive survey by Hensel et al. [85] categorizes the use of TDA in machine learning into two main interventional strategies: (i) using topological signatures as features for downstream tasks, and (ii) incorporating topological losses or inductive biases directly into model architectures. On one hand, persistent diagrams and other topological summaries can be embedded into vector spaces using kernel methods or learned representations, allowing them to serve as informative features for conventional models. On the other hand, recent work focuses on embedding topological priors directly into learning objectives, either by penalizing undesirable topological artifacts (e.g., spurious loops in generative models) or by encouraging desired structures through topological regularization. These strategies not only enhance interpretability but also improve robustness and generalization, particularly in settings with limited or noisy data. As such, TDA offers both a complementary perspective and a powerful toolkit for enriching modern machine learning architectures.

In supervised learning, topological features have improved performance by incorporating topology-aware layers into machine learning pipelines [86, 33]. Topological priors have also been incorporated into neural networks to enforce structural constraints, such as encouraging the presence (or absence) of loops in generated

1. Introduction

data, with differentiable analogues of persistent homology allowing gradients to flow through topological computations [42]. Another development is the introduction of *topological autoencoders*, which explicitly preserve topological features in the learned latent spaces. This leads to improved interpretability and a greater degree of disentanglement in the representations [124]. These tools have enabled end-to-end training of deep learning models where topological reasoning is a core component. Recent work has extended TDA to more complex data types and tasks. One major direction is *topological graph representation learning*, which leverages persistent homology to capture multi-scale structures that are not readily visible to local message-passing schemes used in graph neural networks (GNNs) [90, 39]. Furthermore, *temporal extensions* of TDA have been proposed, allowing for shape-aware analysis in dynamic data streams [146].

As the field matures, TDA offers a compelling complement to geometric deep learning, providing a mathematically grounded framework for incorporating structural insights into modern ML pipelines. While persistent homology excels at capturing global topological features of the data [63, 214], a comprehensive understanding often also requires attention to local geometric structures such as curvature, density variations, and local connectivity patterns—elements that play a pivotal role in tasks like clustering [65], anomaly detection [34], or manifold learning [12]. Recent developments in topological methods aim to bridge this gap by incorporating tools sensitive to both local and global aspects, such as multiscale sheaf theory [49]. These approaches enrich the traditional TDA pipeline, offering a more nuanced view of data geometry and topology. Taken together, they highlight the vast potential of topological methods not only for capturing shape, but also for informing and enhancing machine learning models with mathematically grounded structural insights on different locality scales.

1.1. Motivation

Modern machine learning systems have achieved remarkable successes across a variety of domains—from computer vision and natural language processing to genomics and dynamical systems modeling. However, the increasing complexity and heterogeneity of real-world data challenge traditional representation learning techniques. These models, often optimized in high-dimensional Euclidean spaces, struggle when the underlying data geometry deviates from manifold assumptions, or when the data exhibits singularities, discontinuities, or multi-scale structural patterns.

This thesis is concerned with the development of robust topological representation learning, meaning methods that leverage topological structures to build representations that are not only expressive but also more resilient to perturbations. A representation learning paradigm that incorporates topological priors has the potential to address these issues by capturing intrinsic structures that remain stable across transformations. A core insight is that topological structures encode features that are often invariant under common sources of noise or deformation. Yet, despite the theoretical stability guarantees offered by TDA, such methods have traditionally

played only a peripheral role in modern machine learning pipelines. However, the robustness of machine learning models is not only a theoretical concern but also a pressing practical issue. In domains such as computational biology, medical imaging, and physics-based simulations, data is often high-dimensional and inherently noisy. Topological representations often provide a natural means to handle such challenges by offering invariance to deformations, occlusions, and missing data.

The goal of this research is to integrate topological thinking into the foundations of representation learning. Rather than treating topology as an afterthought, e.g. through post hoc analysis of neural embeddings, this work explores the design of representation learning methods that are informed by and adaptive to the topological structure of data from the outset. In particular, we focus on local and global topological descriptors that reflect intrinsic spatial patterns, and show their utility in data science and machine learning tasks.

More broadly, the methodological contributions of this thesis emerge from the synthesis of algebraic topology, geometry and machine learning. Topological methods are particularly well-suited to the challenges of modern data analysis: they are coordinate-free, naturally multi-scale, and grounded in a rich mathematical theory with provable stability properties.

The long-term vision of this line of research is to bridge the gap between statistical learning theory and geometric-topological insights. As data grows in complexity and dimensionality, it becomes ever more essential to develop representation learning methods that are not only empirically powerful but also geometrically aware and theoretically principled. By introducing and validating new topological descriptors and inference frameworks, this thesis contributes to that goal, by opening up new avenues for machine learning models that are more interpretable and aligned with the structural properties of the data they are designed to understand.

1.2. Outline

This thesis is structured into the following chapters.

Chapter 2: Algebraic Topology

This chapter provides the necessary mathematical and conceptual background for the thesis. It introduces foundational tools from algebraic topology, including topological spaces, simplicial complexes, homology, and the Euler characteristic. Special attention is given to local homology, as well as the topology of manifolds and stratified spaces. The goal is to equip the reader with a rigorous toolkit for reasoning about both global and local structure in complex data.

Chapter 3: Algebraic Topology and Data

This chapter establishes the connection between algebraic topology and data analysis. It introduces persistent homology as a method for extracting multiscale topological features from data, and presents core constructions such as persistence diagrams. It also discusses practical aspects including the choice of filtrations, sta-

1. Introduction

bility of persistence, and appropriate distance metrics. Extensions to multiparameter persistence and the Euler Characteristic Transform are covered to motivate later applications in representation learning.

Chapter 4: Local Topology and Singularities

Chapter 4 focuses on the challenges posed by data with singularities or varying local geometry. It introduces Persistent Local Homology and Persistent Intrinsic Dimension as tools to characterize the local topological complexity of datasets. The notion of Euclidicity is proposed as a measure for how Euclidean a local neighborhood is. Through empirical studies, this chapter demonstrates how these descriptors can uncover hidden structure and inform learning tasks sensitive to local irregularities.

Chapter 5: Local Topology and Graph Data

This chapter addresses representation learning on graphs, where conventional graph neural networks may be limited by oversmoothing and local aggregation. It introduces the local Euler Characteristic Transform as a stable, expressive descriptor for graph neighborhoods. This tool is shown to capture both geometric and structural information, and is applied to a variety of benchmark tasks in node classification. The chapter also proposes a rotation-invariant metric for comparing graphs via topological signatures.

Chapter 6: Global Topology and Bayesian Inference

Chapter 6 explores the integration of topological insights into probabilistic modeling. It develops a framework for Bayesian inference in settings where likelihoods are implicitly defined or difficult to compute. Comparison-based posteriors and topology-aware losses are introduced as ways to incorporate structural information into inference. Applications to agent-based systems and physical simulations demonstrate the utility of topological insights in complex domains, enabling robust and interpretable uncertainty quantification.

Chapter 7: Local Geometry and Large Language Models

This chapter applies topological ideas to the latent spaces of large language models. It studies how fine-tuning affects the local dimensionality of contextual embeddings by measuring local intrinsic dimension. These geometric descriptors are shown to be predictive of phenomena such as grokking, overfitting, and training saturation. This chapter highlights the potential of topological tools for unsupervised diagnostics and interpretability in modern machine learning systems.

Conclusion and Outlook

The thesis concludes with a synthesis of its main contributions discovered in the preceding chapters, and a forward-looking discussion of open directions.

2. Algebraic Topology

2.1. Topological Spaces

We start this chapter with the very basic concepts of topology, as we will need them for the remainder of this text.

Definition 2.1.1. A topological space is a set X together with a collection $\mathcal{T}$ of subsets of X , called open sets, such that:

1. The empty set $\emptyset$ and the entire set X are in $\mathcal{T}$.
2. The union of any collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$.
3. The intersection of any finite collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$.

The collection $\mathcal{T}$ is called a topology on X , and the pair $(X, \mathcal{T})$ is called a topological space. When the topology $\mathcal{T}$ is understood from the context, we often abbreviate $(X, \mathcal{T})$ by simply writing X .

A subset $A \subset X$ of a topological space X is said to be *closed* if its complement $X \setminus A$ is open.

Example 2.1.1. Given a topological space $(X, \mathcal{T})$ and a subset $A \subset X$, the subspace topology on A is the collection $\mathcal{T}_A = \{U \cap A \mid U \in \mathcal{T}\}$. The pair $(A, \mathcal{T}_A)$ is then a topological space called a subspace of X .

Definition 2.1.2. A continuous map between two topological spaces $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ is a function $f : X \rightarrow Y$ such that for every open set $V \in \mathcal{S}$, the preimage $f^{-1}(V)$ is in $\mathcal{T}$. That is, f preserves the structure of open sets.

Definition 2.1.3. A homeomorphism is a continuous map $f : X \rightarrow Y$ between two topological spaces X and Y that has a continuous inverse $f^{-1} : Y \rightarrow X$. If such a map exists, we say that X and Y are homeomorphic and that they share the same topological structure.

2.2. Some Homological Algebra

This section summarizes some basic concepts and results from homological algebra which we will need for the remainder of this chapter. The objects of this section are purely algebraic and the results have found its utility in a variety of mathematical

2. Algebraic Topology

disciplines, such as topology, geometry and number theory. Algebraic chain complexes are a fundamental concept in homological algebra, describing a sequence of abelian groups (or modules) linked by homomorphisms, with the property that the composition of consecutive homomorphisms is zero. Chain complexes are central to defining homology groups, which usually are used to capture information about an underlying object of interest, encoded in an algebraic way. The idea of this concept is that the underlying object of interest is often difficult to understand, while their corresponding homology groups are easier to work with since they come with additional structure. Proofs of this section are omitted as they can be found in every comprehensive text about homological algebra.

Definition 2.2.1. A chain complex $(C_\bullet, \partial_\bullet)$ is a sequence of abelian groups (or modules) $\{C_n\}_{n \in \mathbb{Z}}$ together with homomorphisms called boundary maps $\partial_n : C_n \rightarrow C_{n-1}$ for each $n \in \mathbb{Z}$, such that:

$$\partial_{n-1} \circ \partial_n = 0 \quad \text{for all } n \in \mathbb{Z}.$$

This condition, often expressed as $\partial^2 = 0$, implies that the image of each boundary map ∂_n is contained in the kernel of ∂_{n-1} , which allows for the definition of homology groups.

The chain complex $(C_\bullet, \partial_\bullet)$ is typically represented as:

$$\cdots \rightarrow C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \rightarrow \cdots$$

In this context:

- The elements of C_n are called n -chains.
- The subgroups $Z_n = \ker(\partial_n) \subseteq C_n$ are called n -cycles.
- The subgroups $B_n = \text{im}(\partial_{n+1}) \subseteq C_n$ are called n -boundaries.

The homology groups $H_n(C_\bullet) = Z_n/B_n$ measure the difference between cycles and boundaries in each degree n and are key invariants of the chain complex.

Definition 2.2.2. Let $(C_\bullet, \partial_\bullet^C)$ and $(D_\bullet, \partial_\bullet^D)$ be two chain complexes, where $C_\bullet = \{C_n\}_{n \in \mathbb{Z}}$ and $D_\bullet = \{D_n\}_{n \in \mathbb{Z}}$ are sequences of abelian groups (or modules) with boundary maps $\partial_n^C : C_n \rightarrow C_{n-1}$ and $\partial_n^D : D_n \rightarrow D_{n-1}$, respectively. A morphism of chain complexes $f_\bullet : C_\bullet \rightarrow D_\bullet$ is a sequence of homomorphisms $\{f_n : C_n \rightarrow D_n\}_{n \in \mathbb{Z}}$ such that the following diagram commutes for each $n \in \mathbb{Z}$:

$$\begin{array}{ccc} C_n & \xrightarrow{\partial_n^C} & C_{n-1} \\ f_n \downarrow & & \downarrow f_{n-1} \\ D_n & \xrightarrow{\partial_n^D} & D_{n-1} \end{array}$$

This commutativity condition, $f_{n-1} \circ \partial_n^C = \partial_n^D \circ f_n$, ensures that $f_\bullet$ respects the chain structure of $C_\bullet$ and $D_\bullet$.

Definition 2.2.3. A sequence of morphisms between chain complexes

$$\cdots \rightarrow A_{\bullet} \xrightarrow{f_{\bullet}} B_{\bullet} \xrightarrow{g_{\bullet}} C_{\bullet} \rightarrow \cdots$$

is called *exact* if the sequence of individual chain groups

$$\cdots \rightarrow A_n \xrightarrow{f_n} B_n \xrightarrow{g_n} C_n \rightarrow \cdots$$

is exact at each degree n , meaning that:

$$\text{im}(f_n) = \ker(g_n) \quad \text{for all } n \in \mathbb{Z}.$$

Exact sequences of chain complexes play a central role in homological algebra, as they allow for the construction of long exact sequences in homology and facilitate the study of homological properties across different complexes.

We now formulate the well-known Snake Lemma, which is a fundamental result in homological algebra that provides a long exact sequence associated with a commutative diagram of exact sequences. It is particularly useful for studying the relationship between kernels and cokernels in abelian categories, such as in the context of modules or abelian groups.

Proposition 2.2.1 (Snake Lemma). *Consider a commutative diagram of abelian groups (or modules) with exact rows:*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A' & \xrightarrow{f'} & A & \xrightarrow{f} & A'' & \longrightarrow & 0 \\ & & \downarrow g' & & \downarrow g & & \downarrow g'' & & \\ 0 & \longrightarrow & B' & \xrightarrow{h'} & B & \xrightarrow{h} & B'' & \longrightarrow & 0 \end{array}$$

where the rows are exact sequences and the vertical maps g' , g , and g'' are homomorphisms.

Then there exists a long exact sequence of the form:

$$0 \rightarrow \ker(g') \rightarrow \ker(g) \rightarrow \ker(g'') \xrightarrow{\delta} \text{coker}(g') \rightarrow \text{coker}(g) \rightarrow \text{coker}(g'') \rightarrow 0.$$

Here, δ is called the connecting homomorphism and is constructed from the data of the commutative diagram.

The Snake Lemma is widely used to analyze the structure of modules and abelian groups in the context of homological algebra and is foundational for constructing long exact sequences in various (co-)homology theories.

As a direct consequence of the Snake Lemma, we note that a short exact sequence of chain complexes gives rise to a long exact sequence on homology.

Corollary 2.2.1. *Let*

$$0 \rightarrow A_{\bullet} \xrightarrow{\alpha_{\bullet}} B_{\bullet} \xrightarrow{\beta_{\bullet}} C_{\bullet} \rightarrow 0$$

2. Algebraic Topology

be a short exact sequence of chain complexes, where $A_\bullet$, $B_\bullet$, and $C_\bullet$ are chain complexes, and each map $\alpha_n : A_n \rightarrow B_n$ and $\beta_n : B_n \rightarrow C_n$ is a chain map such that the sequence is exact at each degree n . Then, for each integer n , there is a long exact sequence in homology:

$$\cdots \rightarrow H_n(A_\bullet) \xrightarrow{H_n(\alpha)} H_n(B_\bullet) \xrightarrow{H_n(\beta)} H_n(C_\bullet) \xrightarrow{\delta_n} H_{n-1}(A_\bullet) \rightarrow \cdots$$

where δ_n is the connecting homomorphism induced by the Snake Lemma.

This long exact sequence describes the relationship between the homology groups $H_n(A_\bullet)$, $H_n(B_\bullet)$, and $H_n(C_\bullet)$ in terms of the original short exact sequence of chain complexes.

2.3. Simplicial Complexes

In this section, we introduce the fundamental concepts of simplices and simplicial complexes. These structures generalize the notion of triangles and tetrahedra to any number of dimensions, forming the basis to study a plethora of interesting objects in topology and geometry.

Definition 2.3.1. An n -simplex is the smallest convex set in a Euclidean space $\mathbb{R}^m$ that contains $n + 1$ points $v_0, \dots, v_n$, which are affinely independent.

Affine independence means that the points do not all lie in a hyperplane of dimension less than n . Equivalently, the difference vectors $v_1 - v_0, \dots, v_n - v_0$ are linearly independent, ensuring that these points uniquely determine an n -dimensional object. We call the points v_i the *vertices* of the simplex, and the simplex itself is denoted by $[v_0, \dots, v_n]$.

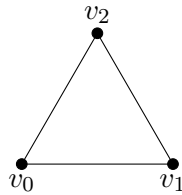
Example 2.3.1. 1. A 0-simplex: A single point.



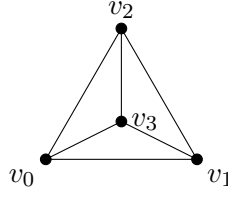
2. A 1-simplex: A line segment connecting two points.



3. A 2-simplex: A triangle with three vertices and the edges between them.



4. A 3-simplex: A tetrahedron with four vertices, six edges, and four triangular faces.



More generally, the *standard n -simplex* Δ^n is defined as

$$\Delta^n = \left\{ (t_0, \dots, t_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^n t_i = 1 \text{ and } t_i \geq 0 \text{ for all } i \right\},$$

where the vertices are given by the unit vectors along each coordinate axis. Any point inside this standard simplex can be represented as a convex combination of the vertices with weights t_i , known as the *barycentric coordinates* of the point.

If we remove one of the $n + 1$ vertices of an n -simplex $[v_0, \dots, v_n]$, the remaining vertices span an $(n - 1)$ -simplex, which is referred to as a *face* of $[v_0, \dots, v_n]$. Each subset of the vertices of an n -simplex thus defines a face of some dimension between 0 and $n - 1$. The collection of all $(n - 1)$ -faces of an n -simplex forms its *boundary*. The *open simplex* $\mathring{\Delta}^n$ is the n -simplex with its boundary removed, representing the interior of the simplex.

Definition 2.3.2. A *simplicial complex* K is a set of simplices such that:

- If a simplex σ is in K , then all faces of σ are also in K .
- The intersection of any two simplices in K is either empty or a shared face.

Informally, a *simplicial complex* is a collection of simplices that are glued together along shared faces, forming a more complex structure. This construction is widely used to approximate topological spaces by breaking them into simpler pieces. For example, the two-dimensional sphere S^2 can be represented by a simplicial complex composed of triangular faces. By breaking down complicated shapes into simpler, combinatorial parts, this view allows for analysis using linear algebra and combinatorics. Moreover, simplicial complexes serve as the foundation for computational topology in data analysis, where they are used to study the shape of data.

Definition 2.3.3. For a simplicial complex K , we define its *dimension* $\dim(K)$ to be the maximum dimension of all simplices present in K . If there is no such maximum, we define $\dim(K) := \infty$.

2.4. Simplicial Homology

Simplicial homology provides a way to study simplicial complexes by examining their properties, focusing on their shape. It assigns algebraic objects to these com-

2. Algebraic Topology

plexes, such as abelian groups or vector spaces, to capture features like connected components, holes, and voids of various dimensions.

Definition 2.4.1. Let $C_n(K)$ denote the group of n -chains, which is the free abelian group generated by the oriented n -simplices in K . An element of $C_n(K)$ has the form

$$c = \sum_i a_i \sigma_i,$$

where $a_i \in \mathbb{Z}$ and σ_i are the oriented n -simplices in K . Here, the choice of orientation affects the signs of the terms in the chain.

For example, a 1-chain in K would be a linear combination of the oriented 1-simplices (edges), and a 2-chain would be a linear combination of the oriented 2-simplices (triangles).

Definition 2.4.2. The boundary operator $\partial_n : C_n(K) \rightarrow C_{n-1}(K)$ is defined on each oriented n -simplex $\sigma = [v_0, v_1, \dots, v_n]$ in a simplicial complex K by the formula:

$$\partial_n(\sigma) = \sum_{i=0}^n (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_n], \quad (2.1)$$

where $[v_0, \dots, \hat{v}_i, \dots, v_n]$ denotes the $(n-1)$ -simplex obtained by removing the vertex v_i from σ .

The sign $(-1)^i$ ensures that the orientations of the faces are consistent with the orientation of the n -simplex. The boundary operator ∂_n on $c \in C_n(K)$ is given by linear extension of Eq.2.1. This defines a sequence of chain groups connected by boundary operators:

$$(C_\bullet(K), \partial_\bullet) := \dots \xrightarrow{\partial_{n+1}} C_n(K) \xrightarrow{\partial_n} C_{n-1}(K) \xrightarrow{\partial_{n-1}} \dots \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} 0,$$

which is known as the *geometric chain complex*.

Example 2.4.1. • For a 1-simplex $[v_0, v_1]$, we have:

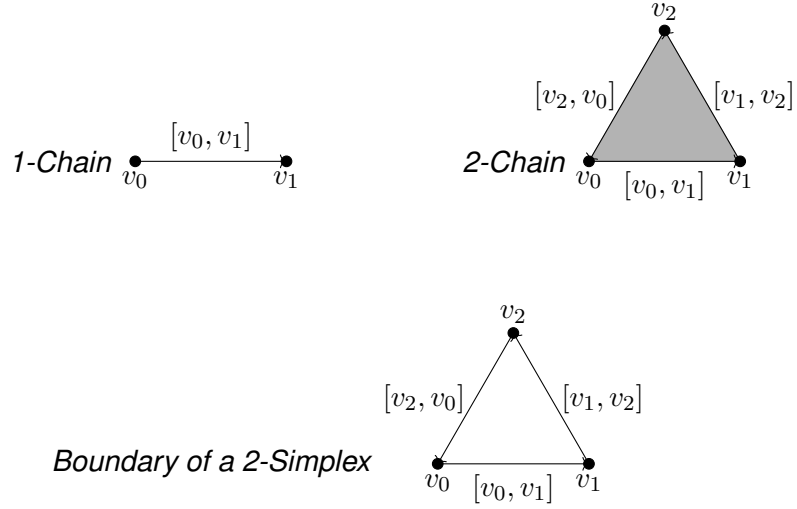
$$\partial_1([v_0, v_1]) = v_1 - v_0.$$

• For a 2-simplex $[v_0, v_1, v_2]$, we have:

$$\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1].$$

In other words, the boundary of a 1-simplex (edge) is the difference between its two endpoints, while the boundary of a 2-simplex (triangle) is the sum of its three oriented edges.

2.4. Simplicial Homology



Example 2.4.2.

In the diagrams above:

- The 1-chain is an edge $[v_0, v_1]$, representing a linear combination of oriented 1-simplices.
- The 2-chain is a triangle $[v_0, v_1, v_2]$, showing a combination of oriented edges and its filled face.
- The boundary of the 2-simplex is the sum of its oriented edges.

The boundary operator satisfies the property that applying it twice yields zero, i.e.,

$$\partial_{n-1} \circ \partial_n = 0.$$

This is because the boundary of any boundary is empty, which can be shown by observing that each $(n-2)$ -face of an n -simplex appears twice in the expression for $\partial_{n-1} \circ \partial_n$, but with opposite signs. We can therefore define the following:

Definition 2.4.3. The n -th homology group $H_n(K)$ is defined as the quotient group

$$H_n(K) = \frac{\ker \partial_n}{\operatorname{im} \partial_{n+1}}.$$

Intuitively, $H_n(K)$ measures the n -dimensional "holes" in K . Elements in $\ker \partial_n$ are called n -cycles, and they represent closed structures, while elements in $\operatorname{im} \partial_{n+1}$ are called n -boundaries and represent boundaries of higher-dimensional simplices.

Proposition 2.4.1. The 0-th simplicial homology group $H_0(K)$ of a connected simplicial complex K is isomorphic to $\mathbb{Z}$.

Proof. The 0-chains $C_0(K)$ are formal sums of the vertices of K with integer coefficients:

$$C_0(K) = \left\{ \sum_i a_i v_i \mid a_i \in \mathbb{Z}, v_i \text{ a vertex of } K \right\}.$$

2. Algebraic Topology

The boundary operator $\partial_0 : C_0(K) \rightarrow 0$ maps every 0-chain to zero, so the kernel of ∂_0 is the entire group $C_0(K)$. Thus, we have:

$$Z_0(K) := \ker(\partial_0) = C_0(K).$$

The group of 0-boundaries $B_0(K) := \text{im}(\partial_1)$ is generated by the boundaries of 1-simplices in K .

The 0-th homology group is defined as the quotient group:

$$H_0(K) = Z_0(K)/B_0(K).$$

This group captures the equivalence classes of 0-chains modulo boundaries. Specifically, two vertices v and w are considered equivalent if there exists a chain of 1-simplices connecting them. Since K is connected, there exists a path (a chain of 1-simplices) between any pair of vertices in K . This implies that all vertices are equivalent under the relation induced by $B_0(K)$, so there is a single equivalence class of 0-chains. We can represent this single class by any vertex $v_0 \in K$.

Consequently,

$$H_0(K) \cong \mathbb{Z}.$$

□

To illustrate the concept of homology, we compute the homology groups of several popular spaces. Since all of the spaces in these examples are connected, Prop. 2.4.1 immediately provides us with the 0-th homology groups. We notice that all of the spaces under consideration admit a triangulation so that the framework of simplicial homology applies in this setting.

Example 2.4.3. For a single point P , the homology groups are:

$$H_0(P) \cong \mathbb{Z}, \quad H_n(P) = 0 \text{ for } n > 0$$

since there are no higher-dimensional chains than in degree 0.

Example 2.4.4. The n -dimensional sphere S^n is the set of all points in $\mathbb{R}^{n+1}$ that are a fixed distance from the origin. It is defined by:

$$S^n = \{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} : x_1^2 + \dots + x_{n+1}^2 = 1\}.$$

The homology groups of S^n read:

$$H_0(S^n) \cong \mathbb{Z}, \quad H_n(S^n) \cong \mathbb{Z}, \quad H_k(S^n) = 0 \text{ for } k \neq 0, n.$$

The $H_n(S^n) = \mathbb{Z}$ group corresponds to the n -dimensional "void" or boundary of the n -sphere. For all other dimensions $k \neq 0, n$, the homology groups vanish.

Example 2.4.5. The homology groups of the 2-dimensional torus $T^2 = S^1 \times S^1$ are:

$$H_0(T^2) \cong \mathbb{Z}, \quad H_1(T^2) \cong \mathbb{Z} \oplus \mathbb{Z}, \quad H_2(T^2) \cong \mathbb{Z}, \quad H_n(T^2) = 0 \text{ for } n > 2.$$

2.4. Simplicial Homology

There are two independent one-dimensional loops (one for each S^1 factor), leading to $H_1(T^2) = \mathbb{Z} \oplus \mathbb{Z}$. The H_2 group detects the two-dimensional "void" enclosed by the torus, so $H_2(T^2) = \mathbb{Z}$. There are no higher-dimensional chains, so $H_n(T^2) = 0$ for $n > 2$.

Proposition 2.4.2. *Let K and L be two simplicial complexes, and let $K \sqcup L$ be their disjoint union. Then*

$$H_n(K \sqcup L) \cong H_n(K) \oplus H_n(L) \quad \text{for all } n \geq 0.$$

Proof. The n -chains of the disjoint union $K \sqcup L$ consist of formal sums of n -simplices from both K and L . Therefore, the chain group $C_n(K \sqcup L)$ can be decomposed as:

$$C_n(K \sqcup L) = C_n(K) \oplus C_n(L).$$

Here, $C_n(K)$ and $C_n(L)$ represent the n -chains of K and L , respectively.

The boundary operator $\partial_n^{K \sqcup L} : C_n(K \sqcup L) \rightarrow C_{n-1}(K \sqcup L)$ acts on n -simplices of K and L independently, since K and L have no simplices in common in their disjoint union. Thus, we have:

$$\partial_n^{K \sqcup L} = \partial_n^K \oplus \partial_n^L.$$

This means that if $\sigma \in C_n(K)$ and $\tau \in C_n(L)$, then:

$$\partial_n^{K \sqcup L}(\sigma, \tau) = (\partial_n^K \sigma, \partial_n^L \tau).$$

In other words, the boundary operator on $K \sqcup L$ applies the boundary operators on K and L separately.

We define the cycle group Z_n and the boundary group B_n as follows:

$$\begin{aligned} Z_n(K) &= \ker(\partial_n^K) \quad \text{and} \quad Z_n(L) = \ker(\partial_n^L), \\ B_n(K) &= \text{im}(\partial_{n+1}^K) \quad \text{and} \quad B_n(L) = \text{im}(\partial_{n+1}^L). \end{aligned}$$

Then, for the disjoint union, the cycle and boundary groups also decompose as

$$Z_n(K \sqcup L) = Z_n(K) \oplus Z_n(L) \quad \text{and} \quad B_n(K \sqcup L) = B_n(K) \oplus B_n(L).$$

The n -th homology group $H_n(K \sqcup L)$ is defined as the quotient of the n -cycles by the n -boundaries:

$$H_n(K \sqcup L) = \frac{Z_n(K \sqcup L)}{B_n(K \sqcup L)}.$$

Since $C_n(K \sqcup L) = C_n(K) \oplus C_n(L)$ and the boundary operator respects this decomposition, it follows that:

$$H_n(K \sqcup L) = \frac{Z_n(K) \oplus Z_n(L)}{B_n(K) \oplus B_n(L)} \cong \frac{Z_n(K)}{B_n(K)} \oplus \frac{Z_n(L)}{B_n(L)} = H_n(K) \oplus H_n(L).$$

We conclude that for any two disjoint simplicial complexes K and L , the homology of their disjoint union is isomorphic to the direct sum of the homology groups of K and L :

$$H_n(K \sqcup L) \cong H_n(K) \oplus H_n(L) \quad \text{for all } n \geq 0.$$

□

2. Algebraic Topology

Definition 2.4.4. Let K and L be simplicial complexes. A simplicial map $f : K \rightarrow L$ is a map such that the image of each simplex under f is a simplex of L .

A simplicial map $f : K \rightarrow L$ induces homomorphisms $f_* : C_n(K) \rightarrow C_n(L)$ for each n , by sending an n -simplex $\sigma \in K$ to $f(\sigma)$ and extending linearly. Since f_* is a chain morphism, it defines homomorphisms $f_* = H_n(f) : H_n(K) \rightarrow H_n(L)$. If $f : K \rightarrow L$ and $g : L \rightarrow M$ are simplicial maps, then $g \circ f$ induces a map $(g \circ f)_* = g_* \circ f_* : H_n(K) \rightarrow H_n(M)$. Moreover, if $f : K \rightarrow L$ is a simplicial isomorphism, then $f_* : H_n(K) \rightarrow H_n(L)$ is an isomorphism for all n .

The latter property is often particularly useful as it provides a criterion to identify non-homeomorphic spaces: if the homology groups of two given spaces do not agree up to isomorphism, the two underlying spaces cannot be homeomorphic. In practice, this is a very powerful observation as the calculation of homology groups is often feasible using combinatorial considerations, while understanding the geometry and topology of the underlying space can be very difficult a priori.

2.5. Relative Homology

Relative simplicial homology extends the concept of simplicial homology to pairs of simplicial complexes. Given a simplicial complex K and a subcomplex $L \subseteq K$, the relative homology $H_n(K, L)$ captures information about the structure of K in relation to L . This allows us to study the features of K that are not present in L , focusing on the parts of K that "lie outside" L .

For a simplicial complex K with a subcomplex $L \subseteq K$, the relative chain group $C_n(K, L)$ is defined as the quotient of the chain groups:

$$C_n(K, L) = C_n(K) / C_n(L),$$

where $C_n(K)$ and $C_n(L)$ are the n -chain groups of K and L , respectively. Elements of $C_n(K, L)$ are called *relative n -chains* and can be viewed as n -chains in K that are allowed to "ignore" any chains that lie entirely in L .

The boundary operator on $C_n(K)$ induces a boundary operator $\partial_n : C_n(K, L) \rightarrow C_{n-1}(K, L)$ on the relative chain groups. This boundary operator is well-defined because the boundary of a chain in L remains in L , ensuring that $\partial_n(C_n(L)) \subseteq C_{n-1}(L)$. Therefore, ∂_n descends to a map on the quotient $C_n(K, L)$. Consequently, $(C_\bullet(K, L), \partial_\bullet)$ forms a chain complex, and we notice that the sequence

$$0 \rightarrow C_\bullet(L) \xrightarrow{i_\bullet} C_\bullet(K) \xrightarrow{j_\bullet} C_\bullet(K, L) \rightarrow 0$$

is exact.

The n -th *relative homology group* $H_n(K, L)$ is then defined as:

$$H_n(K, L) = H_n(C_\bullet(K, L)),$$

and by Cor.2.2.1 we obtain a long exact sequence

$$\cdots \rightarrow H_n(L) \xrightarrow{i_*} H_n(K) \xrightarrow{j_*} H_n(K, L) \xrightarrow{\partial_*} H_{n-1}(L) \rightarrow \cdots \quad (2.2)$$

This latter long exact sequence often enables us to compute homology groups of spaces, given the homology groups of a given subspace or vice versa. This generic property is therefore particularly useful for practical considerations.

2.6. Local Homology

Local homology provides a tool for analyzing the structure of a topological space "locally" around a fixed point. For a simplicial complex K , local homology can give insight into how K behaves in the neighborhood of a chosen vertex, detecting the presence of "holes" or other topological features nearby.

Let K be a simplicial complex, and let v be a vertex in K . To define the local homology group of K at v , we consider neighborhoods of v that capture the local structure of K around v . In order to do so, we first need to introduce the notion of stars and links.

Definition 2.6.1. • $\text{St}(v)$ is the star of v , consisting of all simplices in K that have v as a vertex.

• $\text{Lk}(v)$ is the link of v , consisting of simplices in $\text{St}(v)$ that do not contain v itself.

Intuitively, the star of v represents the "immediate neighborhood" around v , capturing all simplices connected to v in K , while the link of v encodes the boundary of this neighborhood, formed by the simplices that intersect the star of v but do not include v itself.

Definition 2.6.2. The local homology group of K at a vertex v , in degree n , is defined as

$$H_n(K, K - v) := H_n(\text{St}(v), \text{Lk}(v)).$$

Local homology is a useful concept in studying topological invariants and properties of a space at a granular level, as it allows for the detection of localized features that are not necessarily apparent in the global homology of K . For instance, local homology can distinguish points in K that contribute to different topological structures, such as boundaries, isolated points, or singularities.

2.7. Manifolds

In mathematics and physics, *manifolds* serve as one of the most versatile and profound structures for modeling both abstract mathematical spaces and physical

2. Algebraic Topology

phenomena. An n -dimensional manifold is a topological space that locally resembles Euclidean space $\mathbb{R}^n$, but can have a more complex global structure. These spaces have become indispensable across various fields due to their flexibility and rich structure.

Definition 2.7.1. *Let M be a simplicial complex. We say M is an n -dimensional simplicial manifold if, for every vertex v in M , $\text{Lk}(v)$ is an $(n - 1)$ -dimensional simplicial complex that is homeomorphic to an $(n - 1)$ -dimensional sphere S^{n-1} .*

We now provide some common examples of manifolds, ranging from basic shapes to more complex structures.

Example 2.7.1. S^n is an n -dimensional manifold, as it locally resembles $\mathbb{R}^n$ at each point.

Example 2.7.2. The torus T^2 is a surface that can be visualized as the shape of a donut. It can be defined as the product of two circles:

$$T^2 = S^1 \times S^1.$$

This is a 2-dimensional manifold that locally resembles $\mathbb{R}^2$, but its global structure has a more complex topology, featuring two independent cycles. More generally, the product of any two manifolds of dimension m and n is a manifold of dimension $m + n$.

Example 2.7.3. The real projective plane $\mathbb{RP}^2$ is a non-orientable surface that can be thought of as the set of all lines passing through the origin in $\mathbb{R}^3$. We define:

$$\mathbb{RP}^2 = \{\text{lines in } \mathbb{R}^3 \text{ through the origin}\}.$$

It is a 2-dimensional manifold that can also be visualized as a sphere with antipodal points (points directly opposite each other) identified.

Example 2.7.4. The Klein bottle is a non-orientable surface with no distinct "inside" or "outside." It can be constructed by identifying opposite edges of a rectangular region in $\mathbb{R}^2$, with a twist along one pair of edges. Unlike the torus, it cannot be embedded in $\mathbb{R}^3$ without self-intersection. The Klein bottle is a 2-dimensional manifold.

Example 2.7.5. The Euclidean space $\mathbb{R}^n$ is the set of all n -tuples of real numbers:

$$\mathbb{R}^n = \{(x_1, x_2, \dots, x_n) : x_i \in \mathbb{R} \text{ for all } i = 1, \dots, n\}.$$

$\mathbb{R}^n$ is the simplest example of an n -dimensional manifold, as it is globally homeomorphic to Euclidean space.

2.7.1. Local homology of manifolds

The computation of the local homology groups $H_k(M, M - v)$ for a simplicial n -manifold M and a vertex $v \in M$ can be done using the long exact homology sequence of the pair $(St(v), Lk(v))$ and we obtain:

- For $k = n$, the local homology group $H_k(M, M - v)$ is isomorphic to $\mathbb{Z}$, which captures the fact that M locally looks like $\mathbb{R}^n$ around v and has a fundamental class in dimension n .
- For $k \neq n$, the local homology groups $H_k(M, M - v)$ are zero, indicating that there are no lower-dimensional "holes" in the neighborhood of v .

Therefore, we conclude that for a simplicial n -manifold M , the local homology groups at any vertex v are given by:

$$H_k(M, M - v) \cong \begin{cases} \mathbb{Z} & \text{if } k = n, \\ 0 & \text{otherwise.} \end{cases} \quad (2.3)$$

We previously observed that the global homology of two manifolds may differ (as an example take S^2 and T^2), even when their dimension agrees. However, the local homology of a manifold is already determined by its dimension (see Eq.2.3). The local homology thus provides a robust way to confirm the manifold structure at each point in the simplicial complex, and it reflects the fundamental class in top dimension only.

2.8. Stratified Spaces

In the study of spaces with singularities, stratified pseudomanifolds provide a natural framework to capture and formalize these singular structures. This class of spaces generalizes the concept of a manifold by allowing for the controlled presence of singularities, which can be managed through a filtration by manifolds of various dimensions. This notion is particularly well-suited to the context of simplicial complexes, where stratification can be defined in terms of simplices of different dimensions.

Definition 2.8.1. A stratification of a simplicial complex K is a filtration

$$K_0 \subset K_1 \subset \cdots \subset K_n = K$$

where each K_i is a subcomplex, and each difference $K_i \setminus K_{i-1}$ is a locally closed subset of K , called the i -th stratum. Each stratum $K_i \setminus K_{i-1}$ is a manifold of dimension i .

Definition 2.8.2. A 0-dimensional stratified (simplicial) pseudomanifold is a countable set of points with the discrete topology. A simplicial complex X of dimension n is called an n -dimensional stratified (simplicial) pseudomanifold if it has a filtration

$$\emptyset = X_{-1} \subset X_0 \subset X_1 \subset \cdots \subset X_{n-1} \subset X_n = X,$$

where each $X_i \setminus X_{i-1}$ is an i -dimensional simplicial manifold (possibly with boundary), called the i -th stratum of X , and:

1. The top stratum $X_n \setminus X_{n-1}$ is dense in X and is an n -dimensional manifold.

2. Algebraic Topology

2. For each $i < n$, the stratum $X_i \setminus X_{i-1}$ has a neighborhood that is homeomorphic to a product of the form $\mathbb{R}^i \times C(L)$, where $C(L)$ denotes the cone on a lower-dimensional stratified (simplicial) pseudomanifold L .

The filtration and local product structure of stratified pseudomanifolds make them well-suited for studying spaces with controlled singularities, where each stratum locally resembles a manifold. In particular, stratified pseudomanifolds enable the extension of manifold-based tools and concepts to settings where singularities naturally occur, such as the study of topological invariants of singular spaces [8].

2.9. Euler Characteristic

The *Euler characteristic* is a fundamental topological invariant that provides a global measure of a topological space's structure. In the context of simplicial complexes, the Euler characteristic is particularly useful for analyzing the combinatorial structure of the complex. Defined by counting vertices, edges, faces, and higher-dimensional simplices, the Euler characteristic connects geometric data to global topological properties and is preserved under topological transformations.

Definition 2.9.1. Let K be a finite simplicial complex. For each n -dimensional simplex in K , let f_n denote the number of n -simplices in K . The Euler characteristic $\chi(K)$ of K is defined as:

$$\chi(K) = \sum_{n=0}^{\infty} (-1)^n f_n.$$

Intuitively, the Euler characteristic sums the number of n -dimensional simplices with alternating signs, thus taking into account the structure of the simplicial complex across different dimensions. For a simple complex, such as a polygonal shape in $\mathbb{R}^2$, this sum reduces to counting vertices minus edges plus faces, often yielding intuitive results.

The Euler characteristic is central in topology, as it captures important global information. It also has significant applications in algebraic topology, where it is expressed in terms of homology groups. Specifically, if K is a simplicial complex, then the Euler characteristic can also be defined using the ranks of its homology groups $H_n(K)$, as follows:

$$\chi(K) = \sum_{n=0}^{\infty} (-1)^n \text{rank } H_n(K). \quad (2.4)$$

This homological formulation of the Euler characteristic provides a powerful tool for understanding topological invariants in more abstract spaces. Since homology groups are themselves topological invariants, the Euler characteristic inherits this property, ensuring its stability under homeomorphisms (and even homotopy-equivalences). Moreover, this formulation generalizes well beyond simplicial com-

plexes to other topological spaces where homology groups are defined, thereby creating a bridge between combinatorial and algebraic approaches to topology.

Example 2.9.1. Consider a simple example, the boundary of a triangle. This boundary is a simplicial complex K with three vertices, three edges, and no faces. Thus, we have $f_0 = 3$, $f_1 = 3$, and $f_2 = 0$. The Euler characteristic is calculated as:

$$\chi(K) = f_0 - f_1 + f_2 = 3 - 3 + 0 = 0.$$

This result is consistent with the fact that the boundary of a triangle is topologically equivalent to a circle, which has an Euler characteristic of zero.

The Euler characteristic is invariant under homeomorphisms, which makes it a useful tool for distinguishing topological spaces up to deformation. For example, surfaces such as the torus, the sphere, and the projective plane all have distinct Euler characteristics, reflecting their differing topological structures. This makes the Euler characteristic a valuable tool in classifying surfaces and higher-dimensional complexes.

In summary, the Euler characteristic offers a combinatorial approach to capturing topological information, particularly when applied to simplicial complexes. Whether calculated through counting simplices or homology ranks, it remains one of the most accessible and insightful invariants in topology, bridging geometry, combinatorics, and algebra.

2.10. Local-Global Principles in Topology

In topology, the study of properties and structures of spaces often benefits from an interplay between local and global perspectives. Local-global principles are central to this approach, enabling us to deduce global properties of a topological space from local information and vice versa. This framework is not only foundational in topology but also finds applications across a wide range of mathematical fields, including geometry, analysis and number theory.

A fundamental example of a local-global principle is the classification of manifolds. While a manifold is defined as a space that locally resembles Euclidean space, its global structure can exhibit complex properties, such as different shapes that do not appear locally. Understanding the global characteristics of a manifold often involves piecing together information from these local Euclidean structures.

The local-global principle can also be observed in the study of covering spaces and fundamental groups. By understanding the local properties of covering maps, one can infer global information about the fundamental group of a space, which in turn reveals significant topological characteristics. For instance, connectedness and simple connectedness, which are global properties, can be analyzed through the existence of universal covers and the behavior of loops in local neighborhoods.

Another significant example of local-global principles is found in cohomology theories,

2. Algebraic Topology

which provide a bridge between local and global properties of spaces. Cohomology assigns algebraic invariants to spaces, and these invariants can be computed either locally (for instance, on open sets in a covering) or globally for the entire space. The interplay between local and global cohomology is especially crucial in the context of sheaf cohomology, where local sections and the process of "gluing" across overlapping open sets leads to global sections that reveal global topological information.

Definition 2.10.1. *Given a topological space X and an open cover $\{U_\alpha\}$ of X , a sheaf on X is an assignment of data to each open set in X , subject to two conditions:*

- *For each inclusion of open sets $V \subset U$, there is a restriction map from the data on U to the data on V .*
- *If $U = \bigcup_\alpha U_\alpha$, and data are given for each U_α that agree on overlaps, then there exists a unique extension to all of U .*

The notion of local-global principles also appears in Morse theory, where one studies smooth functions on a manifold to deduce its topology. Morse functions allow us to understand the global topology of a manifold by examining critical points and their indices. As a result, local critical point analysis yields insights into the global structure of the manifold, such as its homology.

Overall, local-global principles in topology serve as powerful tools to bridge the gap between local and global perspectives. By relating local properties to global structures, these principles offer a deeper understanding of the underlying object of interest, enabling mathematicians to make significant connections across various disciplines.

3. Algebraic Topology and Data

3.1. Persistent Homology

Persistent homology is a technique from topological data analysis that extends traditional homology to study the shapes and topological features of data across multiple scales. In many applications, data is represented in a metric space where topological features emerge depending on the scale at which the data is viewed. Persistent homology captures these features by examining how homology groups change as the scale parameter varies, creating a multi-scale summary of the data's topological structure.

Definition 3.1.1. A filtration of a topological space X is a nested sequence of subspaces indexed by a totally ordered set A . That is, we have

$$\{X_\alpha\}_{\alpha \in A}, \quad \text{where } X_\alpha \subseteq X_\beta \quad \text{for all } \alpha \leq \beta \text{ in } A.$$

If $A = \mathbb{R}_{\geq 0}$, the filtration provides a continuous view of X across a real-valued scale, allowing the study of topological features as they evolve smoothly. In other cases, A may be a discrete set or any other totally ordered set, adapting the filtration to specific contexts and applications.

We now explain the concept of persistent homology, which is generic in the sense that it applies to any filtration of simplicial complexes and does not depend on a specific filtration.

Definition 3.1.2. Let $\{X_\alpha\}_{\alpha \in A}$ be a filtration of simplicial complexes indexed by a totally ordered set A , typically $A = \mathbb{R}_{\geq 0}$. For each $n \geq 0$ and $\alpha \leq \beta$ in A , the inclusion $X_\alpha \hookrightarrow X_\beta$ induces a map on homology, $f_{\alpha,\beta}^n : H_n(X_\alpha) \rightarrow H_n(X_\beta)$. The collection of homology groups $\{H_n(X_\alpha)\}_{\alpha \in A}$, together with the maps $f_{\alpha,\beta}^n$ for each $\alpha \leq \beta$, is called the n -th persistent homology of the filtration $\{X_\alpha\}_{\alpha \in A}$.

The persistent homology captures the n -dimensional topological features of X that are not only present but also persist across multiple scales α in the filtration. Each feature in has a birth scale α (meaning it first appears in $H_n(X_\alpha)$) and may have a death scale β at which it vanishes in $H_n(X_\beta)$. Features that persist over a large range of α -values (i.e., have long lifetimes) are often interpreted as significant or robust structural properties of X , while features that only appear over a short range are typically considered noise or artifacts of the data or scale.

3.2. Decomposition of Persistence Modules

In the study of persistent homology, so-called persistence modules play a crucial role in understanding the evolution of homological features across scales. A key result in this context is the decomposition theorem, which provides a complete classification of these modules through interval modules.

Definition 3.2.1. Let $\mathbb{R}$ denote the real numbers, considered as an ordered set. A persistence module V is a family of vector spaces $\{V_t\}_{t \in \mathbb{R}}$ indexed by $t \in \mathbb{R}$, along with a collection of linear maps $\varphi_{s \leq t} : V_s \rightarrow V_t$ for $s \leq t$, called structure maps, such that:

$$\varphi_{t \leq t} = \text{id}_{V_t} \quad \text{and} \quad \varphi_{r \leq t} = \varphi_{s \leq t} \circ \varphi_{r \leq s} \quad \text{for all } r \leq s \leq t.$$

Definition 3.2.2. An interval module is a persistence module that is concentrated on a single interval $I = [a, b) \subset \mathbb{R} \cup \{\infty\}$, where $a \leq b$. Formally, for an interval $I = [a, b)$, the interval module $\mathbb{I}(I)$ is defined by:

$$\mathbb{I}(I)_t = \begin{cases} \mathbb{K} & \text{if } t \in [a, b), \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbb{K}$ is the base field (e.g., $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$), and the structure maps $\varphi_{s \leq t}$ are the identity maps if $s, t \in [a, b)$ and zero otherwise.

The decomposition theorem for persistence modules states that every one-dimensional persistence module can be decomposed as a direct sum of interval modules.

Theorem 3.2.1 (Decomposition Theorem for One-Dimensional Persistence Modules). Let V be a persistence module over $\mathbb{R}$. Then there exists a unique (up to isomorphism) decomposition

$$V \cong \bigoplus_{i=1}^n \mathbb{I}(I_i)$$

where I_i are intervals in $\mathbb{R}$ and $\mathbb{I}(I_i)$ are the corresponding interval modules. Each I_i is called a persistence interval and represents a homological feature of V that persists over I_i .

3.3. Persistence Diagrams

The decomposition theorem implies that one-dimensional persistence modules can be completely described by their persistence intervals. This result is fundamental in persistent homology, as it allows for the characterization of topological features in terms of birth and death times, represented by the endpoints of each interval I_i . The collection of intervals $\{I_i\}$ is commonly visualized using a *persistence diagram*.

Definition 3.3.1. For each homology class that appears at scale $\alpha = \alpha_{\text{birth}}$ and disappears at scale $\alpha = \alpha_{\text{death}}$, we define its persistence as the interval $[\alpha_{\text{birth}}, \alpha_{\text{death}})$. This interval is often represented as a point $(\alpha_{\text{birth}}, \alpha_{\text{death}})$ in a two-dimensional plot, known as a persistence diagram.

The persistence diagram provides a concise summary of topological features across scales, with each point in the diagram representing a homology class and its lifespan. Features that persist over a wide range of scales are typically considered more significant, while those that appear only for a narrow range of scales may correspond to noise.

3.4. Vietoris-Rips and Čech Complexes

For data analysis, we often start with a finite set of points in a metric space and define a sequence of simplicial complexes that approximate the data at different scales. By constructing these complexes at varying thresholds, we obtain a filtration for which we can calculate its persistent homology. We now give two famous examples of such constructions that we will need in what follows.

Definition 3.4.1. *Let X be a set of points in a metric space with distance function d , and let $\epsilon > 0$ be a real number. The Vietoris-Rips complex $\mathcal{V}_\epsilon(X)$ is the simplicial complex with vertex set X such that a finite subset $\{x_0, x_1, \dots, x_k\} \subseteq X$ spans a k -simplex if and only if $d(x_i, x_j) \leq \epsilon$ for all $0 \leq i < j \leq k$.*

In other words, a simplex in the Vietoris-Rips complex exists when all pairs of points in that subset are within a distance ϵ of each other. This complex captures connectivity information in the data and provides a computationally efficient approximation of the underlying topology. One of the main advantages of the Vietoris-Rips complex is that it can be constructed from pairwise distances, making it particularly suitable for large datasets.

Definition 3.4.2. *Let X be a set of points in a metric space with distance function d , and let $\epsilon > 0$ be a real number. The Čech complex $\mathcal{C}_\epsilon(X)$ is the simplicial complex with vertex set X such that a finite subset $\sigma = \{x_0, x_1, \dots, x_k\} \subseteq X$ spans a k -simplex if and only if the set of ϵ -balls centered at the vertices of σ has a nonempty intersection.*

3.5. The Nerve Theorem

We now formulate the well-known Nerve Theorem, which is a foundational result in algebraic topology that relates the topology of a space to the topology of a combinatorial structure known as the *nerve* of a covering. This theorem provides a powerful tool for studying complex spaces by breaking them down into simpler, overlapping pieces and analyzing the relationships between these pieces. We start with the definition of a nerve of a covering.

Definition 3.5.1. *Let X be a topological space, and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be an open covering of X ; that is, $X = \bigcup_{\alpha \in A} U_\alpha$ with each $U_\alpha \subseteq X$ open. The nerve of the covering $\mathcal{U}$, denoted $N(\mathcal{U})$, is an abstract simplicial complex defined as follows:*

3. Algebraic Topology and Data

- The vertices of $N(\mathcal{U})$ correspond to the elements of $\mathcal{U}$.
- A finite subset $\{U_{\alpha_0}, U_{\alpha_1}, \dots, U_{\alpha_k}\} \subseteq \mathcal{U}$ spans a k -simplex in $N(\mathcal{U})$ if and only if the intersection $U_{\alpha_0} \cap U_{\alpha_1} \cap \dots \cap U_{\alpha_k} \neq \emptyset$.

Theorem 3.5.1 (Nerve Theorem). *Let X be a topological space, and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be an open covering of X . Suppose that for every finite nonempty intersection $U_{\alpha_0} \cap U_{\alpha_1} \cap \dots \cap U_{\alpha_k}$ of elements of $\mathcal{U}$, the intersection is either empty or contractible. Then the nerve $N(\mathcal{U})$ is homotopy equivalent to X ; that is,*

$$N(\mathcal{U}) \simeq X.$$

The condition that each nonempty intersection $U_{\alpha_0} \cap U_{\alpha_1} \cap \dots \cap U_{\alpha_k}$ is contractible ensures that each piece of the covering is simple enough in terms of its topological structure. When this condition holds, the nerve complex provides an accurate combinatorial model of the topological space. The Nerve Theorem serves as the theoretical foundation for the Čech complex since it ensures that for a discrete set of points X , the Čech complex $\mathcal{C}_\epsilon(X)$ is homotopy-equivalent to the union of balls with radius ϵ , centered at the points in X . Therefore, in order to calculate the homology of the union of balls with radius ϵ centered at the points in X (which we can view as an approximation of X itself, controlled by a thickening parameter ϵ), we can simply calculate the homology of $\mathcal{C}_\epsilon(X)$ which can be achieved combinatorially.

As we vary ϵ , we obtain a filtration of simplicial complexes for both the Vietoris-Rips and Čech constructions, leading to multi-scale insights into the data through persistent homology.

Definition 3.5.2. *Let X be a set of points in a metric space, and let $\epsilon \geq 0$ be a scale parameter. The n -th Vietoris-Rips persistent homology of X , $\text{PH}_n(\mathcal{V}(X))$, is the n -th persistent homology induced by the filtration $\{\mathcal{V}(X, \epsilon)\}_{\epsilon \geq 0}$, where $\mathcal{V}(X, \epsilon)$ is the Vietoris-Rips complex at scale ϵ .*

Definition 3.5.3. *Similarly, the n -th Čech persistent homology of X , $\text{PH}_n(\mathcal{C}(X))$, is the n -th persistent homology induced by the filtration $\{\mathcal{C}(X, \epsilon)\}_{\epsilon \geq 0}$, where $\mathcal{C}(X, \epsilon)$ denotes the Čech complex at scale ϵ .*

The persistence of features in either the Vietoris-Rips or Čech persistent homology provides insight into the underlying shape and structure of X . Features that persist over a long range of ϵ -values are typically considered to reflect significant topological information, while those that appear and disappear quickly are often treated as noise. Persistent homology, therefore, enables a multi-scale analysis of topological features in data. As we saw earlier, it is well justified to use the Čech complex as an approximation for the underlying discrete data we start with. However, in practice one often prefers to work with the Vietoris-Rips complex instead since it is more efficiently computable. The following theorem relates the two complexes, and therefore provides a justification to use the Vietoris-Rips construction.

Theorem 3.5.2. *Let X be a set of points in a metric space, and let $\mathcal{V}(X, \epsilon)$ and $\mathcal{C}(X, \epsilon)$ denote the Vietoris-Rips and Čech complexes on X at scale ϵ . Then for any*

3.6. Wasserstein and Bottleneck Distances

$\epsilon \geq 0$, we have the inclusion

$$\mathcal{C}(X, \epsilon) \subseteq \mathcal{V}(X, \sqrt{2}\epsilon) \subseteq \mathcal{C}(X, \sqrt{2}\epsilon).$$

This inclusion establishes a connection between Vietoris-Rips and Čech persistent homology, as it implies that the homology groups $H_n(\mathcal{V}(X, \epsilon))$ approximate $H_n(\mathcal{C}(X, \epsilon))$ up to a factor of $\sqrt{2}$ in the scale parameter ϵ .

3.6. Wasserstein and Bottleneck Distances

In persistent homology, distances between persistence diagrams provide a means to quantify the similarity of topological features across different datasets. Two commonly used metrics are the *Wasserstein distance* and the *bottleneck distance*. These metrics enable us to formalize the stability of persistent homology under perturbations of the underlying data, a property that is crucial for applications in data analysis.

Definition 3.6.1. Let D_1 and D_2 be two persistence diagrams. For a real number $p \geq 1$, the Wasserstein distance of order p between D_1 and D_2 is defined by

$$W_p(D_1, D_2) = \left(\inf_{\gamma: D_1 \rightarrow D_2} \sum_{x \in D_1} \|x - \gamma(x)\|^p \right)^{1/p},$$

where the infimum is taken over all bijections γ between the points of D_1 and D_2 , and $\|\cdot\|$ denotes the L^∞ -norm on the persistence plane.

Definition 3.6.2. The bottleneck distance between two persistence diagrams D_1 and D_2 is given by

$$d_B(D_1, D_2) = \inf_{\gamma: D_1 \rightarrow D_2} \sup_{x \in D_1} \|x - \gamma(x)\|,$$

where the infimum is again taken over all bijections γ between the points of D_1 and D_2 . The bottleneck distance can be viewed as the L^∞ -version of the Wasserstein distance.

3.7. Stability of Persistent Homology

In order to acknowledge that persistent homology is a robust tool, it is crucial that the persistent homology of a space is *stable* with respect to small changes in the underlying data. This stability guarantees that similar datasets yield similar persistence diagrams, ensuring reliable interpretations even in the presence of noise or measurement error.

3. Algebraic Topology and Data

Definition 3.7.1. Let X and Y be two non-empty subsets of a metric space (M, d) . The Hausdorff distance $d_H(X, Y)$ between X and Y is defined by

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}.$$

The Hausdorff distance measures how far each point in one set is from the closest point in the other set, capturing the maximum discrepancy between the two sets.

Theorem 3.7.1 (Stability of Vietoris-Rips Persistent Homology). Let X and Y be two point clouds in a metric space, and let $d_H(X, Y)$ denote the Hausdorff distance between X and Y . Then, for each dimension n and any $p \geq 1$, the Vietoris-Rips persistence diagrams $D_n(X)$ and $D_n(Y)$ satisfy

$$d_B(D_n(X), D_n(Y)) \leq 2 d_H(X, Y).$$

This result is known as the stability theorem for Vietoris-Rips persistent homology. It implies that small changes in the data result in small changes in the persistence diagrams, ensuring the robustness of persistent homology with respect to noise.

In practice, persistent homology has been highly effective in analyzing complex data sets, providing insights into the shape and structure of data that can be visualized through persistence diagrams. Persistent homology is widely applied in various fields, from shape recognition in computer vision to detecting structural patterns in biological data. Persistent homology provides a robust framework for understanding data, allowing for the extraction of multi-scale topological information and enabling comparisons between different data sets by examining their topological features. Through persistence diagrams, it bridges combinatorial topology and data analysis, offering a mathematically rigorous approach to understanding the "shape" of data across scales.

3.8. Multiparameter Persistence

In many applications, topological data analysis (TDA) requires the analysis of shapes or spaces across multiple parameters, capturing more complex structures than single-parameter persistence can convey. Multiparameter persistence extends the notion of persistent homology to functions that depend on several parameters simultaneously, providing a richer and more nuanced understanding of the data.

Definition 3.8.1. Let X be a topological space and let $f : X \rightarrow \mathbb{R}^k$ be a continuous function, where $k \geq 2$. For each vector $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k$, define a sublevel set $X_\alpha = f^{-1}((-\infty, \alpha_1] \times \dots \times (-\infty, \alpha_k])$. The collection of these sublevel sets $\{X_\alpha\}_{\alpha \in \mathbb{R}^k}$ forms a multiparameter filtration of X , with a partial ordering defined by $\alpha \leq \beta$ if $\alpha_i \leq \beta_i$ for all $i = 1, \dots, k$.

Definition 3.8.2. The multiparameter persistent homology of a multiparameter filtration $\{X_\alpha\}_{\alpha \in \mathbb{R}^k}$ is the collection of homology groups $H_n(X_\alpha)$ computed at each

3.9. Euler Characteristic Transforms

parameter α , along with the homomorphisms $\phi_{\alpha \leq \beta} : H_n(X_\alpha) \rightarrow H_n(X_\beta)$ induced by the inclusions $X_\alpha \subseteq X_\beta$ whenever $\alpha \leq \beta$.

Unlike the single-parameter setting, multiparameter persistence does not yield a straightforward barcode or persistence diagram representation. The structure of multiparameter persistence modules is more complex and can be visualized as a grid or network of homology groups connected by maps, depending on the partial order.

Definition 3.8.3. A multiparameter persistence module is a family of vector spaces $\{V_\alpha\}_{\alpha \in \mathbb{R}^k}$ indexed by $\mathbb{R}^k$ and connected by linear maps $\phi_{\alpha \leq \beta} : V_\alpha \rightarrow V_\beta$ for all $\alpha \leq \beta$. A multiparameter persistence module is said to be finitely generated if there exists a finite set of generators that span all vector spaces V_α under the action of the linear maps.

One challenge of multiparameter persistence is that the structure of a multiparameter persistence module is generally more intricate and not as easily decomposed as in the single-parameter case. While a single-parameter persistence module can often be decomposed into a sum of interval modules, such a decomposition is not typically possible in the multiparameter setting. This distinction introduces new mathematical complexities and requires novel methods for analysis and computation. Multiparameter persistence modules thus provide a way to study data across multiple scales and parameters simultaneously. Although these modules present significant theoretical and computational challenges, they also offer deeper insights into the structure and features of high-dimensional data.

3.9. Euler Characteristic Transforms

The Euler characteristic transform (ECT) is a method in topological data analysis used to study and classify shapes based on the Euler characteristic. The ECT enables shape characterization through multi-directional views, leveraging both geometric and topological features.

Definition 3.9.1. Let $X \subset \mathbb{R}^n$ be a compact shape. The Euler characteristic transform (ECT) is defined by integrating the Euler characteristic of X over families of sublevel sets in different directions. For a given unit vector $u \in S^{n-1}$ and a real parameter t , the slice of X in direction u at height t is the sublevel set

$$X_{u,t} = \{x \in X \mid \langle x, u \rangle \leq t\},$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^n$.

Definition 3.9.2. The Euler characteristic transform of X is then the function

$$\text{ECT}(X) : S^{n-1} \times \mathbb{R} \rightarrow \mathbb{Z} \tag{3.1}$$

$$(u, t) \mapsto \chi(X_{u,t}). \tag{3.2}$$

3. Algebraic Topology and Data

The function $\text{ECT}(X)$ encodes the Euler characteristic of the sublevel sets of X across different directions u and thresholds t , thereby capturing multi-directional information about X .

The Euler characteristic transform provides a topological signature of X by capturing the evolution of topological features across sublevel sets in various directions. This makes it useful for distinguishing between different shapes based on topological and geometrical properties.

Theorem 3.9.1. *Let $X \subset \mathbb{R}^n$ be a constructible subset. The Euler Characteristic Transform (ECT) map $\text{ECT}(X)$ uniquely determines X up to a constructible subset equivalence. In other words, if two constructible subsets $X, Y \subset \mathbb{R}^n$ have the same ECT, then they are equivalent as constructible sets.*

The Euler characteristic transform is therefore particularly useful in applications involving shape recognition, classification, and comparison in various fields. By using the ECT, one can represent and analyze shapes in a way that is robust to deformations and noise, while capturing essential characteristics of the underlying object.

Part I.

**Topological Representation
Learning**

4. Local Topology and Singularities

Topology, as a field, has long been concerned with understanding spaces through their intrinsic shape properties, often employing techniques that either focus on local structures (e.g., local homology) or global invariants (e.g., homology, homotopy groups etc.). The primary challenge in many applications, particularly in data science and machine learning, is how to reconcile these perspectives into a coherent analytical framework. This chapter provides a local perspective, establishing how singularities manifest as disruptions in the local structure of spaces.

The study of singularities in data and their impact on representation learning highlights a crucial distinction between purely manifold-based assumptions and more general topological structures. The methods introduced in this chapter serve as essential tools for capturing deviations from Euclidean behavior at a local scale.

4.1. Introduction

The field of representation learning has grown rapidly over the last years, with various methodologies proposed to address a plethora of problems.

Many of these methods are grounded in a common assumption: the data lies close to a manifold with a low intrinsic dimension. More formally, while the input data may reside in a high-dimensional ambient space of dimension N , it is assumed that there exists an n -dimensional manifold (with $n \ll N$) that best captures the underlying structure of the data. This assumption, known as the *manifold hypothesis*, is well-suited for certain types of data, such as natural images, where local neighborhoods exhibit consistency in structure [30].

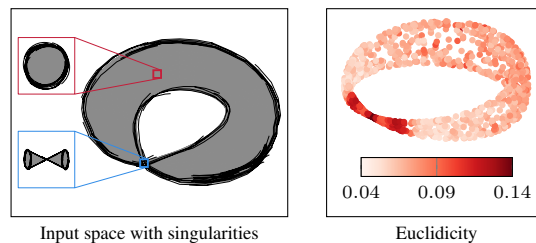


Figure 4.1.: Overview of our method. Given a space with *singularities*, *Euclidity* measures the deviation from a space to a Euclidean model space. Here, *Euclidity* highlights the singularity at the ‘pinch point.’ This figure was recreated from Von Rohrscheidt and Rieck [193].

4. Local Topology and Singularities

Despite the successes of manifold-based methods, the manifold hypothesis is increasingly recognized as a limiting assumption, particularly when dealing with complex data sets from domains like genomics, social networks, or medical imaging [23]. For such data, there is substantial evidence that singularities—points or regions in the data where the manifold structure fails—are not mere artifacts or noise. Instead, these singularities often carry important information about the underlying generative processes and structural nuances within the data. Examples of singular structures in data include intersections, folds, or “pinch points,” where the smoothness of the manifold is interrupted.

A classic illustration of a singularity is the *pinched torus*, where a specific region of a torus, typically a meridian, is compressed to a single point, resulting in a *pinch point* that disrupts the manifold structure. At this pinch point, the local neighborhood is no longer homeomorphic to Euclidean space $\mathbb{R}^2$, even though other regions of the torus retain this property (see Fig. 4.1 for an illustration). Singularities like these violate the basic assumptions of manifold learning methods, which rely on locally uniform structures to capture the intrinsic geometry of data. Consequently, traditional techniques based on the manifold hypothesis often fail to detect or adapt to such non-manifold regions [151, 167].

Singularities pose unique challenges for manifold learning, as they introduce regions where the intrinsic dimensionality of the data varies discontinuously. Standard manifold learning algorithms, such as Isomap, t-SNE, or UMAP, assume a global manifold and therefore struggle with such local inconsistencies. While singularities might initially appear as anomalies or noise, they frequently encode essential structural information [94]. For instance, in biological data, a singularity may represent a point of cellular differentiation or a branching event in evolutionary processes, providing valuable insights into underlying biological mechanisms. Therefore, detecting and analyzing singular regions is crucial for a comprehensive understanding of data, especially in complex domains where such structures are expected to arise.

To address the limitations of manifold-based methods in the presence of singularities, we propose an *unsupervised representation learning framework for detecting singular regions in point cloud data*. Our framework is designed to be agnostic to specific geometric or stochastic properties of the underlying data and requires only a notion of intrinsic dimensionality within neighborhoods, which can vary across different points. This flexibility enables our method to capture non-manifold regions accurately, offering a more robust alternative to traditional manifold learning techniques.

At the core of our approach is a novel formulation of *persistent local homology* (PLH), a method that assesses neighborhood shapes at multiple scales of locality. PLH enables the detection of singular regions by examining changes in the topological structure of local neighborhoods, which is often a direct indication of singularity presence. Using PLH, we estimate the intrinsic dimension of a point locally, capturing variations in the dimensionality that signify departures from manifold structure of the underlying dataset, and define *Euclidicity*, a novel measure that quantifies the deviation of a point in a dataset from being Euclidean, effectively highlighting singular regions.

The *Euclidicity* measure provides a complementary perspective to traditional manifold learning by highlighting areas where the manifold hypothesis does not hold, thus allowing for a more nuanced understanding of the data structure. We establish theoretical guarantees on the approximation quality of this approach for certain classes of spaces, including manifolds and spaces with isolated singularities. Furthermore, through experiments on various data sets, from synthetic spaces with known singularities to high-dimensional image data, we demonstrate the effectiveness of this perspective in revealing the underlying structure of complex data. By focusing on singular regions, our framework opens new avenues for understanding complex datasets where traditional manifold-based techniques fall short. The results of this chapter are essentially based on [193].

4.2. Related Work

Manifold learning traditionally focuses on developing algorithms that extract geometric and topological information under the assumption that data lie on or near a low-dimensional manifold. This assumption, known as the *manifold hypothesis*, has served as a foundational concept in unsupervised and semi-supervised learning for tasks where data are expected to exhibit a locally uniform structure. However, recent work has started to challenge this assumption, exploring cases where data may not conform to a single manifold.

One significant extension of the manifold hypothesis is the *union of manifolds hypothesis*, introduced by Brown et al. [23]. This hypothesis generalizes the traditional manifold assumption to data spaces that can be modeled as disjoint unions of manifolds, allowing for variation in intrinsic dimension across different connected components. While this approach provides a more flexible framework by permitting non-uniform intrinsic dimensions, it explicitly excludes singularities, assuming instead that each component is a manifold. By contrast, our method allows for singular regions, enabling a more comprehensive analysis of complex data that may not adhere strictly to manifold structure.

Our work adopts a ‘singularity-centric’ perspective, arguing that the multi-scale analysis of local geometry and topology is crucial for understanding data with potential singularities. In this regard, topological data analysis (TDA) has become increasingly relevant in machine learning, particularly through techniques like *persistent homology*, which captures the topological features of data at multiple scales [84]. Persistent homology allows us to track changes in topological features across varying resolutions, making it well-suited for uncovering complex structures in data [26, 184].

The notion of examining data at multiple scales is deeply rooted in *scale-space theory*, a concept that dates back to seminal works in computer vision [104, 109]. In scale-space theory, image structures are analyzed at different levels of granularity to capture both local and global information. Although this concept was initially developed for image processing, it has inspired similar approaches in TDA, where

4. Local Topology and Singularities

persistent homology is applied to analyze topological features across scales. Early applications of persistent homology focused primarily on function simplification over manifold domains [62], and it is only more recently that researchers have begun applying it to detect geometric singularities in data.

While persistent homology has been widely used for global unsupervised data analysis tasks, such as clustering and dimensionality reduction [37, 149, 150, 151], these applications primarily target global structures, whereas singularities are inherently local phenomena. A notable exception is the work by Wang et al. [196], which applies persistent homology to analyze local branching and global circular features in data, indicating the potential of TDA for capturing more nuanced topological structures.

Several recent studies take a local perspective on homology, aiming to detect intrinsic dimensionality variations or specific singular structures within data. In particular, Rieck et al. [153] employ a predefined stratification of data combined with *persistent intersection homology* [14], which extends homology to accommodate stratified spaces. Persistent intersection homology provides a framework for handling spaces with a mixture of manifold and non-manifold regions, though it relies on known stratifications, which limits its applicability to general, unstructured data.

Additionally, Fasy and Wang [69] and Bendich [13] developed variants of *local homology* using persistence, focusing on local topological information to detect structural anomalies. These methods reveal intricate topological features in neighborhood data but typically assume a well-defined intrinsic dimension. By contrast, our approach does not necessarily require prior knowledge of the intrinsic dimension and adapts to its variations across the data.

Another related line of work involves *persistent local homology* as employed by Stolz et al. [173], who approximate local homology by computing the absolute homology of an annular region around each point. However, the practical utility of this method remains unclear since the choice of locality parameters which determine the approximation accuracy of local homology is not discussed, and local homology is only approximated using a fixed scale of locality. Moreover, this technique presumes prior knowledge of the intrinsic dimension, which limits its adaptability in cases where dimensionality may vary.

Furthermore, Bendich et al. [15] introduce *persistence vineyards*—a technique that uses continuous families of persistence diagrams to track the local homology of points in stratified spaces. This approach provides a dynamic way to monitor homological changes but is computationally intensive and generally applied to spaces with predefined stratifications. Finally, Dey, Fan, and Wang [53] leverage local homology to estimate the *global* intrinsic dimension of noisy manifolds, a goal aligned with manifold learning but limited in its capacity to detect singularities.

The presented methods indicate a growing interest in topology-driven techniques for detecting singularities and analyzing intrinsic dimensionality. However, existing approaches often assume either a global intrinsic dimension or predefined strat-

ifications, which restricts their ability to handle arbitrary singular structures. Our work diverges from these assumptions by introducing a flexible, unsupervised framework that adapts to local variations in intrinsic dimensionality, enabling effective detection of singular regions without prior knowledge of the data’s topological properties. This framework contributes to the emerging field of singularity analysis in high-dimensional data, offering a novel perspective that integrates persistent local homology with a measure of *Euclidicity* to capture deviations from Euclidean behavior at singular points.

4.3. Methods

Our framework, TARDIS (Topological Algorithm for Robust Discovery of Singularities), is designed to analyze the intrinsic geometric structure of data by leveraging local topological features. It consists of two key components:

- (i) a method for computing the *local intrinsic dimension* of the data, capturing variations in dimensionality across different regions of the space, and (ii) *Euclidicity*, a novel measure that quantifies deviations from an idealized Euclidean structure at multiple scales.

TARDIS is based on the assumption that the intrinsic dimension of data may vary across different regions of the dataset. Therefore, instead of considering a single global dimension, we employ *local measurements*—i.e., measurements performed within small neighborhoods around individual data points. Since no canonical choice for the size of these neighborhoods exists, TARDIS systematically analyzes data across multiple scales to ensure robustness.

Our core idea is to construct a collection of local (*punctured*) neighborhoods at varying scales and to extract their topological features. This process allows us to approximate local topological invariants, specifically local homology, which we use to infer the intrinsic dimensionality of the space. Moreover, by comparing these features to corresponding features of Euclidean model spaces, we can detect singularities and structural anomalies.

4.3.1. Persistent Local Homology

For the formal description of TARDIS, we assume that the input data is represented as a finite metric space $(\mathbb{X}, d)$, where $\mathbb{X}$ is a set of points and d is a metric defining pairwise distances between points. Given a point $x \in \mathbb{X}$, we define the *intrinsic annulus* of x with respect to radii r and s as:

$$B_r^s(x) := \{y \in \mathbb{X} \mid r \leq d(x, y) \leq s\}.$$

This region consists of all points whose distances from x lie within the interval $[r, s]$, effectively forming an annular neighborhood around x .

4. Local Topology and Singularities

To extract topological features, we employ a filtration procedure $\mathcal{F}$ that takes a finite metric space as input and outputs an ascending sequence of topological spaces, such as the Vietoris–Rips filtration. Applying $\mathcal{F}$ to the intrinsic annulus of x yields a tri-filtration:

$$(\mathcal{F}(B_r^s(x), t))_{r,s,t},$$

where t represents the filtration parameter determined by $\mathcal{F}$. This tri-filtration exhibits covariance in s and t , but contravariance in r , leading to the notation:

$$\mathcal{F}(B_\bullet^s(x), \bullet).$$

Applying homology at dimension i to this filtration results in a tri-parameter persistence module, which we define as the i th *persistent local homology* (PLH) of x :

$$\text{PLH}_i(x; \mathcal{F}) := \text{PH}_i(\mathcal{F}(B_\bullet^s(x), \bullet)).$$

Figure 4.2 provides an intuitive visualization of the intrinsic annulus construction and an example of the filtration process.

To the best of our knowledge, this is the first instance where PLH has been considered in the context of multi-parameter persistence. Since the Vietoris–Rips filtration is a standard approach in topological data analysis (TDA), we adopt the shorthand:

$$\text{PLH}_i(x) := \text{PLH}_i(x; \mathcal{V}).$$

Before leveraging PLH for singularity detection, we first establish that our formulation satisfies stability properties similar to those in classical persistent homology. In particular, we demonstrate that small perturbations in input parameters induce only small changes in PLH, thus ensuring robustness in real-world applications.

Theorem 4.3.1. *Given a finite metric space $\mathbb{X}$ and $x \in \mathbb{X}$, let $B_r^s(x)$ and $B_{r'}^{s'}(x)$ denote two intrinsic annuli with $|r - r'| \leq \epsilon_1$ and $|s - s'| \leq \epsilon_2$. Furthermore, let $\mathcal{D}, \mathcal{D}'$ denote the persistence diagrams corresponding to $\text{PH}_i(\mathcal{V}(B_r^s(x), \bullet))$ and $\text{PH}_i(\mathcal{V}(B_{r'}^{s'}(x), \bullet))$. Then $\frac{1}{2} d_B(\mathcal{D}, \mathcal{D}') \leq \max\{\epsilon_1, \epsilon_2\}$.*

Proof. The Hausdorff distance of two non-empty subsets $A, B \subset \mathbb{X}$ is $d_H(A, B) := \inf\{\epsilon \geq 0 \mid A \subset B_\epsilon, B \subset A_\epsilon\}$, where $A_\epsilon = \cup_{a \in A} \{x \in \mathbb{X}; d(x, a) \leq \epsilon\}$ denotes the ϵ -thickening of A in X . Set $\epsilon := \max\{\epsilon_1, \epsilon_2\}$. By assumption, $B_r^s(x) \subset B_{r'}^{s'}(x)_\epsilon$ and $B_{r'}^{s'}(x) \subset B_r^s(x)_\epsilon$, i.e. $d_H(B_r^s(x), B_{r'}^{s'}(x)) \leq \epsilon$. Using the geometric stability theorem of persistence diagrams [36], we have $\frac{1}{2} d_B(\mathcal{D}, \mathcal{D}') \leq d_H(B_r^s(x), B_{r'}^{s'}(x))$, which proves the claim. $\square$

The interpretation of Thm.4.3.1 is that our formulation of PLH indeed yields a stable construction in terms of the choice of the locality parameters r and s . This observation is a fundamental feature of our construction as it allows for varying locality scales in a controllable way. Having established this stability, we now come back to our initial goal of estimating the intrinsic dimension based on PLH.

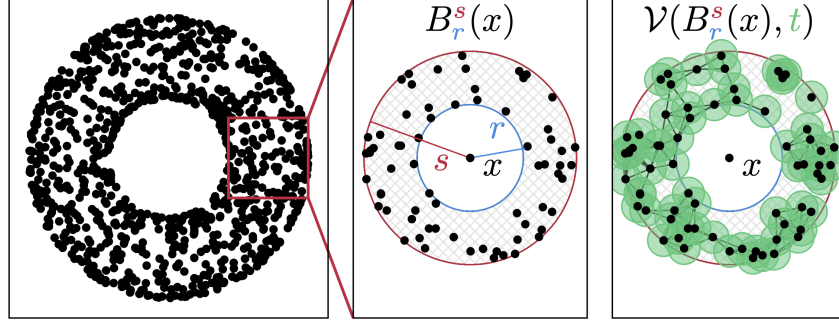


Figure 4.2.: The intrinsic annulus $B_r^s(x)$ around a point x in a metric space $(\mathbb{X}, d)$, as well as one filtration step for some choice of t . By adjusting r and s , we obtain a tri-filtration. This figure was recreated from Von Rohrscheidt and Rieck [193].

4.3.2. Persistent Intrinsic Dimension (PID)

For a finite set of points $\mathbb{X} \subset \mathbb{R}^N$, we define the persistent intrinsic dimension (PID) of $x \in \mathbb{X}$ at scale ϵ as

$$i_x(\epsilon) := \max\{i \in \mathbb{N} \mid \exists r < s < \epsilon \text{ s.t. } \text{PH}_{i-1}(\mathcal{F}(B_r^s(x), \bullet)) \neq 0\}. \quad (4.1)$$

By varying ϵ , this measure characterises the intrinsic dimension of data in a multi-scale fashion. We now prove that we recover the correct dimension in case our data set is sampled from a manifold, in a certain range of locality scale parameters.

Theorem 4.3.2. *Let $M \subset \mathbb{R}^N$ be an n -dimensional compact smooth manifold and let $\mathbb{X} := \{x_1, \dots, x_S\}$ be a collection of uniform samples from M . For a sufficiently large S and $\mathcal{F} = \mathcal{C}$, there exist constants $\epsilon_1, \epsilon_2 > 0$ such that $i_x(\epsilon) = n$ for all $\epsilon_1 < \epsilon < \epsilon_2$ and any point $x \in \mathbb{X}$. Moreover, ϵ_1 can be chosen arbitrarily small by increasing S .*

Proof. Let $x \in \mathbb{X}$ be an arbitrary point. Since M is a manifold, x admits a Euclidean neighbourhood U . Since M is smooth, we can assume U to be arbitrarily close to being flat by shrinking it. Thus, we can find $\epsilon_2 > 0$ with $B_r^s(x) \subset U$ for all $r, s < \epsilon_2$ such that $H_i(\mathcal{C}(B_r^s(x), t)) = 0$ for all $i \geq n$, and all t . Hence, $\text{PH}_i(\mathcal{C}(B_r^s(x), \bullet)) = 0$ for all $i \geq n$, and therefore $i_x(\epsilon_2) \leq n$. By contrast, for S sufficiently large, and r, s as before, there exists a parameter t such that $\mathcal{C}(B_r^s(x), t)$ is homotopy-equivalent to an $(n-1)$ -sphere, and so $H_{n-1}(\mathcal{C}(B_r^s(x), t))$ admits a generator, i.e. it is non-trivial. Consequently, $\text{PH}_{n-1}(\mathcal{C}(B_r^s(x), \bullet)) \neq 0$, and $i_x(\epsilon_2) = n$. By further increasing S , we can ensure that the statement still holds when we decrease ϵ_2 , which proves the two remaining claims. $\square$

Theorem 4.3.2 establishes that for the Čech complex, the function $i_x(\epsilon)$ accurately captures the intrinsic dimension of the underlying manifold M within a specific range of values $\epsilon > 0$, provided that the sample size is sufficiently large. This ensures that, in an appropriate regime, $i_x(\epsilon)$ reliably recovers the correct dimension of M . Moreover, $i_x(\epsilon)$ remains stable within this range, suggesting that examining a

4. Local Topology and Singularities

collection of $i_x(\epsilon)$ for varying values of ϵ can provide a robust method for analyzing the intrinsic dimension at a given point x . Furthermore, leveraging the approximation result from Theorem 3.5.2, we immediately obtain an analogous statement for the Vietoris-Rips complex. This highlights that the intrinsic dimension estimate $i_x(\epsilon)$ extends beyond the Čech complex, ensuring broad applicability. Additionally, we derive the following corollary, which specifically addresses stratified spaces, such as the ‘pinched torus.’ This result implies that our method correctly identifies the intrinsic dimension of individual strata, demonstrating that the persistent intrinsic dimension (PID) approach is capable of handling a large class of ‘non-manifold’ datasets.

Corollary 4.3.1. *Let $X = X_n \supset X_{n-1} \supset X_{n-2} \supset \dots \supset X_{-1} = \emptyset$ be an n -dimensional compact stratified simplicial complex, s.t. $X_i \setminus X_{i-1}$ is smooth for every i . For a fixed i , let $\mathbb{X}_i := \{x_1, \dots, x_S\}$ be a collection of uniform samples from $X_i \setminus X_{i-1}$. For a sufficiently large S and $\mathcal{F} = \mathcal{V}$ or $\mathcal{F} = \mathcal{C}$, there are constants $\epsilon_1, \epsilon_2 > 0$ such that $i_x(\epsilon) = i$ for all $\epsilon_1 < \epsilon < \epsilon_2$ and any point $x \in \mathbb{X}_i$. Moreover, ϵ_1 can be chosen arbitrarily small by increasing S .*

4.3.3. Euclidicity

For our remaining construction, understanding the intrinsic dimension of a neighborhood is fundamental to quantifying its deviation from a purely Euclidean structure. We next introduce the concept of Euclidicity, a measure of how much a given neighborhood differs from being locally Euclidean. A low Euclidicity value signifies that the neighborhood closely resembles a Euclidean space, while a high Euclidicity value indicates regions of singularity or structural complexity within the data set.

Euclidicity is computed without imposing strict manifold assumptions. Instead, it relies on an estimate of the intrinsic dimension n at a given point x . This can be obtained using various dimension estimation methods, such as the previously introduced persistent intrinsic dimension (PID) for different values $i_x(\epsilon)$ across varying locality scales ϵ . Alternatively, other dimension estimation techniques can be incorporated, particularly when prior knowledge about the data structure is available (see [28] for an overview of existing methods).

The core idea behind Euclidicity is to quantify deviations from a Euclidean reference model using persistent local homology (PLH). This involves comparing the topological features of a given neighborhood around x with those of an idealized Euclidean annulus centered at x .

We define a Euclidean annulus $\mathbb{EB}_r^s(x)$ as a set of points uniformly sampled from the region:

$$\{y \in \mathbb{R}^n \mid r \leq \text{dist}(x, y) \leq s\}, \quad (4.2)$$

where r and s are the inner and outer radii of the annulus, respectively. The number of sampled points in $\mathbb{EB}_r^s(x)$ is matched to that in the corresponding intrinsic annulus $B_r^s(x)$ in the given data.

To obtain a multi-scale view, we extend $\mathbb{EB}_r^s(x)$ to larger annuli by sampling additional points, producing a nested sequence $\mathbb{EB}_{r'}^{s'}(x)$ for $r' \leq r$ and $s \leq s'$. Iterating this process generates a tri-filtration:

$$(\mathcal{F}(\mathbb{EB}_r^s(x), t))_{r,s,t} \quad (4.3)$$

for a chosen filtration $\mathcal{F}$, as described in prior discussions.

To formalize the Euclidean reference structure, we define its persistent local homology (PLH) at homology dimension i as:

$$\text{PLH}_i^{\mathbb{E}}(x; \mathcal{F}) := \text{PH}_i(\mathcal{F}(\mathbb{EB}_\bullet^s(x), \bullet)). \quad (4.4)$$

This captures the topological signatures of the Euclidean model space across different scales. Again, for a Vietoris–Rips filtration $\mathcal{V}$, we use a short-form notation, i.e. $\text{PLH}_i^{\mathbb{E}}(x) := \text{PLH}_i^{\mathbb{E}}(x; \mathcal{V})$. Notice that $\text{PLH}_i^{\mathbb{E}}(x)$ implicitly depends on the choice of intrinsic dimension n , and on the samples that are generated randomly. To remove the dependency on the samples, we consider $\text{PLH}_i^{\mathbb{E}}(x)$ to be a sample of a random variable $\mathbf{PLH}_i^{\mathbb{E}}(x)$. Let $D(\cdot, \cdot)$ be a distance measure for 3-parameter persistence modules, such as the *interleaving distance* [108]. We then define the Euclidicity of x , denoted by $\mathfrak{E}(x)$, as the expected value of these distances, i.e.

$$\mathfrak{E}(x) := \mathbb{E}[D(\text{PLH}_{n-1}(x), \mathbf{PLH}_{n-1}^{\mathbb{E}}(x))]. \quad (4.5)$$

This quantity essentially assesses how far x is from admitting a regular Euclidean neighbourhood. The choice of $\mathcal{F}$ and $D(\cdot, \cdot)$ leaves us multiple ways of implementing Eq. (4.5) in practice. We notice that $\mathfrak{E}(x)$ requires different choices, namely (i) a range of locality scales, (ii) a filtration, and (iii) a distance metric between filtrations D . Using a grid Γ of possible radii (r, s) with $r < s$, we approximate Eq. (4.5) using the *mean of the bottleneck distances of fibred Vietoris–Rips barcodes*, i.e.

$$\mathfrak{E}(x) \approx D(\text{PLH}_i(x), \text{PLH}_i^{\mathbb{E}}(x)) := \frac{1}{C} \sum_{(r,s) \in \Gamma} d_B^{r,s} \quad (4.6)$$

where C is equal to the cardinality of the grid,

$$d_B^{r,s} := d_B(\text{PH}_i(\mathcal{V}(B_r^s(x), \bullet)), \text{PH}_i(\mathcal{V}(\mathbb{EB}_r^s(x), \bullet))), \quad (4.7)$$

and with $\text{PLH}_i^{\mathbb{E}}(x)$ denoting a sample from a Euclidean annulus of the same size as the intrinsic annulus around x . We can implement Eq. (4.6) using efficient methods for computing persistent homology [9]. This allows for seamless integration into existing TDA and machine learning frameworks [177, 175].

Furthermore, as established by the following theorem, our approximation of Eq. (4.6) is well-founded: for smooth manifolds, $\mathfrak{E}(x)$ tends to be arbitrarily small in the large-sample regime.

Theorem 4.3.3. *Let $M \subset \mathbb{R}^N$ be a smooth n -dimensional manifold and let $\mathbb{X} \subset M$ be a finite sample of size $S := |\mathbb{X}|$. For a given $\epsilon > 0$, sufficiently large S and a point $x \in \mathbb{X}$, there exists $s_\epsilon > 0$ that only depends on ϵ and the curvature of x (w.r.t. M), such that the approximation of $\mathfrak{E}(x)$ via Eq. (4.6) is bounded above by ϵ , for any grid Γ with maximum outer radius s_ϵ .*

4. Local Topology and Singularities

Proof. Since M is a manifold, every point x admits a Euclidean neighbourhood U . Moreover, as M is smooth, we may assume U to be arbitrarily close to being flat (by potentially shrinking it). Thus, we can find an s_ϵ with the desired properties such that (i) $B_0^{s_\epsilon}(x) \subset U \cap \mathbb{X}$, and (ii) $d_H(B_0^{s_\epsilon}(x), \mathbb{E}B_0^{s_\epsilon}(x)) \leq \frac{\epsilon}{2}$, where the latter condition can be achieved by increasing S if necessary. Consequently, $d_H(B_r^s(x), \mathbb{E}B_r^s(x)) \leq \frac{\epsilon}{2}$ for any $0 \leq r \leq s \leq s_\epsilon$, and by making use of the geometric stability theorem [36], the right-hand side of Eq. (4.6) reduces to

$$\frac{1}{C} \sum_{(r,s) \in \Gamma} d_B(\text{PH}_i(\mathcal{V}(B_r^s(x), \bullet)), \text{PH}_i(\mathcal{V}(\mathbb{E}B_r^s(x), \bullet))) \leq \frac{2}{C} \sum_{(r,s) \in \Gamma} d_H(B_r^s(x), \mathbb{E}B_r^s(x)) \leq \epsilon$$

for any grid Γ with the properties specified in the statement. This concludes the proof. $\square$

One of the key advantages of our formulation is that both the *Persistent Intrinsic Dimension (PID)* and *Euclidicity* can be computed without requiring strong assumptions about the structure or distribution of the input data. Instead, we rely on a minimal and reasonable assumption: that the intrinsic dimension n of the data is significantly lower than the ambient dimension N . This makes our approach particularly well-suited for analyzing high-dimensional data with an underlying low-dimensional structure, which is common in many real-world scenarios.

A crucial aspect of our approach is its local nature, allowing the dimension to vary across the dataset. This flexibility enhances its applicability across diverse datasets and geometric structures. Unlike traditional methods that impose global constraints on dimension, our approach accommodates settings where different regions of a dataset may exhibit varying intrinsic dimensions. This enables a more fine-grained characterization of the underlying data geometry, making our method broadly applicable to problems in data analysis and machine learning.

Moreover, as we have demonstrated, our method is *guaranteed* to produce the correct values for manifolds and stratified simplicial complexes. This theoretical guarantee enhances both the practical utility and expressiveness of our framework. For instance, our approach naturally extends to cases where the dataset consists of unions of manifolds with different dimensions, a setting where many existing dimensionality estimation techniques struggle. In contrast to prior approaches [23], Euclidicity generalizes to a broader class of spaces, enabling a more precise structural analysis of complex datasets.

Our implementation of Euclidicity relies on the *Vietoris–Rips complex*, a fundamental construction in topological data analysis. However, a well-known challenge is that the Vietoris–Rips complex grows exponentially with increasing dimensionality, leading to high memory and computational costs.

This limitation can be alleviated by using alternative filtrations [4, 169]. Importantly, our theoretical results do *not* rely on any specific filtration, meaning that our framework is adaptable to different filtration choices.

Finally, we emphasize that the reliability of the Euclidicity score is fundamentally

dependent on the validity of the estimated intrinsic dimension. If the intrinsic dimension is incorrectly estimated, the comparison is no longer performed with respect to the correct model space, potentially leading to misleading results. Ensuring robust intrinsic dimension estimation is thus an essential step in effectively applying Euclidicity in practice.

4.4. Experiments

We demonstrate the expressive power of TARDIS across various settings, emphasizing its ability to capture key geometric properties of data. Specifically, we show that our method: (i) accurately estimates the intrinsic dimension of datasets, ensuring a correct characterization of their underlying structure, and (ii) effectively detects singularities when analyzing datasets with known geometric irregularities. Additionally, we perform a comparative analysis with one-parameter approaches, highlighting how our multi-scale methodology yields more stable and reliable results. Finally, we investigate Euclidicity scores across both benchmark and real-world datasets, providing empirical evidence that our technique serves as a meaningful measure of the geometric complexity of data.

4.4.1. Parameter Selection

Since Eq. (4.5) inherently incorporates multiple scales of locality, it is necessary to specify an upper bound for the radii ($r_{\min}, r_{\max}, s_{\min}, s_{\max}$) that define the respective annuli in practice.

Given a point x , we found the following procedure to be effective in practice: we set $s_{\max}$, i.e., the maximum outer radius, to the distance to the k th nearest neighbor of x , and $r_{\min}$, i.e., the minimum inner radius, to the smallest non-zero distance to a neighbor of x . Furthermore, we set the minimum outer radius $s_{\min}$ and the maximum inner radius $r_{\max}$ to the distance to the $\lfloor \frac{k}{3} \rfloor$ th nearest neighbor. While we found $k = 50$ to yield sufficiently stable results, spaces with a high intrinsic dimension may require larger values.

The advantage of this parameter selection procedure is its data-driven nature, which automatically adapts to variations in density. Since our approach is inherently *local*, there is a trade-off between ensuring sufficiently large sample sizes to capture topological information and maintaining a small enough scale to preserve locality. Empirically, we found the chosen range to be a suitable compromise.

Regarding the number of steps, we discretize the parameter range using 20 steps by default. Higher values are recommended in cases where there are significant differences between the radii, such as when $s_{\max} \gg r_{\max}$.

4. Local Topology and Singularities

Table 4.1.: Dimensionality estimates for the concatenation of S^1 and S^2 , denoted by $S^1 \vee S^2$. Our PID measure is more capable of detecting the changes in dimensionality that arise from the concatenation.

	METHOD	MIN	$\mu \pm \sigma$	MAX
1D	lpca	1.00	1.42 ± 0.78	3.00
	twoNN	0.83	1.00 ± 0.07	1.20
	DANCo	1.00	1.00 ± 0.01	1.16
	PID	1.00	1.12 ± 0.24	1.97
2D	lpca	2.00	2.88 ± 0.32	3.00
	twoNN	1.01	1.90 ± 0.36	2.53
	DANCo	1.00	2.10 ± 0.32	3.00
	PID	1.52	1.95 ± 0.06	2.08

4.4.2. Persistent Intrinsic Dimension is Expressive

We first analyze the behavior of persistent intrinsic dimension (PID) on samples from a space obtained by concatenating S^1 (a circle) and S^2 (a sphere) at a gluing point. Table 4.1 presents a comparison of PID with state-of-the-art dimensionality estimators.¹ We observe that PID outperforms all other estimators in terms of both mean and standard deviation for the 2D points, correctly indicating that the majority of points admit non-singular 2D neighborhoods. For the 1D points, the mean dimensionality estimate of PID remains close to the ground truth, while its standard deviation and maximum appropriately reflect the fact that some 1D points lie closer to the gluing point.

Fig. 4.3 illustrates this behavior by comparing PID with one of the other estimators. PID proves to be more sensitive in detecting variations in local intrinsic dimension as points approach the gluing region, correctly identifying that points on the sphere exhibit 2D neighborhoods. This aligns with our perspective that dimensionality should be treated as an inherently local property. If such sensitivity is not desirable for a specific dataset, Euclidicity calculations can support *any* dimensionality estimator. However, since these estimators do not provide strong theoretical guarantees like those in Theorem 4.3.2, their selection should ultimately be guided by the characteristics of the dataset in question.

In practice, the sample density may not always be sufficiently high for Theorem 4.3.2 to be directly applicable. As a consequence, spurious homological features may appear in dimensions *higher* than the intrinsic dimension of the underlying space. To mitigate this issue, we recommend considering only those features that exceed a certain persistence threshold relative to the persistence of features in lower dimensions. Specifically, for any data point x and its corresponding intrinsic annulus $B_r^s(x)$, we propose eliminating all topological features whose lifetimes are shorter than the maximum lifetime of features in the dimension directly below. This filtering

¹Method names are taken from the `scikit-dimension` toolkit, see <https://scikit-dimension.readthedocs.io/en/latest/>

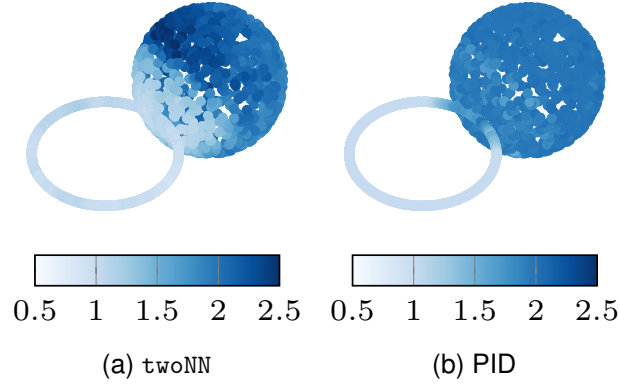


Figure 4.3.: Dimensionality estimates. PID is more nuanced in capturing changes in dimensionality, assigning ≈ 1 to almost all points of the circle, i.e. S^1 , while highlighting that points closer to S^2 exhibit an increase in dimensionality. This figure was recreated from Von Rohrscheidt and Rieck [193].

strategy leads to significantly more stable intrinsic dimension estimates, reducing the risk of overestimation.

4.4.3. Euclidicity Captures Known Singularities

Before applying Euclidicity to high-dimensional spaces with unknown characteristics, we first investigate its behavior on spaces with well-understood singularities. As illustrated in Fig. 4.1, Euclidicity effectively detects the singularity of the ‘pinched torus.’ Crucially, it also identifies that points near the singularity are *not* fully regular, highlighting its sensitivity to local geometric irregularities. This ability to reveal *isolated singularities* despite sampling noise makes Euclidicity particularly useful in practical applications.

A further example of a singular space is the *wedge* of two n -dimensional spheres, $S^n \vee S^n$, formed by gluing them at a single point x_0 . For an appropriate triangulation, this space can be stratified as $X \supset \{x_0\}$. To assess the effectiveness of TARDIS, we apply it to sampled wedge spaces in dimensions $\{2, 3, 4\}$, computing their Euclidicity scores. Since higher-dimensional spaces require denser sampling, we start with 20,000 points in dimension 2 and increase the sample size by a factor of 10 for each additional dimension. For each dataset, we evaluate the Euclidicity of 50 randomly selected points, along with the singular point x_0 .

The results, shown in Fig. 4.4a, confirm that the singular point consistently exhibits a *markedly higher Euclidicity* score than regular points. Additionally, the scores of non-singular points display significant variance, which reflects the imperfect alignment of sampled data with the underlying geometric structure. This reinforces our key argument: Euclidicity provides a continuous measure of local Euclidean structure rather than a binary classification, offering a more nuanced characterization of data geometry.

4. Local Topology and Singularities

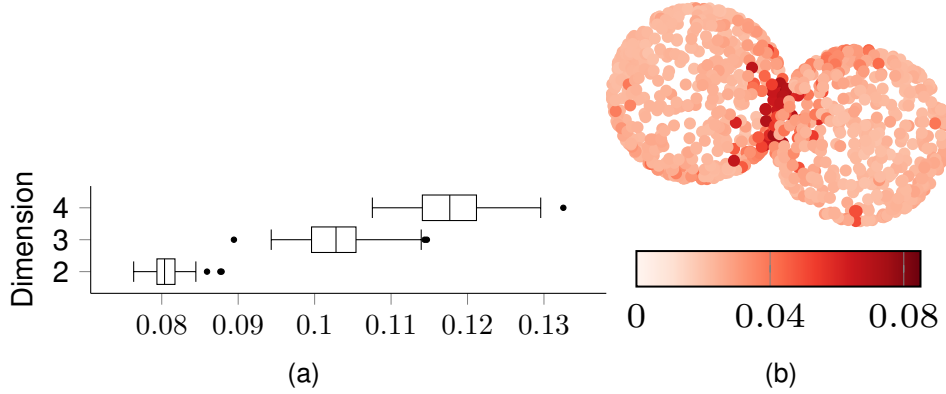


Figure 4.4.: **a**: Euclidity scores of wedged spheres for different dimensions. High values indicate singular points/neighbourhoods. The Euclidity of the singular point always constitutes a clear positive outlier. In 2D, *Euclidity* **b** results in a clearly-delineated singular region. This figure was recreated from Von Rohrscheidt and Rieck [193].

4.4.4. Stability of Euclidity

Theorem 4.3.1 establishes that Euclidity estimates are stable. To validate this claim, we first examine whether Euclidity is *robust to sampling*, a crucial property given the widespread use of mini-batches in machine learning. Repeating the computations for the ‘pinched torus’ across different batches produces highly similar distributions, which, according to Tukey’s range test [182], are statistically indistinguishable at the $\alpha = 0.05$ confidence level. Furthermore, increasing the locality scale allows us to detect singularities even more reliably, albeit at higher computational costs and with a greater portion of the point cloud incorporated.

For a more detailed analysis of these stability properties, see Section A.2.

4.4.5. Comparison to single-parameter approach

We find that Euclidity provides significantly more stable results compared to a single-parameter approach for geometry-based anomaly detection [173]. A detailed analysis in Section A.4 reveals that single-parameter methods, which evaluate data at a fixed global scale, frequently produce ‘false positives’—i.e., points assigned high anomaly scores despite admitting a Euclidean neighborhood. This issue becomes even more pronounced in higher dimensions and sparser datasets, underscoring the advantage of Euclidity’s multi-scale approach.

The preceding experiments validated PID and Euclidity in terms of both *expressivity*, by detecting singularities in spaces with known ground truth, and *robustness*. Building on these insights, our subsequent experiments will extend the analysis to high-dimensional spaces, showcasing potential applications of Euclidity.



(a) MNIST



(b) FASHIONMNIST

Figure 4.5.: Euclidity captures local sample complexity in an unsupervised manner. From left to right: sample images exhibiting low, median, and high Euclidity, respectively. This figure was recreated from Von Rohrscheidt and Rieck [193].

4.4.6. Euclidity of High-Dimensional Spaces

To evaluate TARDIS in an unsupervised setting, we compute Euclidity scores for the MNIST and FASHIONMNIST datasets. Specifically, we select mini-batches of 1000 samples from a randomly drawn subset of 10000 images. Each greyscale image is ‘flattened’ into a high-dimensional vector, treating the entire collection as a point cloud, for which we compute Euclidity values. This approach aligns with standard practices in dimensionality reduction [116, 121].

Following Pope et al. [140], we assume an intrinsic dimension of 10 and set $k = 50$ for local scale estimation. To ensure robustness, all calculations are repeated across five independent subsamples. The resulting Euclidity scores range from [1.1, 5.3] for MNIST and [1.3, 5.6] for FASHIONMNIST.

Notably, the score distributions for the two datasets differ, suggesting distinct geometric properties. For a detailed visualization and further analysis of these distributions, see Section A.5.

Fig. 4.5 presents a selection of 9 images, corresponding to the lowest, median, and highest Euclidity scores, respectively. We observe that images with high Euclidity scores exhibit a greater degree of non-linearity, whereas those with low Euclidity scores tend to have simpler structures. For MNIST, low-Euclidity

4. Local Topology and Singularities

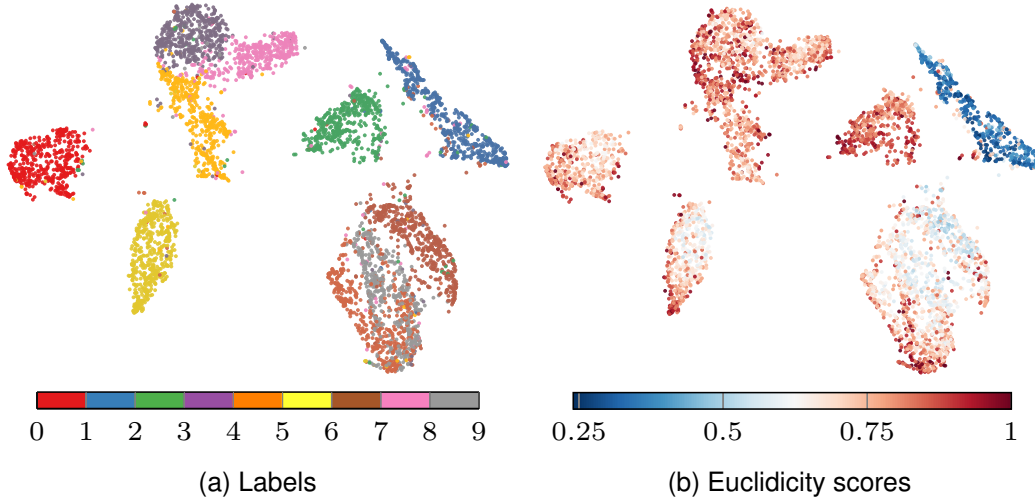


Figure 4.6.: A UMAP [116] embedding of the MNIST data, showing Euclidity scores and class labels. This figure was recreated from Von Rohrscheidt and Rieck [193].

images predominantly correspond to the digit ‘1,’ which features minimal geometric complexity. A similar pattern emerges in FASHIONMNIST, where images with low Euclidity—such as ‘pants’—display relatively simple geometric structures compared to those with higher Euclidity scores.

Since low Euclidity can be interpreted as an indicator of how closely a neighborhood approximates a *locally linear* structure, these findings suggest the presence of simple substructures within the data. This further supports the potential of Euclidity as an unsupervised measure of geometric complexity.

To investigate the relationship between geometric complexity and Euclidity, we focus on MNIST. Fig. 4.6 presents an embedding of the dataset, where labels and *normalized Euclidity scores* (scaled by their maximum value) are highlighted. Notably, Euclidity was computed solely on the raw data, with the embedding serving purely as a visualization tool. We observe that low Euclidity scores are concentrated in clusters of ‘1’s, whereas ‘5’s exhibit both lower-than-average and higher-than-average Euclidity scores—a trend that extends to other digit classes. This supports the interpretation of MNIST as a *union of manifolds*, though these manifolds do not necessarily align with digit classes. In other words, our Euclidity scores suggest that even images with the same label may not reside on the same geometric structure.

To further examine this hypothesis, we analyze the relationship between a classifier’s ability to correctly predict a sample and its corresponding normalized Euclidity score. We train a simple neural network classifier for both MNIST and FASHIONMNIST and compare the Euclidity scores of correctly and incorrectly classified samples. Our results reveal that misclassified samples exhibit significantly higher Euclidity scores than a randomly selected, equally sized subset of correctly classified images.

For MNIST, the average Euclidity score of correctly classified samples is 0.33, whereas misclassified samples have a higher average of 0.39. Similarly, for FASHIONMNIST, correctly classified samples average 0.31, while misclassified samples reach 0.35. Welch’s t-test confirms the statistical significance of these differences. For details on the model architecture and significance testing, see Section A.5.

These findings suggest promising future applications of Euclidity in classification models—either as an unsupervised preprocessing step to improve feature representations or as an additional loss term integrated directly into model architectures to enhance classification performance.

4.4.7. Euclidity of Cytometry Data

To demonstrate the utility of Euclidity in unsupervised representation learning, we apply it to an induced pluripotent stem cell (iPSC) reprogramming dataset [215]. This dataset comprises 33 variables and approximately 220,000 samples, capturing the progression of *fibroblasts* as they diverge and split into two distinct lineages. Previous studies have identified branching structures within this dataset [19], which are most effectively visualized using PHATE [120], a non-linear dimensionality reduction technique.

It is important to note that PHATE is used solely for visualization; all Euclidity calculations are conducted on the original high-dimensional data. To estimate the intrinsic dimensionality, we employ the `twoNN` method, obtaining a mean intrinsic dimension of 16. Parameter selection follows the approach outlined in Section 4.4.6.

Fig. 4.7a presents a PHATE embedding alongside the Euclidity scores computed on the original data. We observe that regions with high Euclidity scores tend to coincide with lower-density areas in the PHATE visualization. Upon further inspection, these points are situated in lower-dimensional subspaces.²

To verify this observation, we use `twoNN` dimensionality estimates, as shown in Fig. 4.8. Specifically, we compute the intrinsic dimensionality for a subsample, finding that the interquartile range for the 1,000 points with the *highest Euclidity* scores lies between 12 and 14, while for the 1,000 points with the *lowest Euclidity* scores, it ranges from 13 to 16. These estimates are obtained using the `twoNN` algorithm with $k = 50$ nearest neighbors.

Since points residing in lower-dimensional subspaces can be interpreted as *singular* within the framework of stratified spaces, this provides further evidence that Euclidity is a valuable tool for detecting non-manifold structures in data.

Moreover, we find that Euclidity estimates remain stable under *subsampling*. Fig. 4.7 illustrates this by showing Euclidity distributions across different subsample sizes, computed on the raw data and visualized via PHATE embeddings.

²However, low-density regions in the PHATE visualization do not necessarily correspond to low-density regions in the original dataset.

4. Local Topology and Singularities

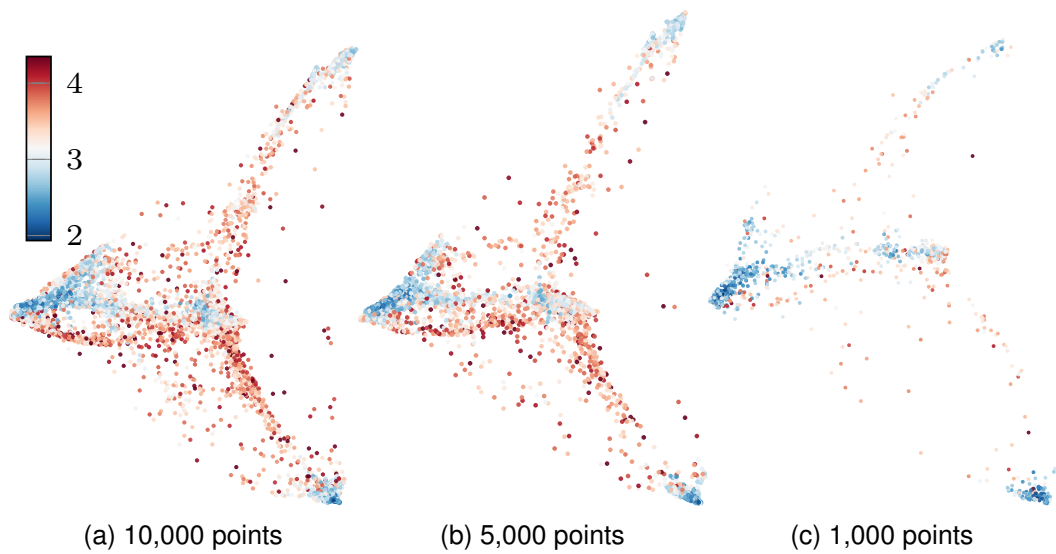


Figure 4.7.: PHATE embeddings of a cytometry data set, with colours based on Euclidity. Euclidity highlights dense non-singular regions and remains stable under subsampling the iPSC data set. Minor variations in the point cloud shape are due to the PHATE embedding algorithm; Euclidity was always calculated on the raw data. This figure was recreated from Von Rohrscheidt and Rieck [193].

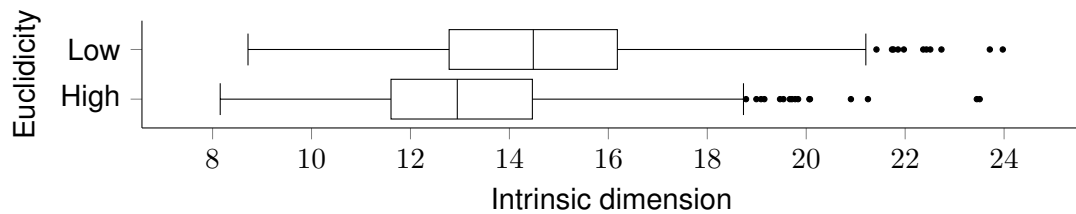


Figure 4.8.: A comparison of intrinsic dimension estimates computed for points in the iPSC dataset that admit low (top) and high (bottom) Euclidity scores. This figure was recreated from Von Rohrscheidt and Rieck [193].

The stability of these distributions confirms that the observed patterns persist across varying sample sizes.

This analysis serves as a proof of concept, demonstrating how Euclidity can be used to ‘probe’ datasets and enhance general unsupervised data analysis workflows.

4.5. Discussion

In this chapter, we introduced TARDIS, a novel framework for the local estimation of two fundamental geometric properties of data embedded in Euclidean space: (i) the intrinsic dimension, quantified through the *persistent intrinsic dimension* (PID), and (ii) the deviation from Euclidean space, captured by *Euclidity*, a multi-scale

measure of manifoldness. Our approach is grounded in a multi-parameter extension of persistent local homology, for which we provided theoretical guarantees in a dense sample regime. We demonstrated empirically that TARDIS exhibits improved stability compared to traditional geometry-based anomaly detection methods that rely on fixed locality scales. Furthermore, Euclidicity was shown to be effective in identifying singular regions within data sets that are known to contain singularities. Notably, our experiments on high-dimensional benchmark and real-world data sets suggest that Euclidicity can serve as an unsupervised measure of geometric complexity, offering insights into the structure of data representations beyond conventional techniques.

Building on the promising results of our study, we outline several key directions for future work. A primary avenue for further research is the incorporation of Euclidicity into machine learning pipelines, particularly in settings where sensitivity to singular structures is beneficial. In light of our results in Section 4.4.6 and Section 4.4.7, we anticipate that Euclidicity will be especially useful in unsupervised learning tasks or as an auxiliary measure in classification settings. By assigning lower confidence scores to data points residing in singular regions, Euclidicity may enhance the robustness and reliability of predictive models. Additionally, our method could be leveraged for the detection of adversarial examples, a problem where understanding the underlying topology of the data space is known to be critical [95]. A second promising direction involves improving the theoretical and computational properties of Euclidicity itself. One approach is to explore alternative distance measures for three-parameter persistence modules, i.e., incorporating different metrics into the computation of Euclidicity as formalized in Eq. (4.6). The goal of this refinement would be to enhance stability, expressivity, and computational efficiency. Moreover, we hypothesize that replacing the Vietoris–Rips filtration with alternative constructions, such as witness complexes or other topological filtrations, may reduce the number of samples required for reliable Euclidicity estimation [4, 169]. A broader research question that naturally emerges from our findings concerns the approximation of an underlying manifold from data that inherently exhibits singularities. More precisely, we seek to investigate the feasibility of constructing the *best* manifold approximation for a given singular data set. This has potential implications for generative modeling and representation learning, as singularities could be mitigated or explicitly accounted for during model training. If an appropriate loss function is defined, Euclidicity may serve as a regularizer that encourages embeddings to respect the topological structure of the data. Ultimately, this could pave the way toward more robust and trustworthy representations in high-dimensional machine learning applications.

5. Local Topology and Graph Data

This chapter introduces a novel, local method that encodes a fingerprint of local graph data, bridging the gap between topology and modern graph learning techniques. The approach is based on a local perspective of Euler Characteristic Transforms, a geometrical-topological technique that captures shape information along all directions in the embedding space. In the language of topology, the given method characterizes local neighborhoods of featured graphs, providing a way to encode and compare the local topological structure of graphs embedded in Euclidean space. The latter enables a richer understanding of local graph features, extending beyond classical combinatorial descriptors. In this sense, the proposed methodology serves as a versatile and expressive framework that unites local topological data analysis with modern graph representation learning, providing a powerful new tool for studying complex graph-structured data locally.

5.1. Introduction

Graph Neural Networks (GNNs) have emerged as a powerful class of models for learning representations on graph-structured data. Traditionally, GNNs rely on message-passing schemes to iteratively aggregate node features from their local neighborhoods. While effective for many tasks, this approach can lead to the loss of crucial local information, as the aggregation process inherently diffuses and obscures the original node representations [180]. This limitation becomes particularly problematic in scenarios where the preservation of fine-grained local structure is critical for downstream tasks such as node classification, link prediction, and anomaly detection.

To address this challenge, we propose the use of the *Euler Characteristic Transform* (ECT), a geometrical-topological invariant known for its advantageous computational and expressivity properties [185]. The ECT captures topological features of data through sublevel set filtrations applied across multiple directions. Notably, for so-called constructible sets, the ECT is *invertible*, meaning that the original data can be reconstructed from its transform [48]. This invertibility, combined with its reliance on weighted summation, enables the ECT to serve as an efficient and powerful tool for representation learning [155].

In this chapter, we extend the ECT to local neighborhoods, introducing the *Local Euler Characteristic Transform* (ℓ -ECT), a novel framework that enhances representation learning on graph-based data. The ℓ -ECT is specifically designed to preserve local structural details while maintaining global interpretability, capturing

5. Local Topology and Graph Data

both the spatial and topological properties of node neighborhoods. By focusing on the localized application of the ECT, our method ensures that critical structural information is retained during aggregation, which is particularly advantageous in graph-based learning tasks.

Although the ℓ -ECT is derived from the Euler characteristic, it serves as a distinctive *fingerprint* of local neighborhoods and should be viewed as a geometrical-topological approach rather than a purely combinatorial method. Our approach directly addresses the challenge of neighborhood aggregation in featured graphs, ensuring lossless representation of local structures. Through theoretical analysis, we demonstrate that ℓ -ECTs maintain key local properties, providing a more refined representation that can be effectively utilized for downstream tasks such as node classification and graph-based anomaly detection.

Empirical evaluations further highlight the effectiveness of our method, particularly in scenarios where standard node aggregation techniques obscure essential differences. Specifically, ℓ -ECTs prove to be highly beneficial in graphs with high heterophily, where traditional message-passing-based GNNs struggle to capture meaningful interactions. Furthermore, due to its natural vectorized representation, the ℓ -ECT framework is readily compatible with a broad range of machine learning models, thereby facilitating both enhanced performance and improved interpretability.

Our main contributions can be summarized as follows: (i) We formally construct ℓ -ECTs within the framework of embedded simplicial complexes and graphs, providing a theoretical investigation of their expressivity in the context of featured graphs. (ii) We empirically demonstrate that ℓ -ECTs serve as a powerful and interpretable tool for node classification, often outperforming standard GNN architectures, particularly in challenging graph learning tasks. (iii) We introduce an efficiently computable rotation-invariant metric based on ℓ -ECTs, which enables the spatial alignment of geometric graphs and enhances their usability in geometric learning applications.

Through these contributions, we establish ℓ -ECTs as a versatile and expressive framework that bridges the gap between topological data analysis and modern graph-based machine learning models, paving the way for more robust and interpretable graph learning methodologies. The results of this chapter are essentially based on [156].

5.2. Related Work

Graph Neural Networks (GNNs) have fundamentally reshaped the field of graph representation learning by enabling end-to-end learning of node embeddings through message passing [102]. The introduction of message-passing architectures has led to remarkable advancements in numerous graph-based tasks, including node classification, link prediction, and graph classification. However, traditional GNNs exhibit fundamental theoretical limitations, particularly in their ability to distinguish

non-isomorphic graph structures. These limitations have been formally characterized in terms of the expressivity of the Weisfeiler-Lehman test [198].

Beyond expressivity limitations, GNNs are also known to suffer from practical challenges such as *oversmoothing* [205, 163] and *oversquashing* [54]. Oversmoothing refers to the phenomenon where deep GNNs cause node embeddings to become indistinguishable, leading to a loss of discriminative power. Oversquashing, on the other hand, emerges in the presence of long-range dependencies, where information is excessively compressed in a small embedding space, limiting the network’s ability to propagate signals effectively.

To alleviate these issues, various architectural modifications have been proposed. Early approaches focused on modifying the message-passing paradigm using attention mechanisms [189] and sampling-based methods [80] to improve efficiency and expressivity. However, despite these enhancements, GNNs often struggle with learning meaningful representations in heterophilous graphs, where connected nodes tend to have different class labels [212].

A recent line of research has explored incorporating additional inductive biases into GNN architectures. Notable advancements include leveraging geometric information [136, 98] and topological features [90, 190] to enhance model expressivity. The goal of these methods is to improve a model’s ability to distinguish non-isomorphic graph structures, as formalized in recent studies on expressivity [125]. These approaches are part of a broader paradigm shift toward *topological deep learning* [213], an emerging field that aims to integrate higher-order topological structures into machine learning models.

Independent of graph-based learning, the ECT has gained popularity in Topological Data Analysis (TDA) as a tool for capturing shape information in data [185, 78]. The ECT provides a systematic way to extract topological features from data and has been successfully applied in various domains, including shape classification and computer vision. Recent efforts have focused on integrating the ECT into deep-learning architectures [155] and extending its theoretical foundation to include additional invariance properties [48, 115]. However, existing work primarily explores *global* ECT representations, with little attention given to *local* variants, which are crucial for capturing fine-grained topological structures.

In addition to the ECT, other topological tools have been incorporated into graph-learning pipelines. One prominent approach is the use of *persistent homology* (PH), a powerful yet computationally expensive method for extracting multi-scale topological features. Applications of PH in graph-based tasks include graph classification [87] and learning topological descriptors through weighted kernels [207]. A closely related work by Zhao et al. [210] incorporates topological features of graph neighborhoods into a GNN framework, leveraging PH to enhance structural representations.

Despite these advances, the integration of topology into GNNs remains an open challenge. While persistent homology and the ECT provide rich descriptors of

5. Local Topology and Graph Data

graph structure, their computational costs and lack of localized variants hinder their applicability in large-scale learning tasks. To the best of our knowledge, our work is the first to develop *local* variants of the ECT tailored for graph-learning applications. Furthermore, we provide a theoretical analysis of these local variants and empirically demonstrate their effectiveness in node classification tasks, bridging the gap between topological descriptors and modern graph representation learning paradigms.

5.3. Methods

5.3.1. The ECT and its approximation

We recall that the Euler Characteristic Transform (ECT) of an (abstract) simplicial complex $X \subset \mathbb{R}^n$ is a function $\text{ECT}(X) : S^{n-1} \times \mathbb{R} \rightarrow \mathbb{Z}$, given by

$$\text{ECT}(X)(v, t) = \chi(\{x \in X \mid x \cdot v \leq t\}), \quad (5.1)$$

where χ denotes the Euler characteristic. The interpretation of $\text{ECT}(X)$ is that it scans the ambient space of X in every direction and records the Euler characteristic of the sublevel sets. In [78], it is proven that $\text{ECT}(X)$ is invertible, meaning that X can be recovered from $\text{ECT}(X)$, as long as X is a so-called constructible set. The main focus of this work are compact simplicial complexes which are constructible, and therefore the invertibility theorem applies in our setting.

We now give an overview of the methods of this chapter which are fundamentally based on the ECT. We first establish some notation. We denote the approximation of $\text{ECT}(X)$ by $\overline{\text{ECT}}(X)_{(m,l)}$ and define it as

$$\overline{\text{ECT}}(X)_{(m,l)} := \text{ECT}(X)|_{\{v_1, \dots, v_m\} \times \{t_1, \dots, t_l\}} \quad (5.2)$$

for uniformly distributed directions $v_1, \dots, v_m \in S^{n-1}$ and filtration steps $t_1, \dots, t_l \in \mathbb{R}$.

Since X is a compact space, it is possible to select a sequence of points $t_1, \dots, t_l$ that lie within a compact interval $[a, b]$, where the endpoints satisfy $t_1 = a$ and $t_l = b$. Moreover, we impose that the sequence $\{t_i\}_{i=1}^l$ forms a uniform partition of $[a, b]$, ensuring that the interval is evenly discretized. This choice of discretization allows for an efficient computational approximation of the domain of $\text{ECT}(X)$, as the induced representation naturally takes the form of a vector of dimension $m \cdot l$, where m represents the number of sample points in a (potentially high-dimensional) sphere, and l corresponds to the number of partition points along the chosen interval.

Regarding the selection of appropriate values for m and l , we take into account established results on the expected nearest neighbor distances in uniformly sampled high-dimensional spaces. Specifically, for uniform samples on the unit sphere S^n , the expected nearest neighbor distance is known to scale as $\mathcal{O}((\frac{\log m}{m})^{1/n})$ (see

[11]). Simultaneously, the uniform partitioning of a compact interval contributes an additional discretization error that scales as $\mathcal{O}(\frac{1}{l})$. Combining these two factors, the total approximation error can be estimated as

$$\mathcal{O}\left(\left(\frac{\log m}{m}\right)^{1/(n-1)} \frac{1}{l}\right).$$

This formulation provides a theoretical basis for balancing the trade-off between computational complexity and approximation accuracy when discretizing the spatial domain of $\text{ECT}(X)$.

Curry, Mukherjee, and Turner [48] establish that the aforementioned approximation scheme not only provides a computationally efficient representation but also accurately determines the true quantity of interest, provided that certain mild regularity conditions hold and that the parameters m and l are sufficiently large.

We notice that both translations and uniform scalings of X within the ambient space induce a reparametrization of $\text{ECT}(X)$. Consequently, the structural properties of $\text{ECT}(X)$ remain invariant under these transformations, up to a change in its parameterization.

5.3.2. The local ECT (ℓ -ECT)

Given a simplicial complex $X \subset \mathbb{R}^n$ and a vertex $x \in X$, we define the *local Euler Characteristic Transform* (local ECT) of x with respect to a scale parameter $k \geq 0$ as

$$\ell\text{-ECT}_k(x; X) := \text{ECT}(N_k(x; X)), \quad (5.3)$$

where $N_k(x; X)$ denotes an appropriately chosen local neighborhood of x within X , whose scale of locality is governed by the parameter k . Typically, $N_k(x; X)$ is defined as either the full subcomplex of X spanned by the k -hop neighbors of x in X , or alternatively, the full subcomplex spanned by the k -nearest vertices of x in the ambient space.

An important special case arises when X is a 0-dimensional simplicial complex, which corresponds to a point cloud. In this scenario, the full subcomplex of X spanned by the k -nearest vertices of x , denoted as $N_k(x; X)$, reduces to the classical k -nearest neighbor set of x . This case is particularly relevant in practical applications where point clouds serve as the primary data representation, and it shows that embedded simplicial complexes vastly generalize point clouds when considered as a data structure.

At first glance, the construction of local ECTs appears to be a purely topological operation due to its reliance on the Euler Characteristic. However, in light of the invertibility theorem, we recognize that $\ell\text{-ECT}_k(x; X)$ can be interpreted as a *fingerprint* of the local neighborhood of x within X . This perspective is especially valuable as it implies that the local ECT encodes rich geometric and topological information about local structures in a manner that is both expressive and computationally feasible. Moreover, this fingerprint can be approximated efficiently in practice,

5. Local Topology and Graph Data

making it possible to extract local structural representations from combinatorial data embedded in Euclidean space.

The practical computation of this approximation proceeds by sampling directions $v_1, \dots, v_m \in S^{n-1}$ and threshold values $t_1, \dots, t_l \in \mathbb{R}$, and subsequently considering the empirical approximation $\overline{\text{ECT}}(N_k(x; X))_{(m,l)}$ in place of $\ell\text{-ECT}_k(x; X)$. This empirical quantity is computationally tractable, and the associated approximation error can be systematically controlled by adjusting the sample sizes m and l , as discussed previously. Notably, this approximation admits a natural representation as a vector of dimension $m \cdot l$, enabling the encoding of local structural information in a manner that is both approximately lossless and amenable to machine learning applications. Consequently, these representations can be directly leveraged for downstream tasks, such as classification, clustering, or regression, where capturing local geometric and topological features is of critical importance.

5.3.3. ℓ -ECT in representation learning

Our formulation of local Euler Characteristic Transforms (ℓ -ECT) provides a natural and expressive representation of local neighborhoods in embedded simplicial complexes. A particularly important special case arises in the context of *featured graphs*, i.e., graphs in which each node is associated with a feature vector. Such data structures form the foundation of a wide range of modern graph learning tasks, including node classification, graph classification, and graph regression. The predominant class of methods used to address these problems are *message-passing graph neural networks* (GNNs), which iteratively aggregate information from neighboring nodes.

Here, we propose an alternative approach for processing featured graph data based on ℓ -ECT, demonstrating that ℓ -ECT representations encode sufficient information to perform message passing. We provide a rigorous justification for this claim in the following sections.

Definition 5.3.1. *A featured graph is an undirected graph $\mathcal{G}$ in which each node $v \in \mathcal{G}$ is associated with a feature vector $x(v) \in \mathbb{R}^n$. The set of nodes in $\mathcal{G}$ is denoted by $V(\mathcal{G})$, and the set of edges by $E(\mathcal{G})$.*

A featured graph $\mathcal{G}$ can naturally be interpreted as an *embedded graph* in $\mathbb{R}^n$, where each node is represented by its feature vector as a point in $\mathbb{R}^n$. The edges of the embedded graph are then defined by retaining the connectivity structure of $\mathcal{G}$, meaning that an edge exists between two embedded points if and only if there is an edge between the corresponding nodes in $\mathcal{G}$. This construction yields a *graph isomorphism* between $\mathcal{G}$ and its Euclidean embedding if and only if, for every pair of distinct nodes $v, w \in V(\mathcal{G})$, their corresponding feature vectors satisfy $x(v) \neq x(w)$. In practice, this condition can always be ensured by perturbing feature vectors with arbitrarily small amounts of Gaussian noise. Consequently, without loss of generality, we may restrict our attention to featured graphs that admit an isomorphic Euclidean embedding.

We now establish that ℓ -ECT serves as an expressive representation for graph learning, capturing both structural and feature information.

Theorem 5.3.1. *Let $\mathcal{G}$ be a featured graph, and let $\{\ell\text{-ECT}_1(x; \mathcal{G})\}_x$ denote the collection of local Euler Characteristic Transforms with respect to the 1-hop neighborhoods in $\mathcal{G}$. Then, the set $\{\ell\text{-ECT}_1(x; \mathcal{G})\}_x$ encodes all the necessary (non-learnable) information required to perform a single message-passing step in a standard GNN framework. Specifically, for any given vertex $x \in \mathcal{G}$, one can reconstruct the feature vectors of all nodes within its 1-hop neighborhood solely from $\ell\text{-ECT}_1(x; \mathcal{G})$.*

Proof. By the remark before, we may assume that the natural embedding of $\mathcal{G}$ into $\mathbb{R}^n$ is a graph isomorphism. Then by the invertibility theorem, the 1-hop neighborhood of a point x in the embedding of $\mathcal{G}$ can be reconstructed from $\ell\text{-ECT}_1(x; \mathcal{G})$. Therefore, the feature vectors of x and its 1-hop neighbors can be deduced from $\ell\text{-ECT}_1(x; \mathcal{G})$, which is the only non-learnable information one needs to perform a message passing step. $\square$

Theorem 5.3.1 establishes that the collection $\{\ell\text{-ECT}_1(x; \mathcal{G})\}_x$ encodes sufficient information to perform a standard message-passing update on a featured graph. The advantage of leveraging ℓ -ECT representations over traditional message-passing techniques lies in their ability to incorporate higher-order structural and feature-based information. Specifically, extending the local representation to include $\{\ell\text{-ECT}_k(x; \mathcal{G})\}_x$ for $k \geq 2$ allows us to encode information about progressively larger neighborhoods, capturing both topological and feature relationships beyond direct neighbors. Such information is usually only implicitly available in message-passing based architectures since message passing diffuses neighboring information consecutively, for each message passing step (i.e. each layer) which is performed.

Unlike standard message-passing approaches, which require multiple iterative steps to propagate information beyond immediate neighbors, ℓ -ECT naturally encodes multi-scale neighborhood information in a single representation. This property is particularly advantageous for graph learning tasks where long-range dependencies play a crucial role, as it circumvents the inherent limitations of iterative aggregation in GNNs, where distant node information is progressively diluted. As a result, ℓ -ECT-based representations provide a principled and efficient alternative for learning structured data in graphs, while preserving the ability to extract meaningful topological and feature relationships.

5.3.4. Subgraph counting

Chen et al. [40] investigate the expressive power of message-passing neural networks in the context of subgraph counting. Specifically, their results establish that message-passing neural networks are inherently limited in their ability to count induced subgraphs. In particular, they prove that message-passing architectures *cannot* correctly compute the number of occurrences of any connected substructure consisting of three or more nodes as an induced subgraph. This limitation arises due

5. Local Topology and Graph Data

to the aggregation-based nature of message passing, which prevents the network from distinguishing between certain structural configurations at a local level.

By contrast, we demonstrate that Euler Characteristic Transforms (ECT) for featured graphs, as well as their local variants, provide a fundamentally more expressive representation that enables subgraph counting. To formalize this claim, we first introduce the necessary concepts.

Definition 5.3.2. *Two featured graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be isomorphic if there exists a bijection $\pi : V(\mathcal{G}_1) \rightarrow V(\mathcal{G}_2)$ such that an edge $(v, w) \in E(\mathcal{G}_1)$ exists if and only if $(\pi(v), \pi(w)) \in E(\mathcal{G}_2)$, and such that for all $v \in V(\mathcal{G}_1)$, the feature vectors satisfy $x(v) = x(\pi(v))$.*

A featured graph $\mathcal{G}_S$ is called a *subgraph* of $\mathcal{G}$ if its node set satisfies $V(\mathcal{G}_S) \subseteq V(\mathcal{G})$ and its edge set satisfies $E(\mathcal{G}_S) \subseteq E(\mathcal{G})$, with the additional requirement that the respective node features remain unaltered under the induced embedding. Furthermore, $\mathcal{G}_S$ is an *induced subgraph* of $\mathcal{G}$ if, in addition, the edge set is defined as $E(\mathcal{G}_S) = E(\mathcal{G}) \cap (V(\mathcal{G}_S) \times V(\mathcal{G}_S))$, meaning that $\mathcal{G}_S$ contains all edges between its nodes that exist in $\mathcal{G}$.

Given a featured graph $\mathcal{G}$ and a candidate subgraph $\mathcal{G}_S$, we define two key counting functions:

- $C_{Sub}(\mathcal{G}; \mathcal{G}_S)$ denotes the total number of subgraphs of $\mathcal{G}$ that are isomorphic to $\mathcal{G}_S$.
- $C_{Ind}(\mathcal{G}; \mathcal{G}_S)$ denotes the total number of induced subgraphs of $\mathcal{G}$ that are isomorphic to $\mathcal{G}_S$.

We now establish a fundamental result linking ECT representations to featured graph isomorphism.

Theorem 5.3.2. *Two featured graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are isomorphic if and only if their Euler Characteristic Transforms are equal, i.e.,*

$$\text{ECT}(\mathcal{G}_1) = \text{ECT}(\mathcal{G}_2).$$

Proof. When two featured graphs are isomorphic in the sense of Def.5.3.2, their respective Euclidean embeddings produce equal ECTs by construction since the node feature vectors of two corresponding points under the isomorphism are equal. On the other hand, let us assume that $\text{ECT}(\mathcal{G}_1) = \text{ECT}(\mathcal{G}_2)$. Then by the invertibility theorem, the Euclidean embeddings of $\mathcal{G}_1$ and $\mathcal{G}_2$ are equal. Therefore, the only information that may tell apart the two graphs are their node labels, but this means that $\mathcal{G}_1$ and $\mathcal{G}_2$ are isomorphic. $\square$

A direct consequence of Theorem 5.3.2 is the following corollary:

Corollary 5.3.1. *Euler Characteristic Transforms (ECT) can be used to perform subgraph counting.*

This result highlights a crucial distinction between ECT-based methods and traditional message-passing graph neural networks. While message passing fundamentally relies on local feature aggregation, which inherently limits its ability to count specific substructures, ECT encodes a richer structural and topological representation. As a result, ECT-based methods for graph representation learning are provably more expressive than message-passing approaches in the context of subgraph counting.

5.3.5. Rotation-invariant metric based on local ECTs

The aforementioned invariance properties of Euler Characteristic Transforms (ECT) with respect to translations and scalings naturally raise the question of whether local Euler Characteristic Transforms (ℓ -ECT) can be effectively used to compare the local neighborhoods of two distinct points in a simplicial complex. Ideally, a meaningful local comparison should be independent of the choice of the coordinate system. However, a fundamental limitation arises from the fact that ECT is sensitive to rotations: rotating the underlying simplicial complex misaligns the directional sampling in S^{n-1} , leading to a distortion in the computed transform. Consequently, this sensitivity presents a major obstruction to utilizing ℓ -ECT as a local similarity measure. To address this issue, we construct a *rotation-invariant metric* that allows for a meaningful comparison of local structures.

Let $X, Y \subset \mathbb{R}^n$ be two finite simplicial complexes. Since X and Y are finite, their Euler Characteristic Transforms, $\text{ECT}(X)$ and $\text{ECT}(Y)$, take on only finitely many values. This allows us to define a similarity measure $d_{\text{ECT}}(X, Y)$ between X and Y as follows:

$$d_{\text{ECT}}(X, Y) := \inf_{\rho \in \text{SO}(n)} \|\text{ECT}(X) - \text{ECT}(\rho Y)\|_{\infty}, \quad (5.4)$$

where $\text{SO}(n)$ denotes the special orthogonal group, representing all possible rotations in $\mathbb{R}^n$. This formulation ensures that $d_{\text{ECT}}(X, Y)$ captures structural similarity while remaining invariant to rotational transformations.

We now establish the mathematical validity of d_{ECT} by proving that it satisfies the formal definition of a metric.

Theorem 5.3.3. *The function d_{ECT} defines a metric on the collection of rotation equivalence classes of finite simplicial complexes embedded in $\mathbb{R}^n$.*

Proof. $d_{\text{ECT}}(X, X) = 0$ holds for ρ being the identity. Now assume that $d_{\text{ECT}}(X, Y) = 0$. Then there exists $\rho \in \text{SO}(n)$ with $\|\text{ECT}(X) - \text{ECT}(\rho Y)\|_{\infty} = 0$. As $\|\bullet\|_{\infty}$ is a norm, it follows that $\text{ECT}(X) = \text{ECT}(\rho Y)$, and by the invertibility theorem we obtain $X = \rho Y$. This shows the first property of a metric (note that positivity follows from $\|\bullet\|_{\infty}$). For symmetry, note that $\|\text{ECT}(X) - \text{ECT}(\rho Y)\|_{\infty} = \|\text{ECT}(\rho^{-1}X) - \text{ECT}(Y)\|_{\infty}$, since rotations are invertible. For the triangle inequality, let Z be another finite simplicial complex. $d_{\text{ECT}}(X, Z)$ then reads

5. Local Topology and Graph Data

$$\begin{aligned}
& \inf_{\rho \in SO(n)} \|(\text{ECT}(X) - \text{ECT}(\rho Z))\|_\infty \\
& \leq \inf_{\rho, \rho' \in SO(n)} \|(\text{ECT}(X) - \text{ECT}(\rho' Y))\|_\infty + \|(\text{ECT}(\rho' Y) - \text{ECT}(\rho Z))\|_\infty \\
& = \inf_{\rho, \rho' \in SO(n)} \|(\text{ECT}(X) - \text{ECT}(\rho' Y))\|_\infty + \|(\text{ECT}(Y) - \text{ECT}((\rho')^{-1} \rho Z))\|_\infty \\
& = \inf_{\rho \in SO(n)} \|(\text{ECT}(X) - \text{ECT}(\rho Y))\|_\infty + \inf_{\rho \in SO(n)} \|(\text{ECT}(Y) - \text{ECT}(\rho Z))\|_\infty
\end{aligned}$$

But this final term is precisely the definition of $d_{\text{ECT}}(X, Y) + d_{\text{ECT}}(Y, Z)$. $\square$

Theorem 5.3.3 ensures that d_{ECT} provides a well-defined and mathematically sound similarity measure that remains invariant under rotations. In particular, given a simplicial complex $X \subset \mathbb{R}^n$ and two points $x, y \in X$, we obtain a rotation-invariant measure for comparing their local neighborhoods by defining

$$d_{\text{ECT}}^k(x, y; X) := \inf_{\rho \in SO(n)} \|\ell\text{-ECT}_k(x; X) - \ell\text{-ECT}_k(y; \rho X)\|_\infty. \quad (5.5)$$

In practical applications, we approximate $d_{\text{ECT}}(X, Y)$ using a finite set of sampled directions $v_1, \dots, v_m \in S^{n-1}$ and threshold values $t_1, \dots, t_l \in \mathbb{R}$, leading to the empirical approximation

$$\inf_{\rho \in SO(n)} \|\overline{\text{ECT}}(X)_{(m,l)} - \overline{\text{ECT}}(\rho Y)_{(m,l)}\|_\infty. \quad (5.6)$$

A similar approximation is employed for the local version $d_{\text{ECT}}^k(x, y; X)$. Notably, as previously discussed, the approximated Euler Characteristic Transforms in Equation (5.6) admit a natural vector representation. Consequently, the $\|\bullet\|_\infty$ norm in Equation (5.6) corresponds to the maximum of the entry-wise absolute differences between the two respective representation vectors.

This formulation has significant computational advantages: the approximation in Equation (5.6) is efficiently computable for a wide range of data spaces, making it practical for large-scale applications. However, in our following experimental setup, we employ the Euclidean metric instead of the supremum norm for differentiability reasons, enabling the use of gradient-based optimization techniques.

5.4. Experiments

In this section, we present a series of experiments designed to empirically assess the performance of the local Euler Characteristic Transform ($\ell\text{-ECT}$) in the context of graph representation learning, with a particular focus on node classification tasks. Our primary objective is to evaluate the ability of $\ell\text{-ECT}$ -based representations to effectively capture structural information and compare their performance against traditional message-passing mechanisms. We pay special attention to scenarios

exhibiting high heterophily, where standard message-passing methods often struggle, although our evaluation also includes other settings for a comprehensive analysis.

To this end, we evaluate ℓ -ECT-based models against several widely used graph neural network architectures, including graph attention networks (GATs), graph convolutional networks (GCNs), and the graph isomorphism network (GIN). Additionally, we consider H2GCN [212], a model specifically designed to handle heterophilous graphs. By comparing ℓ -ECT-based representations with these architectures, we aim to highlight the advantages of our approach in learning expressive node embeddings that are robust to varying graph structures.

Furthermore, we demonstrate the practical applicability of the rotation-invariant metric introduced in Section 5.3 through a series of experiments on spatial alignment of graph data.

5.4.1. ℓ -ECTs in Graph Representation Learning

The theoretical link between message passing and local Euler Characteristic Transforms, as established in Theorem 5.3.1, motivates us to empirically validate the expressivity of ℓ -ECT representations in the context of node classification. Specifically, given a featured graph $\mathcal{G}$ and a fixed neighborhood scale $k \geq 0$, we assign the corresponding local Euler Characteristic Transform $\ell\text{-ECT}_k(x; \mathcal{G})$ to each node $v \in \mathcal{G}$. This representation, together with the respective node feature vector, serves as the input to a model for node classification.

For our experiments, we employ XGBoost as the classification model, as we found it to consistently outperform more complex models, such as neural networks, in this setting. However, we emphasize that the choice of the downstream model is flexible and can be determined by the user. It is important to clarify that our objective is not to establish a new state-of-the-art method for benchmarking graph datasets. Instead, we aim to demonstrate that an ℓ -ECT-based approach yields results that are *competitive with and often superior to* common message-passing-based graph learning techniques, such as graph attention networks (GATs) and graph convolutional networks (GCNs). Notably, our method exhibits strong performance across both homophilous and heterophilous graph structures, whereas message-passing approaches typically struggle in highly heterophilous settings.

One key advantage of ℓ -ECT-based representations is their effectiveness in scenarios where aggregating information from neighboring nodes is suboptimal or even detrimental. This is particularly relevant for graphs with high heterophily, where neighboring nodes often belong to different classes or exhibit vastly different structural properties. In such cases, ℓ -ECT-based models can significantly outperform message-passing methods, which rely on local aggregation and thus inherently assume that neighboring nodes share similar properties.

The fundamental distinction between ℓ -ECT and message-passing-based methods lies in how local graph information is incorporated. Traditional GNNs rely on a

5. Local Topology and Graph Data

discrete diffusion process, where node features are iteratively aggregated along the graph structure. While this mechanism is beneficial for many graph learning tasks, it can act as a major bottleneck when learning representations for tasks where node features should remain independent of their neighbors. In contrast, ℓ -ECT overcomes this limitation by directly encoding local graph information without requiring explicit diffusion through the network. This makes it particularly well-suited for tasks where local structural and feature-based information is of primary importance.

A second major advantage of ℓ -ECT-based approaches is their *model agnosticism*. Unlike message-passing-based architectures, which typically require specialized neural network models, ℓ -ECT-based representations can be utilized with a wide range of machine learning models, including tree-based methods and classical statistical approaches. This flexibility simplifies model selection and hyperparameter tuning, allowing practitioners to leverage their graph data without requiring deep expertise in GNN training and optimization.

As a consequence, ℓ -ECT facilitates *interpretability*, which is a critical aspect in domains where regulatory requirements impose transparency constraints on machine learning models. GNN-based approaches usually operate as black-box models, making it difficult to understand the decision-making process. In contrast, ℓ -ECT-based methods together with a tree-based learning algorithm naturally enable interpretability: Tree-based models, such as XGBoost, provide feature importance values that directly indicate which aspects of the ℓ -ECT representation contribute most to classification decisions. The individual entries of the ℓ -ECT vector correspond to specific directional evaluations in the calculation of ℓ -ECT, allowing for a spatial interpretation of how the model arrives at its predictions.

A detailed discussion on interpretability, along with an ablation study on the importance of different directions in ℓ -ECT, is presented in Appendix B.1. In particular, we show that removing directional components from ℓ -ECT representations generally leads to a decline in performance, further underscoring the expressivity of ℓ -ECT in node classification tasks.

Finally, we highlight another practical benefit of ℓ -ECT-based methods: they enable the use of models that do not require a separate validation split within the dataset. Many deep learning-based approaches, including GNNs, necessitate a validation set for tuning hyperparameters, which reduces the amount of data available for training and testing. By contrast, ℓ -ECT-based methods can leverage models that do not require explicit validation, thus making a larger portion of the dataset available for learning, which can be particularly advantageous when dealing with limited data.

5.4.2. Implementation details

For our experiments, we consider a featured graph $\mathcal{G}$, where each node $x \in V(\mathcal{G})$ is associated with a class label via an assignment $V(\mathcal{G}) \rightarrow \mathcal{Y}$, with $\mathcal{Y}$ denoting the set of possible classes in the underlying node classification task. Given a fixed

neighborhood scale parameter $k \geq 0$, we define $N_k(x; \mathcal{G})$ as the k -hop neighborhood of x in $\mathcal{G}$. The local Euler Characteristic Transform $\ell\text{-ECT}_k(x; \mathcal{G})$ is then approximated using its empirical counterpart, $\overline{\text{ECT}}(N_k(x; \mathcal{G}))_{(m,l)}$, based on a finite set of sampled directions and filtration steps, as described in Section 5.3.

In our experiments, we set the number of sampled directions and filtration steps to $m = l = 64$. However, these values can be treated as hyperparameters and adjusted in practice to optimize performance. The resulting $m \cdot l$ -dimensional vector representation $\overline{\text{ECT}}(N_k(x; \mathcal{G}))_{(m,l)}$, together with the original feature vector of node x , serves as input to the classification model.

For classification, we employ a simple XGBoost classifier without extensive hyperparameter tuning. This choice is intended to demonstrate the effectiveness of our method *out of the box*, without requiring complex model selection or optimization.

To ensure a fair comparison, we follow the experimental setup outlined by Platonov et al. [139] when implementing our baseline models. Specifically, we consider standard graph convolutional networks (GCNs) and graph attention networks (GATs), enhancing them with a two-layer multilayer perceptron (MLP) after each graph neighborhood aggregation layer, skip connections, as proposed by He et al. [81], and layer normalization, following Ba, Kiros, and Hinton [6]. For each run, the respective model is trained for 1000 epochs, and we report the test accuracy corresponding to the model state that achieves the highest validation accuracy during training. The performance of our baseline models is consistent with the results reported in Platonov et al. [139], confirming the reliability of our experimental setup.

5.4.3. WebKB datasets

We begin our experimental analysis with the WebKB datasets, originally introduced by Pei et al. [136]. These datasets, comprising *Cornell*, *Wisconsin*, and *Texas*, are known for their challenging classification tasks, particularly due to their heterophilous graph structures.

Our results indicate that the $\ell\text{-ECT}$ -based approach significantly outperforms the baseline GNN architectures across all three datasets. Notably, the best performance is achieved by combining both $\ell\text{-ECT}_1$ and $\ell\text{-ECT}_2$ for the *Cornell* and *Texas* datasets. However, for the *Wisconsin* dataset, using only $\ell\text{-ECT}_1$ yields the highest accuracy.

Interestingly, while the inclusion of $\ell\text{-ECT}_2$ provides a slight performance boost for *Cornell* and *Texas*, the marginal improvement suggests that one-hop neighborhood alone is already highly informative for these tasks. This finding underscores the expressivity of $\ell\text{-ECT}_1$ in capturing local structural and feature-based properties relevant to node classification.

A summary of our results is presented in Table 5.1.

5. Local Topology and Graph Data

Model	Cornell	Wisconsin	Texas
GCN	$45.0 \pm 2.2 \%$	$44.2 \pm 2.6 \%$	$47.3 \pm 1.5 \%$
GAT	$44.7 \pm 2.9 \%$	$48.2 \pm 2.0 \%$	$51.7 \pm 3.2 \%$
GIN	$46.5 \pm 3.1 \%$	$49.7 \pm 2.5 \%$	$54.2 \pm 2.9 \%$
H2GCN	$66.2 \pm 3.5 \%$	$70.2 \pm 2.3 \%$	$72.3 \pm 3.0 \%$
ℓ -ECT ₁	$66.8 \pm 4.2 \%$	$81.2 \pm 2.9 \%$	$74.6 \pm 0.5 \%$
ℓ -ECT ₂	$67.0 \pm 4.9 \%$	$76.1 \pm 2.8 \%$	$73.8 \pm 2.6 \%$
ℓ -ECT ₁ + ℓ -ECT ₂	$67.1 \pm 4.1 \%$	$78.5 \pm 2.6 \%$	$74.8 \pm 3.1 \%$

Table 5.1.: Performance (accuracy) of different graph learning models across multiple datasets, for 5 training runs each. The results for the ℓ -ECTs are with respect to a simple XGBoost classifier.

5.4.4. Heterophilous datasets

To further evaluate the effectiveness of our ℓ -ECT-based approach in heterophilic graph structures, we conduct experiments on the *Roman Empire* and *Amazon Ratings* datasets, both introduced in Platonov et al. [139]. These datasets are particularly relevant for benchmarking classification methods in the presence of heterophily, where neighboring nodes often belong to different classes.

Our results indicate that combining ℓ -ECT₁ and ℓ -ECT₂ consistently achieves the best performance across both datasets, significantly outperforming the baseline models (see Table 5.2). These findings align with the observations made in Platonov et al. [139], which report that specialized architectures such as H2GCN do not necessarily improve performance over standard graph neural network models in certain heterophilous settings.

Additionally, our analysis reveals a dataset-specific distinction: on the *Roman Empire* dataset, ℓ -ECT₁ outperforms ℓ -ECT₂, whereas the opposite holds for the *Amazon Ratings* dataset, where ℓ -ECT₂ provides better predictive performance. This suggests that one-hop neighborhood information is particularly informative for node classification in *Roman Empire*, while two-hop neighborhood structures play a more significant role in *Amazon Ratings*. These results further emphasize the flexibility of ℓ -ECT representations in adapting to different graph structures and classification tasks.

5.4.5. Amazon dataset

The Amazon dataset, originally introduced by Shchur et al. [168], consists of two benchmark datasets: *Computers* and *Photo*. These datasets provide a valuable testbed for evaluating graph-based learning methods in real-world settings.

Our results indicate that while the graph attention network (GAT) achieves the best performance on the *Computers* dataset, the combination of ℓ -ECT₁ and ℓ -ECT₂ outperforms all baselines on the *Photo* dataset (see Table 5.3). However, it is

Model	Roman Empire	Amazon Ratings
GCN	$73.3 \pm 0.8 \%$	$42.3 \pm 0.7 \%$
GAT	$76.4 \pm 1.2 \%$	$44.6 \pm 0.9 \%$
GIN	$56.8 \pm 1.0 \%$	$44.1 \pm 0.8 \%$
H2GCN	$64.2 \pm 0.9 \%$	$40.1 \pm 0.7 \%$
ℓ -ECT ₁	$80.4 \pm 0.4 \%$	$48.4 \pm 0.3 \%$
ℓ -ECT ₂	$78.0 \pm 0.3 \%$	$49.6 \pm 0.3 \%$
ℓ -ECT ₁ + ℓ -ECT ₂	$81.1 \pm 0.4 \%$	$49.8 \pm 0.3 \%$

Table 5.2.: Performance (in accuracy) of different graph learning models across multiple datasets, for 5 training runs each. The results for the ℓ -ECTs are with respect to a simple XGBoost classifier.

Model	Computers	Photo
GCN	$91.6 \pm 1.6 \%$	$93.6 \pm 1.7 \%$
GAT	$92.4 \pm 1.3 \%$	$94.8 \pm 1.1 \%$
GIN	$55.9 \pm 1.5 \%$	$82.2 \pm 1.3 \%$
H2GCN	$84.5 \pm 1.4 \%$	$92.8 \pm 1.2 \%$
ℓ -ECT ₁	$89.6 \pm 0.3 \%$	$94.1 \pm 0.3 \%$
ℓ -ECT ₂	$90.1 \pm 0.5 \%$	$94.4 \pm 0.7 \%$
ℓ -ECT ₁ + ℓ -ECT ₂	$92.2 \pm 0.6 \%$	$94.9 \pm 0.6 \%$

Table 5.3.: Performance (in accuracy) of different graph learning models across multiple datasets, for 5 training runs each. The results for the ℓ -ECTs are with respect to a simple XGBoost classifier.

5. Local Topology and Graph Data

Model	Actor	Squirrel	Chameleon
GCN	$30.7 \pm 2.1 \%$	$28.9 \pm 1.4 \%$	$42.8 \pm 1.8 \%$
GAT	$31.1 \pm 1.8 \%$	$31.8 \pm 1.3 \%$	$47.3 \pm 1.3 \%$
GIN	$26.5 \pm 2.0 \%$	$35.4 \pm 1.5 \%$	$43.1 \pm 1.7 \%$
H2GCN	$30.7 \pm 1.9 \%$	$40.8 \pm 1.4 \%$	$62.7 \pm 1.6 \%$
ℓ -ECT ₁	$31.4 \pm 1.9 \%$	$35.6 \pm 0.7 \%$	$43.46 \pm 1.7 \%$
ℓ -ECT ₂	$30.1 \pm 1.3 \%$	$35.6 \pm 0.8 \%$	$40.44 \pm 1.5 \%$
ℓ -ECT ₁ + ℓ -ECT ₂	$30.9 \pm 0.7 \%$	$35.3 \pm 1.5 \%$	$43.90 \pm 0.7 \%$

Table 5.4.: Performance (in accuracy) of different graph learning models across multiple datasets, for 5 training runs each. The results for the ℓ -ECTs are with respect to a simple XGBoost classifier. The maximum accuracy for each dataset is highlighted in bold.

important to note that the performance differences between these two models are relatively small, suggesting that both approaches are comparably effective on these datasets.

These results highlight once more that ℓ -ECT-based representations remain competitive with state-of-the-art graph learning models while providing additional advantages in terms of interpretability and flexibility, as discussed in previous sections.

5.4.6. Actor dataset

For the *Actor* dataset, introduced by Pei et al. [136], we observe that the ℓ -ECT₁ model achieves the highest accuracy, outperforming all other baseline models. Interestingly, the combination of ℓ -ECT₁ and ℓ -ECT₂ performs slightly worse, while ℓ -ECT₂ alone yields the lowest accuracy on this dataset. This suggests that one-hop neighborhood information (ℓ -ECT₁) is particularly informative for the *Actor* dataset, while the inclusion of two-hop neighborhood information (ℓ -ECT₂) introduces redundancy or noise that does not contribute positively to classification performance.

5.4.7. WikipediaNetwork dataset

The *WikipediaNetwork* dataset, introduced by Rozemberczki, Allen, and Sarkar [160], consists of two node classification tasks: *Squirrel* and *Chameleon*. These datasets are known for their high heterophily, making them challenging for traditional message-passing-based graph learning methods.

On both *Squirrel* and *Chameleon*, the heterophily-specific architecture H2GCN achieves the best performance. However, our results show that ℓ -ECT₁ and the combination of ℓ -ECT₁ and ℓ -ECT₂ achieve accuracies that are comparable to or even better than all other standard baselines. This highlights the surprising effectiveness of our ℓ -ECT-based approach, even without extensive hyperparameter

Model	Cora	CiteSeer	PubMed
GCN	$88.1 \pm 1.2\%$	$74.6 \pm 1.5\%$	$85.3 \pm 4.7\%$
GAT	$88.3 \pm 1.1\%$	$75.3 \pm 1.5\%$	$85.7 \pm 4.2\%$
H2GCN	$85.4 \pm 1.6\%$	$72.4 \pm 1.9\%$	$86.4 \pm 0.5\%$
GIN	$85.0 \pm 1.5\%$	$72.2 \pm 1.7\%$	$87.0 \pm 0.5\%$
ℓ -ECT ₁	$87.6 \pm 0.6\%$	$72.1 \pm 0.6\%$	$90.2 \pm 0.5\%$
ℓ -ECT ₂	$87.2 \pm 0.7\%$	$72.3 \pm 0.8\%$	$90.3 \pm 0.5\%$
ℓ -ECT ₁ + ℓ -ECT ₂	$87.8 \pm 0.6\%$	$72.5 \pm 0.7\%$	$90.3 \pm 0.5\%$

Table 5.5.: Performance (accuracy) of different graph learning models across datasets, based on 5 training runs. Results for ℓ -ECTs are derived using a simple XGBoost classifier.

tuning, demonstrating its potential as a competitive alternative to message-passing architectures in heterophilous graph settings.

5.4.8. Planetoid datasets

To further evaluate the effectiveness of our ℓ -ECT-based approach, we consider the well-known *Planetoid* datasets, originally introduced by Yang, Cohen, and Salakhudinov [201]. This benchmark consists of three citation networks: *Cora*, *CiteSeer*, and *PubMed*, which are widely used for evaluating graph-based node classification methods.

For our experiments, we trained all models using a random 75/25 split of the data. The results are summarized in Table 5.5. While GCN and GAT achieve slightly better performance on *Cora* and *CiteSeer*, the performance gap between these models and our ℓ -ECT-based approach is surprisingly small. Notably, on the *PubMed* dataset, ℓ -ECT-based models outperform both GCN and GAT, further demonstrating their ability to effectively capture relevant structural information.

These findings suggest that the expressivity of ℓ -ECT, as formally established in Section 5.3, is not merely of theoretical interest but also translates into practical advantages. The results highlight that ℓ -ECT-based representations provide a viable alternative for graph-based learning tasks, offering a model-agnostic approach that is not constrained by specific neural network architectures. Additionally, the interpretability of ℓ -ECT-based methods makes them particularly appealing for applications where understanding the decision-making process is crucial.

5.4.9. Post-hoc evaluation

We now present an evaluation of the experimental results that we have shown before. Our analysis indicates that ℓ -ECT-based approaches offer a versatile, general-purpose solution for node classification, demonstrating their capability to effectively handle heterophilic datasets without requiring additional customization or specialized architectures.

5. Local Topology and Graph Data

To systematically compare the performance of various models across both homophilic and heterophilic graph datasets, we employed *critical difference diagrams*.

A *critical difference diagram* provides a visual representation of the average ranks of multiple models across a set of datasets, facilitating an overall comparison of model performances. Figure B.3 presents the critical difference diagram summarizing the results for all node classification tasks discussed in Section 5.3, covering both homophilic and heterophilic graph datasets.

Our results clearly indicate that ℓ -ECT-based approaches outperform standard graph learning methods as well as heterophily-specific architectures, such as H2GCN, when averaged over all datasets. Notably, the best-performing model, ℓ -ECT₁ + ℓ -ECT₂, achieves an average rank of 2, while the lowest-ranked model, GIN, has an average rank of 5.7. Importantly, even the weakest-performing ℓ -ECT-based model (ℓ -ECT₂) surpasses the best-performing non- ℓ -ECT-based method, GAT. This finding underscores the broad applicability and superior performance of ℓ -ECT-based representations across diverse graph learning scenarios.

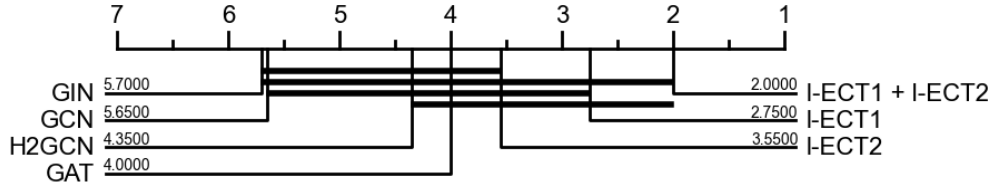


Figure 5.1.: Critical difference diagram depicting the ranking of different models across all node classification tasks in Section 5.4. Even the weakest-performing ℓ -ECT-based approach (ℓ -ECT₂) outperforms all other methods when averaged over all tasks. This figure was recreated from Rohrscheidt and Rieck [156].

To further evaluate the effectiveness of our approach in comparison to methods reported in the literature, we incorporate results from Platonov et al. [139]. In this comparison, we assess model rankings across heterophilic node classification tasks, as presented in Table B.3. Several of the listed models, such as H2GCN, CPGNN, and GPR-GNN, are explicitly designed to address the challenges posed by heterophilic graphs through specialized architectures. In contrast, our ℓ -ECT₁ + ℓ -ECT₂ approach remains a general-purpose method, yet it achieves a competitive rank of 11, matching the performance of GloGNN while outperforming widely adopted architectures such as GT and GAT by a considerable margin.

These results further emphasize the robustness and adaptability of ℓ -ECT-based methods, demonstrating their effectiveness across a wide range of graph benchmarks without requiring specific modifications for heterophilic settings. Combined with the findings illustrated in Figure B.3, this positions ℓ -ECT-based approaches as a highly versatile and powerful solution for node classification tasks.

Model	Rank
H2GCN	18.250
CPGNN	16.750
GPR-GNN	15.250
ResNet	13.750
ℓ -ECT ₁	12.375
ℓ -ECT ₂	12.375
GAT	12.250
GT	11.000
ℓ -ECT ₁ + ℓ -ECT ₂	11.000
GloGNN	11.000
ResNet+SGC	10.750
FAGCN	10.000
JacobiConv	9.750
GCN	9.625
GBK-GNN	9.000
ResNet+adj	7.250
SAGE	5.875
GAT-sep	5.500
GT-sep	5.250
FSGNN	3.000

Table 5.6.: Ranking of various models across heterophilic datasets as reported in Platonov et al. [139], compared to our proposed ℓ -ECT-based approach. This table was recreated from Rohrscheidt and Rieck [156].

5.4.10. Learning Spatial Alignment of Geometric Graphs

In the following, we employ the approach described in Section 5.3 to learn the spatial alignment of two data spaces by re-rotating one into the other. Our objective is to demonstrate that when two datasets differ solely by a rotational transformation, our method is capable of realigning them effectively. Furthermore, we establish that this alignment remains stable in the presence of noise, thereby validating its robustness as a tool for comparing local neighborhoods in data.

Compared to conventional spatial alignment techniques, such as the *iterative closest point* (ICP) algorithm, our approach offers a significant computational advantage. Unlike ICP, which requires the computation of all pairwise distances between points in the respective spaces, our method avoids this computational bottleneck. This is particularly advantageous for large datasets, where pairwise distance computations can be prohibitively expensive. Consequently, our ECT-based alignment procedure provides a more scalable and efficient alternative for spatial alignment tasks.

Additionally, we observe that this method is highly effective for aligning *embedded graph data*, making it particularly relevant for applications involving *geometric graphs*.

To formalize the spatial alignment task, we consider two embedded simplicial com-

5. Local Topology and Graph Data

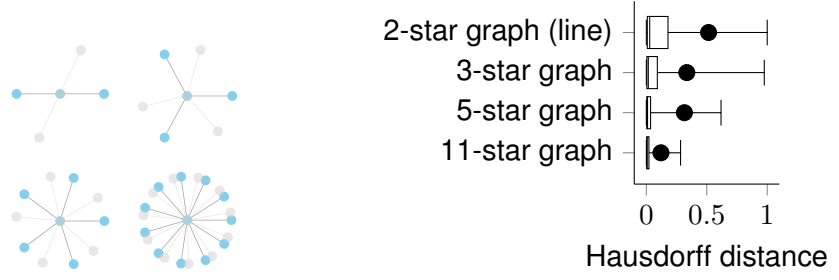


Figure 5.2.: A comparison of the Hausdorff distances of aligned graphs. The black dots represents the Hausdorff distance between the original graph and a randomly rotated version of itself. This figure was recreated from Rohrscheidt and Rieck [156].

plexes $X, Y \subset \mathbb{R}^n$. As established in Section 5.3, the metric properties of d_{ECT} ensure that

$$d_{\text{ECT}}(X, Y) = 0$$

if and only if X and Y differ solely by a rotation. We therefore approximate $d_{\text{ECT}}(X, Y)$ by solving the following learning problem:

$$\min_{\rho \in SO(n)} \|\overline{\text{ECT}}(X)_{(m,l)} - \overline{\text{ECT}}(\rho Y)_{(m,l)}\|_2^2, \quad (5.7)$$

where the approximation is computed using a set of sampled directions $v_1, \dots, v_m \in S^{n-1}$ and filtration steps $t_1, \dots, t_l \in \mathbb{R}$.

As discussed in Section 5.3, the ECT approximations take the form of vectorized representations, making it feasible to optimize the above learning problem using a general-purpose gradient-based optimization algorithm. The advantage of this formulation is twofold: (i) it provides a rotation-invariant loss function for measuring alignment quality, and (ii) it directly yields the optimal rotation parameters required for realigning Y with X . The learned transformation parameters thus represent the optimal rotation that minimizes the ECT-based discrepancy, offering both theoretical soundness and practical utility in data alignment tasks.

Geometric graphs serve as a compelling test case for the utility of our method, particularly in addressing the challenging problem of *graph re-alignment*. In this analysis, we focus on a specific type of geometric graph known as the *k-star graph*. A *k-star graph* is defined as a tree consisting of a single central node connected to k leaf nodes, representing a simple yet structurally significant configuration that encapsulates fundamental geometric properties (see Figure 5.2 for examples of different *k-star* graphs).

To embed the graph into a two-dimensional space, we assign each node a unique 2D vector while ensuring that all leaf nodes are equidistant from the central node, thereby preserving structural symmetry. To introduce variability and evaluate robustness, the embedded graph is then subjected to a random rotation in $\mathbb{R}^2$, simulating

realistic perturbations encountered in practical applications. The primary objective of this experiment is to recover the original orientation of the graph by learning the appropriate rotation matrix using the metric introduced in Section 5.3.

To assess the effectiveness of our approach, we measure the similarity between the original graph and its re-aligned counterpart using the *Hausdorff distance*, a widely used metric that quantifies the maximal deviation between two point sets. To ensure statistical significance, we repeat the learning procedure over 200 independent trials while maintaining consistent initializations for both the graph and the applied rotation matrix.

The results, summarized in Figure 5.2, clearly demonstrate the strength of our method. Specifically, the realignment process consistently achieves small Hausdorff distances, with median values close to zero, indicating near-perfect recovery of the original graph’s orientation. In contrast, the Hausdorff distances between the original graph and its perturbed version remain significantly larger, underscoring the effectiveness of our approach in mitigating rotational distortions.

It is well established that *equivariant graph neural networks* (GNNs) struggle to differentiate the orientation of rotationally symmetric structures [98]. By contrast, our method provides a natural mechanism for encoding rotation-sensitive graph representations while preserving the ability to recover rotation-invariant properties through alignment of the underlying ECT representations.

5.5. Discussion

In this chapter, we introduced the *Local Euler Characteristic Transform* (ℓ -ECT), presenting a novel approach to graph representation learning that preserves local structural information without relying on feature aggregation. Our method addresses fundamental limitations inherent to message-passing neural networks, particularly in scenarios where aggregating neighboring information is suboptimal—such as in heterophilous graph structures. By retaining essential local details, ℓ -ECT-based representations enable more nuanced and expressive embeddings, yielding significant advantages in node classification tasks.

A key strength of our approach is its model-agnostic nature, allowing ℓ -ECT representations to be effectively combined with a wide range of machine learning models, including interpretable approaches such as decision trees and gradient boosting. This flexibility makes ℓ -ECT-based methods particularly suitable for applications where transparency is essential, such as in healthcare, finance, and legal decision-making. Unlike conventional black-box neural networks, which often struggle to provide clear explanations for their predictions, ℓ -ECT-based models offer interpretability while maintaining expressive power, making them a promising alternative for domains with strict regulatory requirements.

Beyond graph representation learning, the ℓ -ECT framework has the potential for applications in areas where local structural information plays a crucial role. Fur-

5. Local Topology and Graph Data

thermore, the ability to extend ℓ -ECT to higher-dimensional simplicial complexes suggests promising applications in simplicial learning, where higher-order relationships in data can be analyzed in a computationally efficient manner.

While this study establishes the theoretical and empirical advantages of ℓ -ECT-based representations, several directions remain for future research. Developing more efficient algorithms for computing ℓ -ECT at scale is an important step toward applying these methods to massive graph datasets. Exploring hybrid approaches that balance local and global information more effectively could further enhance performance, potentially by integrating ℓ -ECT into message-passing neural networks. In this regard, the use of ℓ -ECT as a structural positional encoding within GNN architectures may provide additional benefits by allowing models to better capture spatial relationships in graph data. Moreover, improving sampling techniques for selecting filtration directions could optimize the trade-off between computational cost and representational fidelity.

Overall, the versatility and effectiveness of ℓ -ECT demonstrate its potential as a novel tool in graph representation learning and beyond. By improving its scalability, integrating it with existing deep learning frameworks, and refining its adaptability to complex data structures, ℓ -ECT has the potential to become a powerful asset in the field of geometric and topological machine learning.

6. Global Topology and Bayesian Inference

The methods presented in this chapter mark a significant departure from the preceding ones by shifting from a local to a global perspective on topological machine learning. While the previous chapters introduced local geometrical-topological approaches that focus on capturing fine-grained structure within small neighborhoods of data, the method introduced here leverages global topological properties in the context of Bayesian inference. This shift is motivated by the necessity of understanding large-scale structural patterns in data, which are often beyond the reach of purely local techniques.

Topological descriptors, such as persistence diagrams, inherently encode global information about the shape of data, making them particularly useful in scenarios where local geometric features alone are insufficient. This global perspective is essential for robust inference in cases where the underlying data structure exhibits significant complexity, as seen in generative models of fluid dynamics, collective motion, and percolation processes. Unlike local approaches, which primarily rely on small-scale interactions, the method presented in this chapter integrates topological data analysis into Bayesian inference, allowing for a more holistic understanding of data distributions.

This global perspective enables robust inference in settings where the overall shape of data, rather than its local geometry, plays a crucial role.

6.1. Introduction

Topological machine learning methods provide a framework for analyzing the ‘shape’ of data at multiple scales while maintaining robustness against various types of noise. This robustness and multi-scale perspective have led to the development of hybrid models that combine geometric and topological techniques across diverse application domains, including computer vision [93, 195], graph learning [89, 199, 209], time series analysis [204], and representation learning [32, 122]. The ability of topological descriptors to capture global structural properties of data has made them particularly useful in scenarios where classical geometric or statistical approaches fall short.

Despite their advantages, topological methods face significant challenges when integrated into Bayesian inference frameworks. Many real-world data analysis tasks

6. Global Topology and Bayesian Inference

require the ability to incorporate topological information within a Bayesian setting. Such requirements arise when the ‘shape’ of data provides crucial insights for inference, when uncertainty quantification is essential, or when data availability is too limited for deep learning-based approaches.

A Bayesian approach to inference typically assumes a prior distribution $p(\theta)$ over parameters and uses observed data y to update beliefs via Bayes’ theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \quad (6.1)$$

This formulation provides not only point estimates of parameters but also a full characterization of uncertainty. However, applying this approach to complex generative models presents notable difficulties. The likelihood function $p(y|\theta)$, and consequently the marginal likelihood $p(y) = \int p(y|\theta)p(\theta)d\theta$, are often analytically intractable.

Furthermore, the nature of many real-world problems requires inference techniques that are invariant to isometries (such as translations and rotations), a property that topological methods are well-suited to address [84]. To ensure a successful integration of topology into Bayesian inference, two key requirements must be met: (i) the inference framework must effectively incorporate topological descriptors from the observed data y , and (ii) it must provide a theoretically sound method for defining and updating posterior beliefs.

Both of these criteria are essential for overcoming the current limitations of topological machine learning in complex real-world scenarios. The importance of addressing these issues is particularly pronounced in settings where traditional neural network-based approaches are infeasible due to small data availability (e.g., expensive simulations) or the necessity of robust uncertainty quantification.

In the absence of an explicit likelihood function, we assume that the only access to parameter information comes from generative simulation models, represented by samples $x \sim p(x|\theta)$. Crucially, we do *not* assume that $p(x|\theta)$ is differentiable in θ , precluding the use of standard neural network-based inference techniques. This assumption is particularly relevant in modeling chaotic systems such as collective motion, where tipping points lead to highly non-continuous behavior.

Instead, we introduce a framework that leverages a loss function $\ell(x, y)$, which measures the discrepancy between generated samples x and observed data y . Under this setting, we define the posterior distribution as

$$\pi(\theta|y) = \frac{\int \exp(-\ell(y, x))p(x|\theta)p(\theta)dx}{\iint \exp(-\ell(y, x))p(x|\theta)p(\theta)dx d\theta}. \quad (6.2)$$

This posterior captures uncertainty within the generative model and provides a mechanism for parameter estimation without requiring explicit likelihood functions.

A fundamental challenge in this framework is the selection of a suitable loss function ℓ . Traditional approximate Bayesian computation (ABC) methods often rely on

summary statistics derived from data, but these are rarely sufficient for capturing all relevant information. On the other hand, purely geometric losses, such as the Hausdorff distance, may fail to preserve the necessary information for robust inference.

In this chapter, we (i) introduce a novel framework that integrates Bayesian inference with topological machine learning via efficient geometrical–topological loss terms, (ii) show that the resulting posterior distribution can be interpreted as a solution to a generalized Bayesian optimization problem, (iii) develop sampling algorithms that leverage topology-aware loss functions for Bayesian inference, and (iv) demonstrate the practical utility of our approach on multiple complex generative models.

By adopting a Bayesian perspective, our framework naturally incorporates prior knowledge, allowing for more interpretable and robust inference compared to black-box models. Our work contributes to the growing field of simulation-based inference by providing a principled way to integrate topological information into parameter estimation. The results of this chapter are essentially based on Rohrscheidt, Rieck, and Schmon [158].

6.2. Related Work

Uncertainty quantification in complex models with intractable likelihoods has been an active area of research, leading to the development of various Bayesian inference methods. These approaches seek to approximate the posterior distribution when direct computation is infeasible. While each method introduces a unique perspective on inference, they also come with inherent challenges, including computational tractability, robustness, and theoretical guarantees.

One widely used class of methods is *Approximate Bayesian Computation* (ABC) [176, 142, 10], which circumvents explicit likelihood evaluation by simulating data from the model and accepting parameter values that yield sufficiently similar synthetic observations to real data. The degree of similarity is determined by a predefined discrepancy function and an acceptance threshold.

A key limitation of ABC is the sensitivity to the choice of the rejection threshold. A too-stringent threshold leads to prohibitively high computational costs, while an overly permissive threshold can result in biased posterior estimates with incorrect credible regions [75]. Moreover, rejection-based ABC methods exhibit non-standard asymptotic behavior in the presence of model misspecification [74].

To improve the efficiency and stability of ABC, Wasserstein ABC [18] introduces optimal transport distances as a metric between observed and simulated data. However, while this approach refines the selection process, it still fundamentally relies on rejection sampling, which remains computationally expensive.

Several works have explored alternative formulations of Bayesian inference by constructing generalized posteriors based on discrepancy measures rather than like-

6. Global Topology and Bayesian Inference

likelihood functions. Notable examples include posterior constructions using *maximum mean discrepancy* (MMD) [134, 41], which replace explicit likelihood computation with kernel-based similarity measures. However, these methods rely on theoretical losses of the form $\ell(y, \theta) = E_{p(x|\theta)}[\ell(y, x)]$, which must be approximated using empirical samples, potentially introducing bias and numerical instability.

An alternative approach involves discrepancy-based Bayesian inference, where robustness is achieved by treating observed data as a coarsened version of latent data. Miller and Dunson [119] introduced a Bayesian coarsening framework that enhances robustness by broadening the data-generating process rather than relying on pointwise likelihood evaluations.

Topological Data Analysis (TDA) has recently gained traction for analyzing generative models by capturing global structural properties of data. Prior work has primarily focused on parameter inference for pre-defined probabilistic models of persistence diagrams [114, 129], demonstrating the expressive power of TDA in complex systems.

More directly relevant to our approach, Topaz, Ziegelmeier, and Halverson [179] employed persistent homology to qualitatively analyze swarming models, such as the Vicsek model, revealing topological patterns in emergent collective behavior. However, their analysis remained descriptive rather than inferential.

A closer approach to ours is that of Thorne, Kirk, and Harrington [178], who incorporated topological summary statistics into ABC via rejection sampling. While this method represents one of the first integrations of TDA within the ABC framework, it suffers from multiple parameter choices, including the selection of topological descriptors and rejection thresholds, which require domain expertise.

Our method advances the intersection of Bayesian inference and TDA by directly integrating topological information into posterior construction without relying on rejection sampling or empirical approximations of theoretical losses. Unlike previous TDA-based ABC methods, our approach operates on full persistence information rather than predefined summary statistics, thereby preserving richer structural characteristics of data. Moreover, by formulating inference through a loss-based posterior, we provide a more theoretically grounded framework that mitigates sensitivity to arbitrary parameter choices inherent in rejection-based methods.

6.3. Methods

6.3.1. Comparison-Based Posteriors

Within the Bayesian framework, unknown quantities, such as parameters denoted by θ , are endowed with a probability distribution $p(\theta)$, commonly referred to as the *prior distribution*. This prior encodes assumptions or domain knowledge about the parameters before any data is observed. The process of incorporating observed

data y to update beliefs about θ is formalized through Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \quad (6.3)$$

where $p(y|\theta)$ is the likelihood function, quantifying how probable the observed data is under a given parameter setting, and $p(y)$ is the evidence, which serves as a normalizing constant ensuring that the posterior remains a valid probability distribution.

However, a key challenge in practical applications of Bayesian inference arises when the likelihood function $p(y|\theta)$ is either *unknown* or *intractable*, making exact inference computationally prohibitive. This difficulty is particularly pronounced in complex domains where the data exhibits intricate geometrical and topological structures, making the likelihood analytically unaccessible.

To circumvent this issue, alternative formulations of Bayesian inference have been proposed that relax the reliance on an explicit likelihood function. One such approach involves the concept of *generalized Bayesian inference*, which replaces the traditional likelihood function with a more flexible loss function [20]. This perspective is motivated by viewing Bayesian inference as the solution to an optimization problem, as established in prior work on decision-theoretic foundations of Bayesian methods [47, 56, 203].

Specifically, Bayesian inference can be cast as the minimization of a functional that balances two competing objectives: (i) fidelity to the data, represented through the expected log-likelihood, which we seek to maximize, and (ii) adherence to prior knowledge, encoded via the Kullback–Leibler (KL) divergence between the prior and the posterior candidate $q(\theta)$. Mathematically, this optimization problem is given by:

$$\arg \min_{q \in \mathcal{P}(\Theta)} \{ \mathbb{E}_q[-\log p(y|\theta)] + \text{KL}(q(\theta) \parallel p(\theta)) \}, \quad (6.4)$$

where the solution recovers the classical Bayesian posterior $q^*(\theta) = p(\theta|y)$.

Recent advances have generalized this optimization-based perspective, demonstrating that coherent posterior updates can be achieved even when replacing the negative log-likelihood $-\log p(y|\theta)$ with an arbitrary loss function $l(y, \theta)$ [20, 103]. This leads to the generalized Bayesian posterior:

$$\arg \min_{q \in \mathcal{P}(\Theta)} \{ \mathbb{E}_q[l(y, \theta)] + \text{KL}(q(\theta) \parallel p(\theta)) \}, \quad (6.5)$$

with the corresponding solution given by:

$$q^*(\theta) = \frac{\exp(-l(y, \theta))p(\theta)}{\int \exp(-l(y, \theta))p(\theta)d\theta} =: \pi_{\text{GBI}}(\theta | y). \quad (6.6)$$

A fundamental limitation of this generalization is that the loss function $l(y, \theta)$ still requires an explicit mapping between observations y and parameters θ , which may not be readily available in many practical scenarios. To address this, an

6. Global Topology and Bayesian Inference

alternative strategy involves leveraging expectations over simulated data drawn from a generative model. In particular, given a probabilistic model $p(x | \theta)$ that generates synthetic observations x conditioned on θ , we define the loss function as:

$$l(y, \theta) := \mathbb{E}_{p(x | \theta)}[\ell(y, x)] \approx \frac{1}{N} \sum_{i=1}^N \ell(y, x_i), \quad (6.7)$$

where $x_i \sim p(x | \theta)$ are sampled observations, and $\ell(y, x)$ is a discrepancy function measuring the divergence between the observed data y and the generated sample x .

Replacing $l(y, \theta)$ with a sample-based approximation, however, introduces a fundamental challenge: due to the non-linearity of the exponential function, Jensen's inequality implies that the resulting posterior distribution deviates from the idealized formulation in Eq. (6.6). This discrepancy can lead to systematic biases in the inferred posterior distribution.

In the following, we demonstrate that a principled construction of a valid posterior can still be achieved by directly incorporating the loss function $\ell(y, x)$ into the Bayesian update procedure. This yields an alternative Bayesian inference framework that is both computationally feasible and theoretically well-founded.

Uncertainty estimates based on distance functions have played a fundamental role in Bayesian statistics, particularly as a means of *approximating* posterior distributions. A notable example is *approximate Bayesian computation* (ABC) [176, 142, 10], where the likelihood function is replaced with a similarity-based criterion.

In this section, we demonstrate that a modification of this approach—specifically, a generalized posterior as proposed by Schmon, Cannon, and Knoblauch [166]—can yield a theoretically rigorous posterior update mechanism. The central idea involves leveraging a *loss function* $\ell(y, x)$ to construct a joint posterior that incorporates both simulated data $x \sim p(x | \theta)$ and real observations y . This leads to the joint posterior distribution:

$$\pi(\theta, x | y) \propto \exp(-\ell(y, x)) p(x | \theta) p(\theta). \quad (6.8)$$

Despite its appeal, this formulation presents two primary challenges. First, it is not immediately clear whether Eq. (6.8) provides a meaningful quantification of uncertainty. Second, since x represents *simulated* rather than *observed* data, marginalization over possible values of x is required, leading to the generalized posterior:

$$\pi(\theta | y) = \frac{\int_{\mathbf{X}} \exp(-\ell(y, x)) p(x | \theta) p(\theta) dx}{\iint_{\mathbf{X} \times \Theta} \exp(-\ell(y, x)) p(x | \theta) p(\theta) dx d\theta}. \quad (6.9)$$

We establish below that Eq. (6.9) constitutes a valid (generalized) posterior under an optimization-based formulation. In particular, the following theorem formalizes this result:

Theorem 6.3.1. *The comparison-based posterior*

$$\pi(\theta, x | y) = \frac{\exp(-\ell(y, x)) p(x | \theta) p(\theta)}{\int_{\mathbf{X}} \int_{\Theta} \exp(-\ell(y, x)) p(x | \theta) p(\theta) d\theta dx} \quad (6.10)$$

is the solution $q^*(\theta, x) = \pi(\theta, x | y)$ to the optimization problem:

$$\arg \min \{ \mathbb{E}_q[\ell(y, x)] + \text{KL}(q(\theta, x), p(x|\theta)p(\theta)) \}, \quad (6.11)$$

where the minimization is over all $q \in \mathcal{P}(\Theta \times \mathbf{X})$.

Proof.

$$\begin{aligned} q_B &= \arg \min_q \{ \mathbb{E}_q[\ell(y, x)] + \text{KL}(q(x, \theta), p(x|\theta)\pi(\theta)) \} \\ &= \arg \min_q \left\{ \iint \ell(y, x) q(x, \theta) dx d\theta + \iint \log \frac{q(x, \theta)}{p(x|\theta)\pi(\theta)} q(x, \theta) dx d\theta \right\} \\ &= \arg \min_q \left\{ \iint \log(\exp \ell(y, x)) q(x, \theta) dx d\theta + \iint \log \frac{q(x, \theta)}{p(x|\theta)\pi(\theta)} q(x, \theta) dx d\theta \right\} \\ &= \arg \min_q \left\{ \iint \log \left(\frac{q(x, \theta)}{p(x|\theta)\pi(\theta) \exp(-\ell(y, x))} \right) q(x, \theta) dx d\theta \right\} \\ &= \arg \min_q \left\{ \iint \log \left(\frac{q(x, \theta)}{p(x|\theta)\pi(\theta) \exp(-\ell(y, x)) Z^{-1}} \right) q(x, \theta) dx d\theta \right\} - \log Z \\ &= \arg \min_q \text{KL} \{ q(x, \theta), p(x|\theta)\pi(\theta) \exp(-\ell(y, x)) Z^{-1} \}. \end{aligned}$$

Hence, the minimiser q is such that

$$q(x, \theta) = \frac{p(x | \theta) \pi(\theta) \exp(-\ell(y, x))}{Z}, \quad q(\theta) = \int q(x, \theta) dx$$

and

$$Z := \iint \exp(-\ell(y, x)) p(x | \theta) \pi(\theta) dx d\theta.$$

□

In Eq. (6.11), the first term quantifies the expected loss under the posterior, where the loss function $\ell(y, x)$ measures the discrepancy between observed and simulated data. This flexibility allows for the integration of domain-specific knowledge, such as topological constraints.

The second term in Eq. (6.11) enforces proximity to the prior, akin to standard Bayesian inference, with the key distinction that $p(\theta)$ in conjunction with the generative model $p(x | \theta)$ naturally encodes prior beliefs about data structure. Thus, the KL divergence term ensures consistency between the joint prior and the inferred posterior.

Importantly, Theorem 6.3.1 demonstrates that the comparison-based posterior of Eq. (6.8) is a principled solution to an optimization-based formulation, analogous to the original optimization problem. This provides a strong justification for employing

6. Global Topology and Bayesian Inference

comparison-based posteriors as a form of approximate inference, particularly when a well-defined loss function is available.

While distance-based loss functions have been previously used in settings where parameters are indirectly inferred via simulation, they are often regarded as approximations of an idealized loss:

$$\ell(y, \theta) := \lim_{n \rightarrow \infty} \sum_{i=1}^n \ell(y, x_i), \quad (6.12)$$

where $x_i \sim p(x | \theta)$. Our result, however, establishes that comparison-based posteriors should not merely be viewed as heuristic approximations, but rather as valid generalized Bayesian posteriors in their own right.

Consequently, our approach generalizes classical Bayesian inference by allowing for more flexible modeling of the data generation process. Instead of assuming that observed data is generated precisely according to $p(y | \theta)$, we accommodate settings where the data originates from a distribution that produces values *close* to y in terms of the loss function ℓ . This enhances robustness to model misspecification and makes our framework particularly well-suited for data with complex topological properties.

6.3.2. Inference for Comparison-Based Posteriors

Evaluating posteriors such as those given by Eq. (6.9) analytically is typically *infeasible* due to the high-dimensional integrations involved. One potential alternative is to directly target a variational distribution, as outlined in Eq. (6.11). However, this approach encounters significant challenges.

The first challenge arises from the intractability of the distribution $p(x | \theta)$, which complicates the computation of the KL divergence term in the variational objective. Since $p(x | \theta)$ is often a complex generative process, estimating the KL divergence reliably can be difficult, particularly when the likelihood function is unavailable in closed form.

The second challenge pertains to the computation of gradients. In variational inference, optimizing the variational distribution typically requires propagating gradients through the underlying simulation model. However, many complex simulation models are either non-differentiable with respect to their parameters or computationally expensive, thereby making gradient-based optimization impractical. The high computational cost of simulation-based methods imposes practical limitations. Many simulation models require substantial computational resources to generate a single data point, making large-scale data generation infeasible. This significantly limits the applicability of gradient-based optimization techniques in practice.

To address these issues, Monte Carlo sampling techniques provide a viable alternative. Specifically, methods such as self-normalized importance sampling and pseudo-marginal Markov chain Monte Carlo (MCMC) [5] enable the estimation of

Algorithm 1 Importance Sampling Estimation of the Mean**Require:** Observed data y , test function h **Ensure:** Estimate $\hat{h}(\theta)$

```

1: for  $i \leftarrow 1$  to  $n$  do
2:   Sample  $\theta_i \sim p(\theta)$ 
3:   Sample  $x_i \sim p(x | \theta_i)$ 
4: end for
5: Compute:

```

$$\hat{h}(\theta) \leftarrow \frac{\sum_{i=1}^n h(\theta_i) e^{-w\ell(y, x_i)}}{\sum_{i=1}^n e^{-w\ell(y, x_i)}}$$

```

return  $\hat{h}(\theta)$ 

```

expectations with respect to posteriors of the form in Eq. (6.9) without requiring explicit likelihood evaluations.

Algorithm 1 outlines a self-normalized *importance sampling* procedure. Under mild regularity conditions [e.g. 132, Theorem 9.2], almost-sure convergence can be established for a test function of interest h , yielding the approximation:

$$\frac{\sum_{i=1}^n h(\theta_i) e^{-\ell(y, x_i)}}{\sum_{i=1}^n e^{-\ell(y, x_i)}} \quad (6.13)$$

which converges to

$$\frac{\iint h(\theta) e^{-\ell(y, x)} p(x | \theta) p(\theta) dx d\theta}{\iint e^{-\ell(y, x)} p(x | \theta) p(\theta) dx d\theta} = E_{\pi(\theta | y)}[h(\theta)] \quad (6.14)$$

as $n \rightarrow \infty$.

Another widely used approach is *Markov chain Monte Carlo* (MCMC), which constructs a Markov chain that converges to the target distribution in Eq. (6.9). Algorithmic details are provided in Algorithm 2. These MCMC-based methods are known to converge to the desired posterior distribution under mild regularity conditions [5, Theorem 1].

6.3.3. Losses Based on Geometry & Topology

Having introduced multiple inference algorithms suitable for our approach, we now discuss the construction of topology-based loss functions.

Persistence diagrams can be equipped with a metric structure via optimal transport. Given two persistence diagrams $\mathcal{D}$ and $\mathcal{D}'$ containing features of the same dimensionality, we recall that their p th *Wasserstein distance* is given by

$$W_p(\mathcal{D}, \mathcal{D}') := \left(\inf_{\eta: \mathcal{D} \rightarrow \mathcal{D}'} \sum_{x \in \mathcal{D}} \|x - \eta(x)\|_\infty^p \right)^{\frac{1}{p}}, \quad (6.15)$$

6. Global Topology and Bayesian Inference

Algorithm 2 MCMC Approximation of the Target

Require: Observed data y , start value θ_0 , proposal distribution $q(\vartheta \mid \theta)$

Ensure: Sequence $\theta = (\theta_{1:K})$

```

1: Sample  $X_0 \sim p(x \mid \theta_0)$ 
2: for  $i \leftarrow 1$  to  $K$  do
3:   Sample  $U \sim \text{Unif}[0, 1]$ 
4:   Sample  $\theta' \sim q(\cdot \mid \theta_{i-1})$ 
5:   Sample  $X' \sim p(x \mid \theta')$ 
6:   Compute:

$$a(\theta_{i-1}, \theta') \leftarrow \frac{e^{-\ell(y, X')} p(\theta')}{e^{-\ell(y, X_{i-1})} p(\theta_{i-1})} \frac{q(\theta_{i-1} \mid \theta')}{q(\theta' \mid \theta_{i-1})}$$

7:   if  $U \leq a(\theta_{i-1}, \theta')$  then
8:     Set  $\theta_i \leftarrow \theta'$ ,  $X_i \leftarrow X'$ 
9:   else
10:    Set  $\theta_i \leftarrow \theta_{i-1}$ ,  $X_i \leftarrow X_{i-1}$ 
11:   end if
12: end for
return  $\theta = (\theta_{1:K})$ 

```

where $\eta(\cdot)$ denotes a bijection. Since $\mathcal{D}$ and $\mathcal{D}'$ may have different cardinalities, we extend each diagram with an infinite set of diagonal points (τ, τ) of zero persistence. This ensures the existence of a suitable bijection $\eta(\cdot)$. Modern optimal transport algorithms provide efficient solutions to Eq. (6.15) [73].

While Eq. (6.15) defines a proper mathematical metric, alternative descriptors such as *Betti curves* [152] or *persistence images* [1] offer computationally efficient alternatives. Persistence images, in particular, enable the use of fast L_1 or L_2 distances while retaining topological structure. Although the L_2 distance between persistence images shares properties with the Wasserstein distance [1], it operates on a discretized representation, potentially leading to information loss. Another approach involves employing similarity kernels between persistence diagrams [107, 147], which, while not strict metrics, provide computationally efficient alternatives without requiring vectorization.

To integrate computational topology into Bayesian inference, we propose combining posterior updates with a *topological loss function*. Given observed data y and simulated data $x \sim p(x \mid \theta)$, let $\mathcal{D}_x$ and $\mathcal{D}_y$ be the corresponding persistence diagrams. We define the topological loss function $\ell_T: \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}$ as:

$$\ell_T(y, x) := W_p(\mathcal{D}_y, \mathcal{D}_x). \quad (6.16)$$

This formulation generalizes to different data modalities. For point clouds, we compute persistence diagrams using the Vietoris–Rips complex. For image data, we extract diagrams from *cubical complexes*, e.g. allowing applications to medical imaging such as MRI scans [154].

In our implementation, we use $p = 2$ for the Wasserstein metric unless stated other-

wise. We rely on `giotto-tda` [174], which employs an approximate auction algorithm with worst-case complexity $\mathcal{O}(n \cdot m \cdot \log(\epsilon^{-1}))$, where n and m are the number of points in the respective persistence diagrams and ϵ is a precision parameter.

The Wasserstein distance between persistence diagrams is *stable* in the sense that geometric closeness upper-bounds topological similarity. Given two point clouds Y, X of the same cardinality:

$$W_p(\mathcal{D}_Y, \mathcal{D}_X) \leq C_p W_p(Y, X), \quad (6.17)$$

where C_p is a constant depending on p , and $W_p(Y, X)$ is computed on the original point clouds [172]. Since $W_p(\cdot, \cdot) \leq W_{p'}(\cdot, \cdot)$ for $p' \leq p$, choosing a lower p improves stability, with $p = 2$ striking a balance between robustness and sensitivity.

Furthermore, our loss function ℓ_T is invariant under isometries [60] and remains stable under subsampling [122]. While geometric closeness implies topological similarity, the converse does not hold, justifying the use of topological losses for inference tasks where shape information is critical. We validate this empirically in the subsequent section, showing that topological losses enhance parameter estimation accuracy.

As a baseline, we introduce a geometric loss based on the *Hausdorff distance* between point clouds X and Y . Given a metric space (M, d) and two non-empty subsets $X, Y \subseteq M$, we recall that their Hausdorff distance (w.r.t. M) is given by:

$$\ell_G(X, Y) := \inf\{\epsilon \geq 0 \mid X \subset Y_\epsilon, Y \subset X_\epsilon\}, \quad (6.18)$$

where $X_\epsilon = \cup_{x \in X} \{m \in M; d(m, x) \leq \epsilon\}$ denotes the ϵ -thickening of X in M .

While the Hausdorff distance is computationally simpler, it lacks invariance properties and may be less informative for complex data structures. For $|X| = n$ and $|Y| = m$, its computational complexity is $\mathcal{O}(n \cdot m)$. As we demonstrate in our experiments, it is less effective than topological losses in data generation tasks involving intricate structures.

6.4. Experiments

We empirically validate the effectiveness of our proposed method through three primary experiments, comparing topological losses with their geometric counterparts based on the results from Section 6.3.

The first two experiments focus on parameter inference for chaotic simulation models, which inherently represent an inverse problem. These models serve as ideal test cases for topology-driven Bayesian inference methods, as they exhibit two fundamental challenges: (i) their likelihood functions are *analytically intractable* due to their complex dynamics, and (ii) they are *non-differentiable* owing to phenomena such as tipping points. These properties render both neural network-based approaches and conventional Bayesian inference methods ineffective.

6. Global Topology and Bayesian Inference

Our method overcomes these difficulties by requiring only a relatively small number of simulations (here, 250) to achieve reliable inference. This stands in stark contrast to neural network-based techniques, which generally demand substantially larger datasets for training. This characteristic makes our approach particularly advantageous in settings where simulations are computationally expensive. Each experiment is conducted using 100 observation points for inference, and we assume normal distributions for both the prior and the proposal distribution in the MCMC procedure. The weight parameter w in the importance sampling procedure is fixed at 10.

The third experiment employs a synthetic model to demonstrate the applicability of our method in imaging-based inference tasks. This experiment highlights the flexibility of our approach and its capability to extract meaningful topological features in the context of structured data, such as images.

Across all three experiments, we observe that topological losses consistently outperform their geometric counterparts, reinforcing the utility of our method for parameter estimation.

To further contextualize our results, we compare our approach against standard summary statistics commonly used in Approximate Bayesian Computation (ABC), as detailed in Chapter C. Notably, our topology-driven method demonstrates superior performance relative to traditional L_p -norms applied to summary statistics in the first two non-synthetic experiments. This finding underscores the practical benefits of our approach, particularly for practitioners seeking robust inference techniques in complex generative models.

As a consistency check, we apply our method to infer the radius parameters of two geometric objects: the 2-sphere and the 2-torus. The results of this inference process are summarized in Table C.1 in the appendix, further corroborating the reliability of our proposed approach.

6.4.1. Vicsek Model: Swarm Behaviour

The Vicsek model is a widely studied agent-based model designed to investigate collective motion phenomena such as swarming and flocking. The fundamental principle behind the model is that each agent aligns its direction with its nearest neighbors at each time step. This alignment is subsequently perturbed by an additional stochastic term, introducing an element of randomness to the dynamics. The updated alignment angle determines the direction of movement for the following time step, where all agents move with a constant, time-invariant speed.

Following Vicsek et al. [191], the alignment angle update for an individual agent i is given by:

$$\theta(t+1) = \langle \theta(t) \rangle_r + \Delta\theta, \quad (6.19)$$

where $\langle \theta(t) \rangle_r$ denotes the mean angle of all agents within a distance r of agent i , and $\Delta\theta$ is a noise term sampled uniformly from the interval $[-\eta/2, \eta/2]$ with $0 \leq \eta \leq 2\pi$.

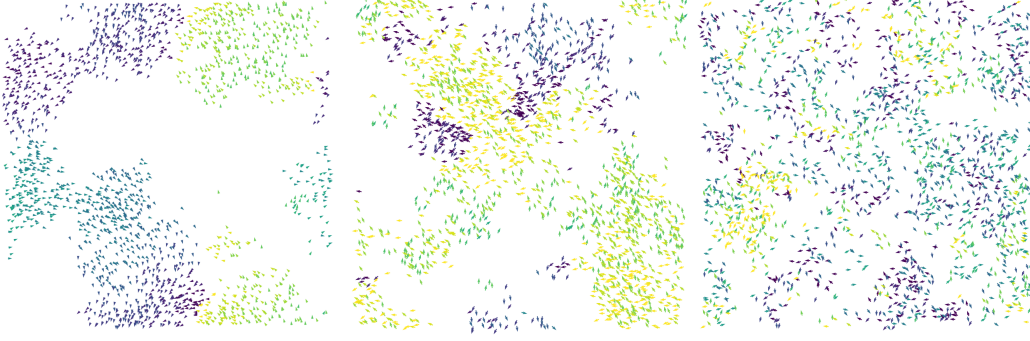


Figure 6.1.: Left to right: simulations of the Vicsek model for $\eta = 0.15, 0.3$ and 0.6 , respectively. For lower values of η swarming behaviour emerges, whereas for higher values of η swarms tend to merge and the model becomes more ergodic. This figure was recreated from Rohrscheidt, Rieck, and Schmon [158].

The agent's updated position is then obtained by moving at a constant speed in the direction of the newly computed alignment angle.

Despite the analytical expression for the position of each agent at any given time step, the complex interdependencies between agents result in emergent behaviors that cannot be captured in closed form. Consequently, the system exhibits highly intricate dynamics over time, rendering its likelihood function *intractable*. This necessitates a simulation-based inference approach, as conventional likelihood-based methods are not applicable in this setting.

To evaluate the effectiveness of topological losses in parameter inference when the likelihood function is *not analytically tractable*, we focus on estimating the noise parameter η after a given number of iterations (time steps) of the Vicsek model.

We distribute 2000 agents across a square domain of edge length $L = 25$, where opposite edges are identified while preserving orientation. As a result, the system evolves on the surface of a 2-torus.

For the inference of η , we conduct 250 simulations per sampling method, aiming to estimate η after 5, 10, and 50 iterations of the model, respectively, with each inference procedure being repeated 5 times for robustness. Given that the Vicsek model produces point cloud data, we compute Vietoris–Rips complexes and compare the associated persistence diagrams. As a baseline, we employ the Hausdorff distance between point clouds, with the corresponding results summarized in Fig. 6.2.

Our experiments demonstrate that the topological loss combined with the MCMC sampling procedure consistently outperforms alternative approaches across all tested settings, yielding highly accurate parameter estimates even after 50 time steps. In contrast, estimators derived using the Hausdorff loss exhibit significantly weaker performance, particularly in regimes approaching ergodicity (i.e., for larger values of η).

6. Global Topology and Bayesian Inference

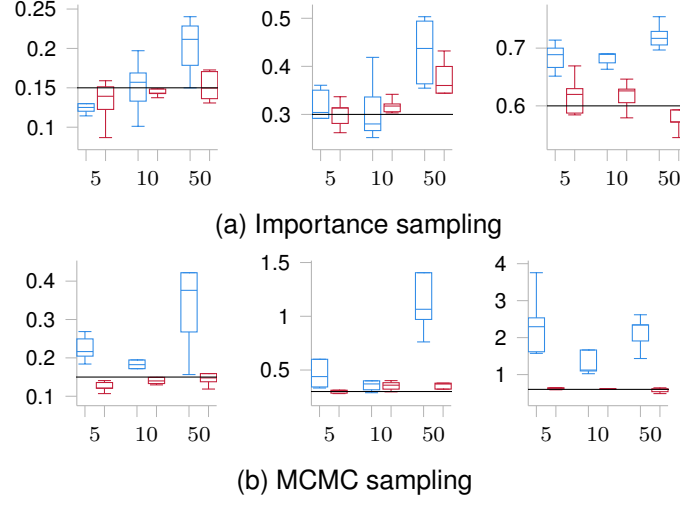


Figure 6.2.: Summary of parameter estimates for the noise parameter η of the Vicsek model, using 5 repetitions of the inference procedure. Each subplot depicts estimates for $\eta \in \{0.15, 0.30, 0.60\}$, respectively. The true parameter is shown as a solid black line. The horizontal axis shows the number of iterations of the Vicsek model. Estimates based on the [geometrical loss \$\ell_G\$](#) are less accurate than the [topological loss \$\ell_T\$](#) , in particular for larger number of iterations. This figure was recreated from Rohrscheidt, Rieck, and Schmon [158].

Finally, we note that our proposed topological loss formulation also surpasses standard rejection sampling techniques that rely on summary statistics, as detailed in Chapter C.

6.4.2. Fluid Dynamics: Lattice Boltzmann Simulation

Lattice Boltzmann methods (LBM) constitute a family of numerical algorithms designed for fluid simulation. Unlike conventional approaches that enforce the conservation of macroscopic properties in a dynamical system, LBM operates on a discrete lattice, where fluid behavior emerges through local propagation and collision processes applied to all particles simultaneously. For a comprehensive introduction, we refer to Krüger et al. [105].

In our setup, the initial state of each particle is initialized uniformly, followed by the addition of independent noise sampled from a normal distribution. This stochastic perturbation introduces variability into the system, allowing for the modeling of realistic fluid dynamics under uncertain conditions.

This portion of randomness is controlled by a global parameter $\eta \geq 0$, where higher values of η correspond to a higher portion of noise in the initialisation. The crucial observation is that the degree of turbulence in the system can be modelled by varying η , as is shown in Fig. 6.3. Our objective is therefore to infer the parameter η , which corresponds to a given observation after a certain amount of iterations in

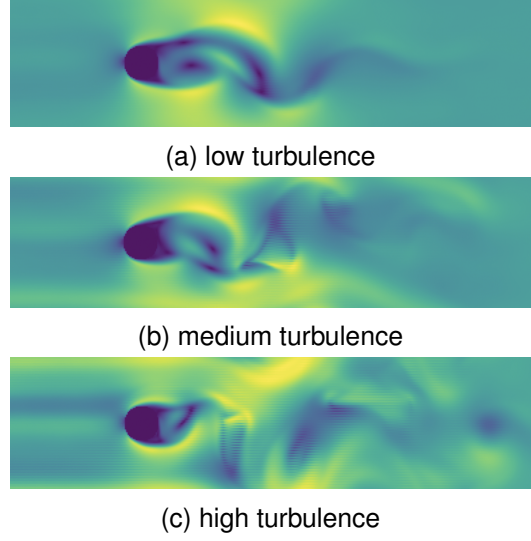


Figure 6.3.: Left to right: simulations of the fluid model after 3000 iterations, for $\eta = 0.2, 0.3$ and 0.4 , respectively. For lower values of η , the fluid flow is less turbulent than for higher values of η . This figure was recreated from Rohrscheidt, Rieck, and Schmon [158].

the model, where we fix this number of iterations to be $t = 3000$ time steps. As the topological loss, we use the Wasserstein distance between the persistence diagrams of the *cubical complexes* corresponding to the observed and the simulated state, respectively, while the Hausdorff distance between the states serves as a baseline comparison. Table 6.1 shows the results; we observe that the topological loss together with importance sampling outperforms the geometrical loss in both configurations, by far. Again, rejection sampling in combination with summary statistics performs less reliably; see Chapter C.

Table 6.1.: Estimating the randomness parameter of the lattice Boltzmann model after 3000 simulation iterations, using 5 repetitions of the inference procedure. The topological loss together with importance sampling outperforms all of the other configurations. This table was recreated from Rohrscheidt, Rieck, and Schmon [158].

Loss	Sampling method	Randomness parameter η		
		0.20	0.30	0.40
ℓ_G	Importance sampling	0.16 ± 0.02	0.22 ± 0.04	0.35 ± 0.13
	MCMC	0.15 ± 0.10	0.91 ± 1.64	0.38 ± 0.25
ℓ_T	Importance sampling	0.21 ± 0.01	0.27 ± 0.00	0.38 ± 0.01
	MCMC	2.28 ± 1.40	1.71 ± 0.81	2.35 ± 1.10

6. Global Topology and Bayesian Inference

Table 6.2.: Estimating the parameter of the Percolation model, using 5 repetitions of the inference procedure. The topological loss together with either importance sampling or MCMC sampling outperforms all other configurations. This table was recreated from Rohrscheidt, Rieck, and Schmon [158].

Loss	Sampling method	Randomness parameter p		
		0.15	0.30	0.60
ℓ_G	Importance sampling	0.18 ± 0.00	0.30 ± 0.47	0.59 ± 0.00
	MCMC	3.18 ± 2.55	1.30 ± 0.47	2.18 ± 1.47
ℓ_T	Importance sampling	0.15 ± 0.00	0.29 ± 0.00	0.59 ± 0.01
	MCMC	0.15 ± 0.02	0.29 ± 0.00	0.59 ± 0.00

6.4.3. Percolation Model: Multi-Scale Structures

In statistical physics, percolation describes the behavior of a network as links are added incrementally. This process can be interpreted as a geometric phase transition, wherein, at a critical threshold, disconnected components merge into large, connected clusters with the addition of a small fraction of links.

In our setting, the percolation model is probabilistic, meaning that links are added at random with probability p . The objective is to infer the value of p based on the observed network structure. We represent this network as a 2D greyscale image of pixel size n^2 , where n is a positive integer. Here, we note that a greyscale image can be interpreted as an undirected graph by defining the set of vertices as the set of pixels and establishing edges between neighboring pixels if both have nonzero intensity values.

Each pixel in the image is assigned a value v , where $v = 0$ with probability $1 - p$, and v is drawn uniformly from the set $\{1, \dots, v_{\max}\}$ with probability p . Here, $v_{\max}$ represents the maximum attainable greyscale value, which we set to 50 in our experiments.

For the inference of p , we perform 250 simulations for both importance sampling and MCMC sampling. Given that we are working with greyscale images, we extract persistence diagrams using *cubical complexes*.

We evaluate our proposed topological loss ℓ_T against commonly used imaging losses, including mean squared error (MSE), root mean squared error (RMSE), the universal image quality index (UQI), the relative average spectral error (RASE), the spatial correlation coefficient (SCC), and the pixel-based visual information fidelity (VIFP). Since SCC and VIFP consistently outperform other metrics and yield comparable results, we only report SCC-based comparisons in Table 6.2 (see [211] for details on SCC construction).

Our results indicate that the topology-based loss ℓ_T , in combination with MCMC sampling, significantly outperforms the alternative methods.

Interestingly, in this particular experiment, rejection sampling using a mean statistic

achieves superior performance. This behavior stems from the specific characteristics of the percolation model, which allow for parameter inference based solely on the *mean* of pixel values. A more detailed discussion is provided in Chapter C.

Across our experiments, topological losses consistently outperform their geometrical counterparts. This observation underscores the value of topological descriptors in simulation-based inference. Complex systems often exhibit emergent collective behavior that is best captured by their *topological structure* rather than geometric properties. Geometric losses, which rely on precise positional information, fail to capture these global patterns, as small perturbations in individual elements may lead to large deviations in geometric metrics. In contrast, topological features tend to be more stable with respect to input parameters, making them a promising tool for a broad range of inference problems.

6.5. Discussion

We have introduced a novel Bayesian approach that integrates topological information into parameter estimation. Building upon the framework of generalized Bayesian inference, our method incorporates a topology-based loss function into the construction of the posterior distribution. This methodology is particularly well-suited for scenarios where the likelihood function is intractable, i.e., only accessible through simulations.

Our approach provides a theoretically grounded posterior belief update that addresses several limitations of traditional distance-based methods, which rely solely on geometric features. By leveraging topological descriptors, we improve inference in complex systems where standard geometric losses fall short.

The primary limitation of our approach lies in its dependence on the specific problem setting. While computational topology offers a powerful tool for certain complex models, its effectiveness in parameter estimation can vary significantly across different applications [112]. A thorough assessment of the suitability of topological features for a given scenario is recommended before application.

Additionally, our current work focuses on parameter spaces of relatively low dimensionality. Investigating the performance of our method in high-dimensional settings remains an open avenue for future research. Furthermore, we do not claim superiority over benchmark Bayesian methods in cases where the likelihood function is explicitly known. Instead, our empirical results demonstrate that our method consistently outperforms geometric approaches and summary statistics in complex system inference, highlighting its potential for applications in fields such as life sciences and dynamical systems modeling.

Our work contributes to the continued advancement of Bayesian inference methods for complex models, offering a novel perspective that bridges topological data analysis and Bayesian statistics. We believe that the principles introduced in this chapter hold broad applicability, particularly for uncertainty quantification in simulation-based

6. Global Topology and Bayesian Inference

inference.

From a machine learning standpoint, the key innovation of this work is the introduction of a framework that fuses topology with Bayesian inference, specifically targeting scenarios where gradient-based methods such as neural networks face fundamental challenges. Future research efforts should focus on refining this approach, particularly in developing efficient estimation techniques for sparse data regimes.

7. Local Geometry and Large Language Models

Understanding how large language models (LLMs) represent, process, and adapt language internally is one of the central open problems in modern machine learning. While most analyses treat models as black boxes evaluated solely by downstream performance, a complementary line of research examines the structure of their embedding spaces. This chapter continues the overarching theme of this thesis, the use of mathematical structure to probe and interpret representation spaces, by shifting the focus from topological to geometric analysis.

Building on the earlier chapters that investigated the utility of global and local topological methods, we now turn to a finer-grained geometric perspective based on local dimensionality. Specifically, we study how the local intrinsic dimension of contextual embeddings evolves through fine-tuning and task-specific adaptation. By treating these embeddings as samples from a (potentially intricate) data space which naturally carries geometry, we analyze how these geometric properties vary across training stages and datasets. This approach reveals interpretable signals that correlate with generalization, overfitting, and structural transitions such as grokking.

7.1. Introduction

Large language models (LLMs) have fundamentally reshaped natural language processing, achieving state-of-the-art results across a broad range of tasks including question answering, summarization, translation, and text generation [52, 143, 24, 181, 96]. At the core of these models lies their ability to learn contextual token representations, i.e. high-dimensional embeddings that reflect the linguistic and semantic context of tokens within a given input. These representations give rise to latent spaces whose structure governs how information is processed, generalized, and encoded across layers and tasks.

Despite their empirical success, a principled understanding of how these latent spaces evolve during training, fine-tuning, and downstream adaptation remains limited. Most current diagnostic tools focus on extrinsic measures such as held-out accuracy, perplexity, or task-specific probing. These methods offer valuable insights but often provide only coarse or post-hoc assessments of model behavior. In contrast, relatively few approaches seek to examine the internal geometry of the embedding spaces directly. Understanding these geometric properties offers the potential for more fine-grained, interpretable, and unsupervised insights into how

7. Local Geometry and Large Language Models

LLMs develop and transform internal structure in response to training signals.

This chapter explores whether changes in the geometry of embedding spaces can serve as diagnostic signals for key training dynamics, such as convergence, overfitting, and generalization, without relying on explicit supervision or labeled validation data. We focus on one particular geometric property: the *local intrinsic dimension*, which quantifies the effective dimensionality of the data space in the neighborhood of a point. To estimate this quantity, we employ a localized version of the TwoNN estimator [67], applied to the contextual token embeddings produced by LLMs at various stages of training. Although these embeddings reside in high-dimensional ambient spaces (often in the range of hundreds or thousands of dimensions), our analysis reveals that the intrinsic dimensionality in local neighborhoods is substantially lower and exhibits strong heterogeneity across different regions of the data space.

The motivation for this investigation is grounded in the broader research questions addressed throughout this thesis, which studies the internal structure of learned representations through topological and geometric methods. In earlier chapters, we introduced global and local topological descriptors to analyze the coarse-scale structure of embedding spaces. These tools capture connectivity patterns and holes in the data space, providing valuable insights into representational complexity and task structure. However, topological tools like homology generally do not account for more fine-grained geometric properties.

The present chapter therefore complements these topological approaches by introducing a local geometric perspective, focusing on intrinsic dimension as a data-driven measure of local complexity. This shift in perspective enables the detection of nuanced changes in the model’s internal representations that are not easily captured by global methods. For instance, we observe that fine-tuning on a new dataset tends to reduce the local intrinsic dimension in a task-specific manner. These reductions often align with improvements in task performance and can reveal to what extent the fine-tuned model generalizes beyond its training data — even without access to labeled validation examples.

Further, we find that intrinsic dimensionality is a useful indicator of several key training phenomena. In synthetic arithmetic tasks, we show that local dimension drops significantly just before the model begins to generalize correctly, a phase transition known as *grokking*, suggesting that dimension can serve as an early geometric signature of generalization. In dialogue state tracking tasks, we observe that stabilization of local dimension coincides with training convergence, while in sequence classification tasks, a characteristic rise in local dimension after an initial drop correlates with overfitting, marking the point at which the model begins to memorize rather than generalize. These empirical patterns emerge consistently across different tasks and model scales, reinforcing the utility of local geometric measures as general-purpose, unsupervised diagnostics. The results of this chapter are essentially based on [162].

7.2. Related Work

This chapter contributes to the growing literature at the intersection of geometric representation analysis, training dynamics, and interpretability in LLMs. Our work distinguishes itself by focusing on the local intrinsic dimension of contextual embeddings, computed across datasets and training stages. This localized perspective enables the unsupervised tracking of generalization dynamics and fine-tuning effects, offering a practical and interpretable diagnostic signal for language model behavior.

Early investigations into the internal structure of contextual language model embeddings revealed evidence of clustering and low-dimensional manifold structure. In particular, Ethayarajh [66] and Cai et al. [27] provided some of the first qualitative analyses of the representational geometry of transformer models, identifying non-isotropic embedding spaces and contextual variation in representation geometry. Subsequent work began to explore intrinsic dimensionality more explicitly. For example, Tulchinskii et al. [183] observed that the average intrinsic dimension of artificially generated texts is lower than that of human-written texts. However, their analysis is restricted to embeddings arising from single text paragraphs, limiting its applicability to dataset-level or training-dynamic questions. Similarly, Valeriani et al. [187] study how the global intrinsic dimension evolves as data is passed through LLMs, offering insights into how semantic information may concentrate at particular network depths.

Other directions of research have taken broader or alternative definitions of intrinsic dimension. Notably, Aghajanyan, Gupta, and Zettlemoyer [3] define intrinsic dimension in terms of the minimal number of directions in parameter space required to recover model performance. Their results suggest that larger models operate effectively in lower-dimensional parameter subspaces, but their approach targets the optimization landscape rather than the latent representation space of input data. In contrast, Viswanathan et al. [192] measure intrinsic dimension at the token level and link it to prediction loss. While this shares our interest in token-level granularity, their embedding space is always formed from the sequence of tokens within a single prompt, whereas our method systematically subsamples embeddings across an entire dataset split, enabling generalization beyond specific inputs and more robust estimation.

Complementary to geometric approaches, a number of studies have examined language model representations from a topological perspective. For instance, Kushnareva et al. [106], Perez and Reinauer [137], and Tulchinskii et al. [183] apply persistent homology to capture the topological features of sequence-level embedding spaces. These approaches characterize the presence of connected components, holes, or higher-order topological features, but do not offer a direct measure of local geometric complexity. Durrani et al. [58] explore how representations evolve during fine-tuning by comparing pre-trained and fine-tuned models using alignment functions and hierarchical clustering. Their findings show that while upper layers adapt to task-specific signals, lower layers tend to preserve general-purpose representations.

7. Local Geometry and Large Language Models

Recent work by Ed-dib, Datbayev, and Aboussalah [55] proposes an adaptive fine-tuning strategy that dynamically adjusts the rank of LoRA modules based on the intrinsic dimensionality of hidden state representations. While this also involves geometric reasoning, their notion of dimension is based on the rank of an information matrix and does not measure the intrinsic structure of the data manifold in latent space, as we do. More broadly, studies such as Chang, Tu, and Bergen [35] have examined representational geometry in multilingual models, analyzing how language-specific and universal features are encoded in a shared latent space. Our focus, in contrast, is not on encoding structure across language boundaries but on geometric signals associated with training dynamics.

Ruppik et al. [162] are among the few who also consider topological descriptors derived from local neighborhoods in contextual embedding spaces. They use these features to improve sequence tagging performance, although their method relies on a fixed ambient corpus and generates features that are not directly comparable across model checkpoints. By contrast, our approach produces intrinsic, dataset-relative estimates that generalize across models and training states, enabling temporal and comparative analysis of LLM behavior.

Finally, while our work aligns with the broader goals of interpretability, it differs substantially from efforts in mechanistic interpretability, which aim to reverse-engineer the internal structure of neural networks through symbolic or functional analysis of specific components (see, e.g., the survey by Ferrando et al. [72]). Instead, our method takes a more atomic view: we analyze the smallest discrete representation unit — the contextual token embedding — and study the geometry of spaces formed by their distribution. This allows us to remain model-agnostic, broadly applicable, and unsupervised, providing a new layer of interpretability grounded in geometric structure.

7.3. Methods

We now develop a geometric framework for analyzing the internal representation spaces of LLMs by estimating their local intrinsic dimensions. The aim is to quantify how the structure of contextual token embeddings evolves during fine-tuning and task-specific adaptation. To this end, we describe how contextual embeddings are extracted, how representative samples are obtained, and how local intrinsic dimensions are computed and compared across datasets and model checkpoints.

7.3.1. Large Language Models

Modern contextual language models can broadly be categorized into two main architectures: masked language models (MLMs) and autoregressive language models (ALMs). MLMs, such as BERT [52] and RoBERTa [110], are trained to predict randomly masked tokens within an input sequence using both left and right context, thus leveraging bidirectional information. In contrast, ALMs like GPT [144]

generate text in a strictly left-to-right fashion, predicting each token based solely on preceding tokens in the sequence. Despite their architectural differences, both model types yield contextualized vector representations for individual tokens at each layer of the network.

These layer-wise embeddings encode rich semantic and syntactic information about tokens in their specific input context. Formally, for each token in a given text sequence, a corresponding vector is produced at each model layer, resulting in a multi-layered tensorial representation of the entire sequence. This structure enables detailed geometric analyses of language model behavior, including the study of intrinsic dimensionality of LLM latent spaces.

In the following, we provide a formal description of how such contextual representations are extracted and how they can be used to compute localized geometric quantities across a dataset. The overall procedure is summarized in Algorithm 3.

7.3.2. Latent Space Modeling

Let $\mathcal{D} = (s_0, s_1, \dots, s_D)$ denote a corpus of $D + 1$ text sequences, and let $\mathcal{M}$ be a contextual language model, either a masked language model (MLM) or an autoregressive language model (ALM), with depth l and an associated tokenizer $\mathcal{T}$. Each input sequence $s_m \in \mathcal{D}$ is first tokenized into a sequence of subword units:

$$\mathcal{T}(s_m) = (t_0^m, \dots, t_{n_m}^m), \quad (7.1)$$

where n_m denotes the number of tokens in the m -th sequence after tokenization.

When processed by the model, each token t_j^m is assigned a contextual representation at every layer $i \in \{0, \dots, l\}$, giving rise to a sequence of vector embeddings:

$$\mathcal{M}_i(s_m) = (\mathcal{M}_i(t_0^m), \dots, \mathcal{M}_i(t_{n_m}^m)). \quad (7.2)$$

Although we use a simplified notation, it is important to emphasize that each embedding $\mathcal{M}_i(t_j^m)$ is inherently context-dependent and reflects the full input sequence s_m rather than the token in isolation.

We distinguish between two modes of generating token representations: *Regular embeddings* are obtained as tokens are processed as-is during a forward pass, preserving the original sequence. *Masked embeddings* are obtained as specific tokens are replaced with a special token, e.g., [MASK], and their embeddings are inferred by the model based solely on contextual clues. This mode applies primarily to MLMs and allows for studying how models impute missing information.

For a fixed layer i , we define the resulting token-level embedding space as

$$\mathbb{T}_i = \{\mathcal{M}_i(t_j^m) \mid m \in \{0, \dots, D\}, j \in I_m\}, \quad (7.3)$$

where I_m denotes the set of token indices in sequence s_m from which we extract embeddings. We treat these vectors as points in a high-dimensional Euclidean

7. Local Geometry and Large Language Models

space, and unless otherwise noted, distances between them are computed using the standard ℓ_2 norm.

Since $\mathbb{T}_i$ can easily comprise millions of vectors, exact geometric computations such as nearest-neighbor queries or intrinsic dimensionality estimation over the full dataset become computationally intractable. To address this, we adopt a two-stage sampling procedure to construct a representative subspace $\mathbb{T} \subset \mathbb{T}_i$.

We first sample a subset of M text sequences from the corpus $\mathcal{D}$, typically via uniform random sampling. After tokenizing and de-duplicating the resulting token set, we randomly select a subset of N contextualized embeddings from the corresponding embedding space to form the final working set $\mathbb{T}$.

For each vector in $\mathbb{T}$, we then compute a local neighborhood $\mathcal{N}_L(t_j; \mathbb{T})$ consisting of its L nearest neighbors within the subset. These neighborhoods serve as the foundation for downstream analyses.

This sampling-based approach approximates the local geometric structure of the full embedding space while remaining computationally efficient. It relies on the assumption that the text corpus is drawn from an underlying data-generating distribution with a stable latent geometry. We empirically assess the robustness of this approximation in our sensitivity analysis in Appendix D.1.

Algorithm 3 Compute Local Dimension Estimates

Require: Text corpus $\mathcal{D}$;

- 1: Embedding model $\mathcal{M}$ (tokenizer $\mathcal{T}$);
- 2: Dimension estimator dim (default $\text{dim} = \text{TwoNN}$);
- 3: Parameters: corpus subsample size $M \in \mathbb{N}$, token subsample size $N \in \mathbb{N}$, local neighborhood size $L \in \mathbb{N}$

Ensure: Vector of token-level local dimension estimates in $\mathbb{R}_{\geq 0}^N$

- 4: Take a random sub-sample $\mathcal{D}_{\sim M}$ of size M from $\mathcal{D}$
 - 5: Compute embeddings for each token in each sequence of $\mathcal{D}_{\sim M}$
 - 6: Remove duplicate embedding vectors
 - 7: Take a token-vector sub-sample of size N ; call the resulting set $\mathbb{T}$
 - 8: **for all** $t_j \in \mathbb{T}$ **do**
 - 9: Find the L nearest neighbors of t_j in $\mathbb{T}$, giving $\mathcal{N}_L(t_j; \mathbb{T})$
 - 10: Compute the local estimate $d_j \leftarrow \text{dim}(\mathcal{N}_L(t_j; \mathbb{T}))$
 - 11: **end for**
 - 12: **return** $(d_j)_{j=1}^N$
-

7.3.3. Comparing Latent Spaces Across Models

To quantify the effect of fine-tuning on the geometric structure of contextual embedding spaces, we consider a pre-trained base model $\mathcal{M}$ alongside its fine-tuned variant $\mathcal{M}^\Delta$, where fine-tuning has been performed on a text corpus Δ . By construction, both models share the same underlying architecture and tokenizer, ensuring that the input-output interface and tokenization behavior remain consistent across

the two models.

This architectural identity induces a canonical pointwise correspondence between the contextual representations generated by the two models. Formally, for any input sequence $s_m \in \mathcal{D}$ and any model layer i , the embeddings

$$\mathcal{M}_i(s_m) = (\mathcal{M}_i(t_0^m), \dots, \mathcal{M}_i(t_{n_m}^m))$$

and

$$\mathcal{M}_i^\Delta(s_m) = (\mathcal{M}_i^\Delta(t_0^m), \dots, \mathcal{M}_i^\Delta(t_{n_m}^m))$$

define a bijective pairing of contextualized token embeddings across models. In particular, this pairing naturally extends to the sampled embedding sets $\mathbb{T} \subset \mathbb{T}_i$ and $\mathbb{T}^\Delta \subset \mathbb{T}_i^\Delta$, provided that sub-sampling is performed identically in both cases.

This setup enables a direct, token-level comparison of embedding geometry before and after fine-tuning. Such comparisons allow us to assess how local neighborhoods, dimensionality, and other structural properties of the latent space evolve as a result of task-specific training. More generally, such a bijection exists whenever the underlying model architectures and tokenizers are identical.

In the following, we formalize the local geometric quantities used for this comparative analysis.

7.3.4. Local Dimension Estimates

To characterize the local geometric structure of embedding spaces, we employ the *Two Nearest Neighbors* (TwoNN) estimator proposed by Facco et al. [67], a lightweight and nonparametric method for estimating the intrinsic dimension of potentially high-dimensional data. The key idea behind TwoNN is to leverage the relative distances of each data point to its nearest and second-nearest neighbors to infer local dimensionality.

Let r_1 and r_2 denote the distances from a point $v \in X$ to its first and second nearest neighbors in a dataset X , respectively. Under mild regularity conditions (assuming that the data lies on a smooth manifold and the local sampling density is approximately uniform at the scale of r_2), the ratio $\frac{r_2}{r_1}$ is distributed according to a Pareto distribution. As shown by Denti et al. [51], this distributional property enables consistent estimation of the intrinsic dimension using maximum likelihood techniques or linear regression on a logarithmic transformation of the ratios.

To adapt the TwoNN estimator into a *local measure*, we proceed as follows. Let $X = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ be a point cloud and fix a locality parameter $L \in \mathbb{N}_{>0}$. For a target point $v \in X$, we define its L -local neighborhood $\mathcal{N}_L(v; X)$ as the set of the L nearest neighbors of v in X .

The local TwoNN estimate at v is then given by:

$$\text{TwoNN}(\mathcal{N}_L(v; X)),$$

7. Local Geometry and Large Language Models

where the estimator is applied to the set consisting of v and its L neighbors. The parameter L thus determines the locality scale of the dimension estimation: small values of L capture fine-grained geometry, while larger values average over a broader neighborhood.

In the context of our analysis, we instantiate $X = \mathbb{T} \subset \mathbb{R}^d$, the subsampled set of contextual token embeddings obtained from a given model $\mathcal{M}$ and corpus $\mathcal{D}$ (as described in the previous section). We compute a vector of local dimension estimates:

$$(\text{TwoNN}(\mathcal{N}_L(v; \mathbb{T})))_{v \in \mathbb{T}} \in \mathbb{R}_{\geq 0}^N,$$

assigning a positive real value to each token embedding in the subsample.

To obtain a summary statistic for a particular model-layer combination, we aggregate this vector using the arithmetic mean, yielding a single scalar that reflects the average local dimensionality across the sampled tokens.

Although the TwoNN estimator is sensitive to hyperparameters such as the locality scale L and the number of sampled points, we demonstrate in Appendix D.1 that our estimates remain stable under a wide range of reasonable configurations. This robustness enables reliable comparisons across datasets, models, and fine-tuning stages.

7.4. Experiments

In this section, we apply our proposed framework to analyze how the latent representation spaces of large language models evolve throughout various training scenarios. Unless noted otherwise, all measurements refer to the model’s final hidden layer, denoted $\mathbb{T}_{-1}$ in the notation from the above section. Results for earlier layers are discussed in Section D.3.2. We address four key research questions in this chapter: How does fine-tuning affect the geometry of embedding spaces? Can local dimension estimates signal the emergence of grokking? Can they reveal the limits of training progress? And finally, can they help detect overfitting?

7.4.1. Fine-Tuning Induces Dataset-Specific Shifts in Local Dimensions

To analyze how fine-tuning alters the internal representations of language models, we investigate changes in the distribution of local intrinsic dimensions across contextual token embeddings. These estimates, derived using the methods introduced before, provide a geometric lens on how models adapt their internal structure to new data.

By studying the distribution of these local dimensions before and after fine-tuning, we obtain a detailed characterization of representational shifts. This approach allows us to move beyond traditional metrics of model performance and instead focus on how the geometry of internal representations evolves during adaptation.

Experimental Setup We conduct our analysis on a set of diverse language datasets that vary in both domain and their overlap with the training data of the base and fine-tuned models. These include the `MultiWOZ2.1` dataset [64], which consists of multi-domain dialogues between humans; the Schema-Guided Dialogue (SGD) dataset [145], comprising human–virtual assistant interactions; a corpus of Reddit comments from 2022 that mention Tesla Inc.; and the Hugging Face `wikitext-103-v1` corpus containing English Wikipedia articles. These datasets were chosen to span multiple levels of domain familiarity: some, such as Wikipedia, are included in the original pre-training data of models like RoBERTa [110], while others, such as `MultiWOZ` and SGD, are used explicitly for fine-tuning. In contrast, Reddit data represents an out-of-domain scenario, with minimal overlap in both topic and style.

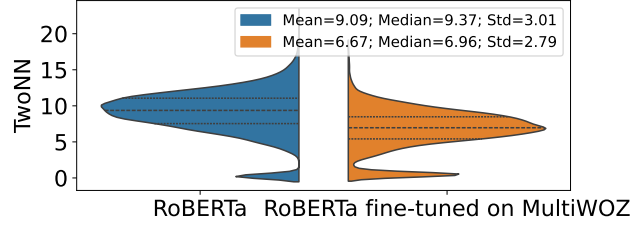
Each dataset is randomly split into training (80%), validation (10%), and test (10%) subsets. For datasets with chronological structure, such as Reddit, we perform random shuffling prior to splitting in order to remove potential temporal bias. Pre-processing of Wikipedia consists of removing section headers, empty lines, and trimming whitespace. Summary statistics, including token counts and sequence lengths, are reported in Table D.1.

Fine-tuning is conducted on the `roberta-base` model using a masked language modeling (MLM) objective with a masking probability of 0.15. We train each model for five epochs on 10,000 examples using a batch size of 8. The learning rate linearly increases to a peak of $5 \cdot 10^{-5}$ over 500 warm-up steps and then decays linearly. We apply a weight decay of 0.01 throughout training to encourage generalization. To evaluate the representational impact of fine-tuning, we embed a random subset of the validation split from each dataset using both the pre-trained and the corresponding fine-tuned model. This ensures that embeddings are drawn from data not seen during the fine-tuning process and enables a controlled comparison. Additional results using autoregressive fine-tuning can be found in Section D.3.1.

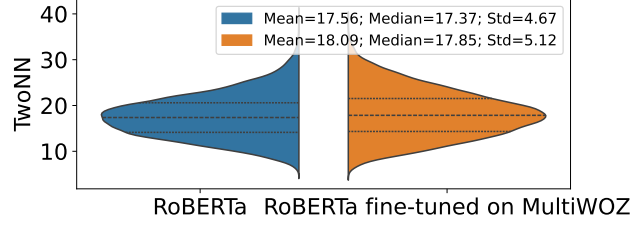
Results Our analysis reveals substantial variation in local intrinsic dimensions across individual tokens, as illustrated by the wide spread of the TwoNN distributions in Figure 7.1. To obtain these estimates, we fix the following parameters: corpus sub-sample size $M = 7000$, token vector sub-sample size $N = 60\,000$, and local neighborhood size $L = 128$. The observed heterogeneity in local dimensionality is consistently present across different models and datasets, reinforcing the notion that contextualized embedding spaces are not globally uniform but rather composed of geometrically diverse subregions. This aligns with prior findings in the literature indicating that data representations rarely reside on a single, well-behaved manifold of constant dimension, but instead exhibit piecewise variability in their local structure [23]. These observations support our methodological choice to prioritize local over global dimensionality estimates, as the latter may obscure important fine-grained structural properties of the embedding space.

A central finding of our study is that fine-tuning leads to a systematic reduction in local intrinsic dimension, but only on the dataset used during the fine-tuning

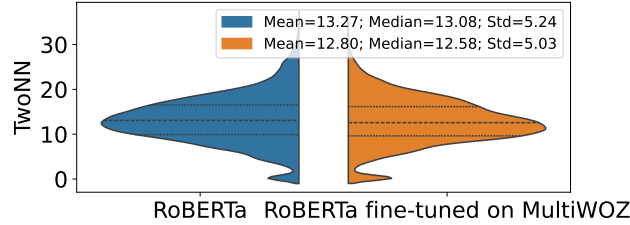
7. Local Geometry and Large Language Models



(a) TwoNN estimates on MultiWOZ validation.



(b) TwoNN estimates on Wikipedia validation.



(c) TwoNN estimates on Reddit validation.

Figure 7.1.: Comparison of the intrinsic dimension (ID) across three data modalities for both models. The distribution of the local estimates over tokens is shown in the violin plot, together with their means and quartiles. The ID of embeddings originating from the fine-tuning distribution (MultiWOZ) differs markedly between models, whereas the IDs for the out-of-distribution corpora (Wikipedia, Reddit) are almost indistinguishable. This figure was recreated from Ruppik et al. [162].

process. As shown in Figure 7.1a, the distribution of TwoNN estimates for the MultiWOZ-validation set exhibits a pronounced downward shift when comparing the fine-tuned RoBERTa model to its pre-trained counterpart. This reduction suggests a condensation or restructuring of the representational geometry that enhances task-specific expressiveness. In contrast, embeddings from unrelated datasets, such as Wikipedia (Figure 7.1b) or Reddit comments (Figure 7.1c), display no substantial changes in their dimension distributions. This dataset-specific effect is further quantified using the standardized mean difference between the local dimension estimates of the fine-tuned and base models. For MultiWOZ, the standardized mean difference reaches 1.19, indicating a large effect, whereas for Reddit and Wikipedia, the respective differences are only 0.08 and 0.10, respectively.

Importantly, we observe that the reduction in local dimension is robust across

the training, validation, and test splits of the fine-tuned dataset. This consistency suggests that the effect reflects a stable structural adaptation in the model’s latent space rather than spurious overfitting or sampling noise. As we demonstrate in our sensitivity analysis (Appendix Section D.1), the phenomenon persists across a wide range of hyperparameter settings and sampling strategies, further validating the reliability of our estimates. Overall, these results underscore the interpretive power of local geometric analysis: fine-tuning appears to act by selectively compressing the embedding space in regions corresponding to in-distribution data, thereby improving the model’s capacity to represent and generalize over the target domain. Conversely, regions of the embedding space associated with unrelated distributions remain effectively untouched, reinforcing the view that local dimension estimates can serve as precise diagnostics for model adaptation and specialization.

7.4.2. Local Dimensions Predict Grokking

The phenomenon known as *grokking* refers to a striking behavior in certain neural networks, where generalization to unseen data emerges only after prolonged training. This counterintuitive effect was first documented by Power et al. [141] in the context of small algorithmic tasks. Despite achieving zero training error early in training, the model initially performs poorly on the validation set and only begins to generalize after a long plateau phase. Predicting whether and when grokking will occur has proven difficult, as it is highly sensitive to architectural and optimization hyperparameters [99]. Understanding the underlying mechanisms of this delayed generalization remains a compelling open question, particularly as it challenges standard assumptions about overfitting and convergence in deep learning.

Experimental Setup To investigate whether local geometric properties of the embedding space can capture signals indicative of grokking, we consider the task of learning an arithmetic operation in a finite group - specifically, addition modulo p . For instance, for $p = 197$, a typical input sequence may take the form [155, ‘o’, 88, ‘=’], where the model is expected to predict the result 46. The input space is fully tokenized such that each integer operand is mapped to a unique token, augmented with a token for the addition operator (‘o’) and one for the equality symbol (‘=’). We construct training sets by sampling a fraction of all valid expressions in the group, varying this fraction between 10% and 50%. The remaining expressions constitute the validation set, allowing us to evaluate generalization behavior under different data regimes.

We train a small decoder-only transformer model from scratch, consisting of two layers, a hidden size of 128, and four attention heads. The model is optimized using the AdamW optimizer [111] with a learning rate of 0.001, following a linear schedule over 400,000 training steps and a warm-up phase of 10 steps. A batch size of 512 and a weight decay of 0.01 are used throughout. The training procedure is deliberately extended far beyond the point of zero training error, replicating the conditions under which grokking is typically observed. This mirrors the original

7. Local Geometry and Large Language Models

experimental protocol introduced by Power et al. [141], ensuring comparability of results.

To probe the internal geometry of the model’s representations during training, we estimate local intrinsic dimensions from the hidden states of the final layer. We set the neighborhood size L to 64 and subsample up to $N = \min(3000, N)$, where N denotes the total number of available token embeddings at the given training step. Because the total number of unique input sequences is relatively small in this task, we include the entire dataset for sequence-level sampling ($M = N$) and apply subsampling only at the token level. The sampled embeddings include tokens corresponding to both operands and special symbols (i.e., ‘o’, ‘=’), as these all contribute to the model’s representation of the compositional arithmetic structure.

Across a wide range of hyperparameter settings and training configurations, we find that our local dimension estimates exhibit qualitatively consistent patterns, suggesting that the observed dynamics are not merely artifacts of sampling noise. This robustness allows us to track how representational complexity evolves during training and whether any measurable geometric transitions precede or accompany the onset of grokking. Detailed results and dimension trajectories over training time are presented in the appendix.

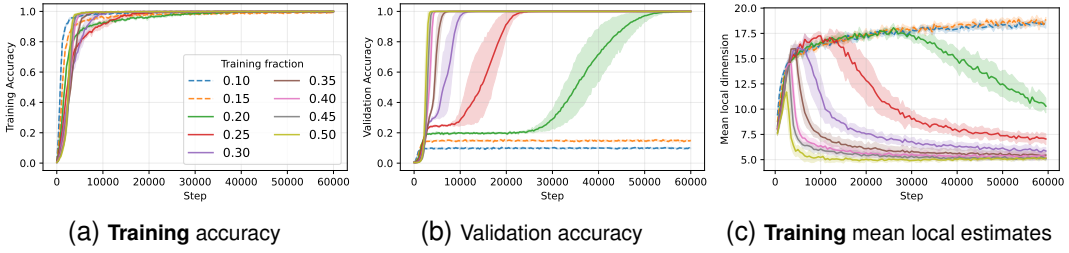


Figure 7.2.: Training a model on addition mod $p = 197$ with different training data fraction selected from $\{0.1; 0.15; 0.2; 0.25; 0.3; 0.4; 0.5\}$. The plots show the development for 60 000 batches, with mean and 95% confidence interval over 5 training seeds per configuration (plots per seed are in Appendix D.3.3). The mean local estimates are computed on the **training split** for the parameters $N = 3000$; $L = 64$. Dashed lines highlight the runs where grokking did *not* occur. This figure was recreated from Ruppik et al. [162].

Results We now analyze how local intrinsic dimension estimates evolve over training and how they relate to the onset of generalization in the grokking regime. In Figure 7.2, we present performance metrics and mean local dimension values over the first 60 000 training steps, averaged across multiple random seeds. Results for individual seeds are provided in Figure D.11 in the appendix to demonstrate the consistency of the observed patterns.

In all configurations, training accuracy reaches near-perfect levels within the first few thousand steps. However, validation accuracy remains low initially and only begins to improve after an extended period, but only in those runs where the training

dataset fraction exceeds 20%. This delayed improvement in validation accuracy is characteristic of grokking. Crucially, based on training loss and accuracy alone, it is not possible to distinguish between runs that will eventually generalize and those that will not. This underscores the insufficiency of conventional performance metrics in diagnosing future generalization in grokking scenarios.

In contrast, local intrinsic dimension estimates offer a more nuanced and predictive signal. As shown in Figure 7.2c, the mean local dimension computed on training data typically increases during the initial training phase. This rise reflects the model's early attempts to fit the training data, possibly by expanding the representational capacity in the embedding space. However, for runs that eventually exhibit grokking behavior, we observe a marked decline in the mean local dimension shortly after this initial growth phase. Importantly, the timing of this drop closely aligns with the point at which validation accuracy begins to rise, as seen in Figure 7.2b.

This alignment suggests that the model undergoes a geometric transition in its internal representations, contracting the latent space in a way that supports abstraction and generalization beyond memorization of the training set. The reduction in local dimension appears to coincide with the reorganization of the embedding space to support structure-preserving generalization over the arithmetic task.

For runs with insufficient training data (specifically, the 10% and 15% fractions, denoted with dashed lines in the figures), validation accuracy remains flat throughout the observed training window. In these cases, the mean local dimension increases early on but then plateaus, showing no sign of the sharp decline characteristic of grokking runs. This stagnation suggests that the model may be overfitting, learning to memorize the training examples without uncovering the underlying group structure necessary for generalization.

Taken together, these results demonstrate that changes in local intrinsic dimensionality can serve as a powerful diagnostic tool in grokking experiments. In particular, a sharp drop in the mean local dimension on training data provides an early, geometry-based indicator of whether a model will generalize, even in the absence of corresponding validation signals at that time.

7.4.3. Local Dimensions Detect Exhaustion of Training Capabilities

Identifying relevant text spans in response to a query, commonly referred to as *span prediction* or *sequence tagging*, is a foundational task in natural language processing, with broad applicability across domains such as information retrieval, question answering, and conversational AI [100]. A particularly prominent use case is retrieval-augmented generation, where relevant text passages must be efficiently selected from a large corpus prior to response generation [68]. In this work, we focus on the application of sequence tagging in dialogue state tracking (DST), a core component in task-oriented dialogue systems. DST models are responsible for interpreting user utterances and maintaining an up-to-date representation of the user's goals and constraints throughout a conversation [202].

7. Local Geometry and Large Language Models

To investigate how local geometric structure evolves in such models, we compute local intrinsic dimension estimates for *TripPy-R*, a state-of-the-art sequence tagging model for dialogue state tracking introduced by Heck et al. [83]. TripPy-R operates on the MultiWOZ2.1 dataset [64], a large-scale benchmark comprising multi-domain dialogues between humans. The model uses a unified encoder built on the RoBERTa architecture, which is fine-tuned during DST training. This encoder processes each dialogue input and produces contextualized token embeddings from its final hidden layer. These embeddings are then passed to task-specific classifiers that determine the slot-value updates for each dialogue turn.

Model training is conducted over 20 epochs using the Adam optimizer, with a linear learning rate schedule that peaks at $5 \cdot 10^{-5}$ and includes a warm-up phase spanning the first epoch. The training objective is to minimize the error in predicting dialogue state updates at each turn, which provides a dense supervision signal that influences all layers of the encoder.

To analyze the model’s internal representations, we compute local intrinsic dimensions based on hidden states from the final encoder layer. The input format mirrors the training and inference pipeline, where each sequence consists of the current user utterance followed by the dialogue history, separated by special tokens. This structure enables the model to leverage context from previous turns to predict the current dialogue state accurately.

For our analysis, we sample $M = 7000$ input sequences from each of the training, validation, and test splits of the dataset. From these, we subsample up to $N = 60\,000$ token embeddings and compute local dimension estimates using a neighborhood size of $L = 128$. These parameters are consistent with those used in previous sections to ensure comparability. Our goal is to examine how the model’s local representational complexity varies across data splits and whether it encodes meaningful distinctions that relate to generalization, robustness, or domain adaptation in dialogue state tracking.

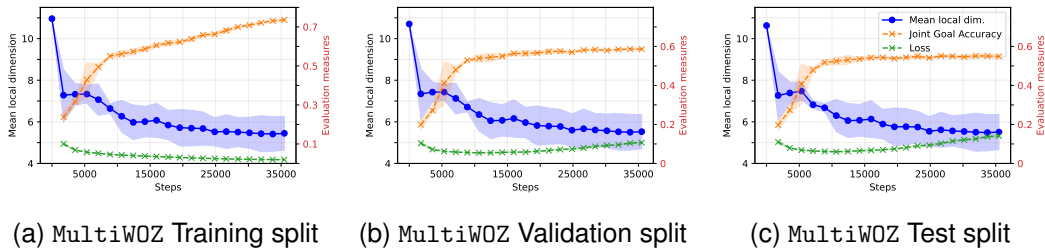


Figure 7.3.: Development of Trippy-R performance measures (model loss in green; joint goal accuracy in orange) compared with mean local dimension estimates (dark blue) evaluated on the training, validation, and test split of the MultiWOZ dataset. We show the mean and standard deviation of the measures evaluated at the end of each epoch over six different model training seeds. This figure was recreated from Ruppik et al. [162].

Results Figure 7.3 illustrates the evolution of two key performance metrics during training: the differentiable model loss used for backpropagation, and the non-differentiable *joint goal accuracy* (JGA), a standard evaluation metric for dialogue state tracking that measures whether the model correctly predicts the full set of slot-value pairs for each turn.

Interestingly, while the loss and JGA exhibit divergent behaviors across data splits, the local intrinsic dimension estimates remain consistent. In particular, the training loss decreases monotonically, as expected, and the JGA on the training set rises steadily throughout the training process. On the validation set, however, the situation is more complex: the validation loss reaches its minimum relatively early—after approximately 7500 batches—well before the JGA converges. This discrepancy is noteworthy because it highlights a limitation of relying on differentiable loss metrics alone for assessing generalization. In this case, the early minimum in validation loss could misleadingly suggest that training should be stopped, even though the model has not yet learned robust, generalizable representations.

In contrast, the local intrinsic dimension estimates tell a different story. Even at the point where validation loss is minimized, the mean local dimension continues to decrease steadily across training, validation, and test splits. This ongoing reduction indicates that the model is still reorganizing its internal representations and has not yet converged. Crucially, this insight can be obtained using training data alone—without the need for a labeled validation set—making local dimension estimates a potentially powerful tool for guiding training schedules in low-resource or semi-supervised settings.

Later in the training process, around step 25 000, the local intrinsic dimension estimates begin to stabilize. This stabilization coincides closely with the convergence of the JGA on both the validation and test sets, suggesting that the model has exhausted its capacity to further improve its performance on the target task. The alignment of this geometric stabilization with downstream performance saturation provides further evidence that local intrinsic dimension can serve as a meaningful proxy for the model’s capacity utilization and learning progress.

This predictive behavior becomes even more pronounced in longer training runs. As shown in Figure D.12 in Section D.3.4, the convergence of the local dimension estimates precedes or matches the plateau of JGA in extended training regimes, reinforcing the view that local geometric signals can serve as early indicators for stopping criteria or model selection.

7.4.4. Local Dimensions Detect Overfitting

We now turn to the more challenging scenario of emotion classification in dialogues, to evaluate whether local intrinsic dimension can serve as an indicator of overfitting. Overfitting is particularly problematic in this setting due to class imbalance and limited signal available in individual utterances. Our aim is to assess whether geometric indicators from the model’s latent space can reveal when a model begins

7. Local Geometry and Large Language Models

to memorize training data at the expense of generalization, even before performance degradation becomes apparent in standard validation metrics.

Experimental Setup Our experiments are conducted on the EmoWOZ dataset [70], which consists of task-oriented dialogues annotated with one of seven emotion labels per user utterance. We use the ERTOD model [71] in its baseline configuration. Specifically, the model comprises a frozen `bert-base-uncased` feature extractor [52] followed by a simple linear classifier. In contrast to more elaborate model variants proposed by the authors, we intentionally disable enhancements such as dialogue history context, data augmentation, and auxiliary prediction heads, in order to isolate the effect of training dynamics on representation quality. This simplified setup also makes overfitting behavior more pronounced and observable.

The model is trained for eight epochs using the AdamW optimizer with a peak learning rate of $2 \cdot 10^{-5}$ and a linear learning rate schedule, including a warm-up phase. After each epoch, we compute local intrinsic dimension estimates from the hidden representations at the final BERT encoder layer. This space directly informs the linear classification layer, and is thus where the model concentrates its representational effort to distinguish between emotion classes.

To estimate local dimensions, we sample $M = 7000$ utterances from each of the training, validation, and test splits. From these, we extract up to $N = 60\,000$ token embeddings and compute TwoNN-based dimension estimates using a neighborhood size of $L = 128$. All inputs are processed in the same form as during training and inference, that is isolated utterances without dialogue history, to ensure consistency between learned representations and evaluated embeddings.

Model performance is evaluated using standard classification metrics, including cross-entropy loss, weighted F1 score, and macro F1 score. Since the “neutral” emotion class constitutes the majority of the data and may dominate aggregate metrics, we exclude it from the F1 evaluations to better capture the model’s performance on underrepresented emotion categories. This design choice allows us to better analyze whether changes in local dimension correlate with the model’s ability to learn generalizable features for minority emotion classes, which is central to assessing overfitting in imbalanced classification tasks.

Results Figure 7.4 illustrates the relationship between the model’s performance metrics and the mean local intrinsic dimension across training epochs. After just one epoch of training, the mean local dimension of the model’s hidden representations drops sharply from an initial value of approximately 9.94 (corresponding to the frozen base model) to around 7.25. In the subsequent epochs, the dimension gradually increases, reaching a value near 8 by the end of training. This non-monotonic trajectory is observed consistently across all three data splits: training, validation, and test. On the training set, the expected learning pattern emerges: cross-entropy loss decreases monotonically, while both the weighted and macro F1-scores improve steadily.

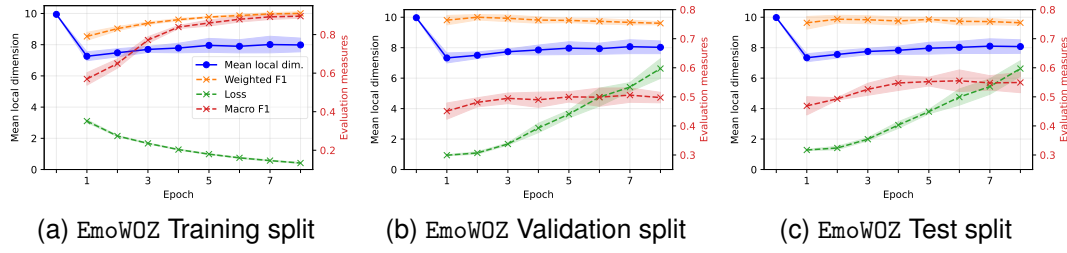


Figure 7.4.: Development of emotion recognition model performance measures (loss in green; weighted F1 in orange; macro F1 in red) compared with mean local dimension estimates (dark blue) evaluated on the training, validation, and test split of the EmoWOZ dataset. We show the mean and standard deviation of the measures evaluated at the end of each epoch over four different training seeds. This figure was recreated from Ruppik et al. [162].

However, the observed minimum in local dimension followed by a gradual rise suggests a more nuanced internal learning dynamic. The steep drop after epoch one indicates that the model rapidly converges to a compact and seemingly efficient latent representation. Yet, the subsequent increase in dimensionality implies that the model begins to expand its representation space, possibly in an attempt to accommodate increasingly specific patterns in the training data. This mirrors the behavior seen in the grokking experiments described in Section 7.4.2, where a rise in local dimension was linked to memorization and overfitting.

This hypothesis is corroborated by the evolution of validation loss, shown in Figure 7.4b. While the model achieves its most compact latent representation after the first epoch, validation loss increases significantly thereafter, providing strong evidence of overfitting. In other words, even as training loss continues to decline, the model’s ability to generalize deteriorates, which is not immediately apparent from the training metrics alone.

The impact on classification performance is subtle and multifaceted. The macro F1-score on the validation set increases and plateaus after approximately three epochs, indicating that the model continues to improve on minority classes. In contrast, the weighted F1-score which is dominated by the majority “neutral” class, declines in the third epoch. This divergence suggests that the model is learning to prioritize correct predictions for underrepresented classes, potentially by overfitting to rare training examples, at the cost of performance on majority examples. The trade-off between macro and weighted F1 thus reflects a shift in the decision boundary that favors minority class recall but reduces the overall classification stability.

Crucially, this behavior is not evident from conventional loss curves alone. Instead, the evolution of local dimension provides early and interpretable signals of representational drift, overfitting, and emerging class trade-offs. These signals can be observed using the training data exclusively, highlighting the potential utility of local dimension estimates as a diagnostic tool in low-supervision settings, where labeled validation data may be scarce or unavailable.

7.4.5. Computational Resources

The computational requirements of our analysis are moderate and compatible with standard research hardware. Fine-tuning the `roberta-base` and `gpt2-medium` models on one of our selected datasets can be performed efficiently on a single NVIDIA V100 or GTX1080Ti (12GB VRAM) GPU. For the datasets and training schedules considered in this work, fine-tuning typically completes within a few hours. Once trained, extracting token embeddings from a specific layer of the model involves only forward passes through the network. For a dataset of moderate size, embedding extraction takes approximately 10 minutes per layer on the same hardware.

The estimation of local intrinsic dimensions via the TwoNN method is more computationally demanding, as it is CPU-bound and scales with both the size of the sampled embeddings and their ambient dimensionality. For a typical configuration involving tens of thousands of high-dimensional token embeddings (with ambient dimensions in the hundreds), the computation can be completed in approximately 20 minutes on a dual-core Intel Xeon E5-2640v4 (Broadwell, 2.40GHz) machine with 32GB of RAM. These experiments use the `scikit-dimension` Python package [7], which provides efficient and well-validated implementations of dimension estimators including TwoNN.

In more compute-intensive settings such as the grokking experiments, where local dimensions are repeatedly evaluated throughout long training runs, the total cost accumulates accordingly. For the full range of training steps considered, a typical run requires approximately 48 CPU core hours. However, precise timings may vary depending on implementation details and the choice of hyperparameters, especially the local neighborhood size and subsample size used in each estimation round.

The theoretical complexity of the TwoNN estimation is driven primarily by the neighborhood computation step. For a fixed and small neighborhood size L and ambient embedding dimension d , computing exact k -nearest neighbors over a subsample of size N requires $\mathcal{O}(dN^2)$ operations. In our setup, we use exact search via pairwise distance computations, implemented efficiently using vectorized operations. Once neighborhoods are determined, the estimation step itself is computationally lightweight: the local dimension is estimated by fitting a straight line to the log-transformed empirical distribution of nearest-neighbor distances, which involves only a single linear regression per point.

Overall, our pipeline is optimized to support fast iteration during empirical analysis while remaining accessible to research groups without access to large-scale computational infrastructure.

7.4.6. Limitations

While our framework offers a flexible and informative tool for analyzing representational geometry in neural models, it is not without limitations. One of the primary drawbacks lies in its computational cost. Obtaining reliable local dimension estimates requires the extraction of a sufficiently large sub-sample of latent representations

through forward passes, followed by neighborhood construction and subsequent dimension estimation. Although these steps are executed independently of the training process, and thus can be performed asynchronously or in parallel to fine-tuning, the overhead remains non-trivial, particularly for large models or datasets.

A promising direction for future research lies in improving the computational efficiency of these estimates. This could involve optimizing nearest-neighbor computations, leveraging approximate methods, or designing new local descriptors that maintain the desirable stability properties of current estimators while reducing computational burden. Additionally, integrating these estimates more tightly into the training loop could open avenues for dynamic model diagnostics during training.

Another important limitation is rooted in the statistical assumptions underlying the dimension estimation itself. The TwoNN estimator, which we apply locally within each neighborhood, is known to yield theoretically accurate estimates only under restrictive conditions. In particular, it assumes that the point samples within each neighborhood are drawn from a homogeneous Poisson process on a manifold of constant intrinsic dimension. In practice, these assumptions are unlikely to hold exactly in the complex and often anisotropic latent spaces of neural networks. However, for our purposes, this limitation is not critical: our primary interest lies in capturing *relative changes* in dimension across time, layers, or training regimes, rather than computing ground-truth intrinsic dimensions. The qualitative trends and shifts observed remain robust, even if absolute values are biased.

Finally, although the TwoNN estimator is technically non-parametric, its output is still sensitive to certain hyperparameter choices. In particular, the size of the latent sub-sample (N) and the neighborhood size (L) influence the absolute values of the resulting estimates. This dependence implies that comparisons of local dimension values across different model architectures or experimental setups must be made with care. In fine-tuning experiments, it is more meaningful to consider *relative changes* in dimension compared to the initial pre-trained model, rather than interpreting absolute values in isolation. For cross-model comparisons or large-scale benchmarking, further work is needed to establish calibration protocols or normalization schemes that facilitate consistent interpretation.

7.4.7. Discussion

In this work, we introduced a novel geometric lens on the internal training dynamics of large language models by quantifying the local intrinsic dimension of contextual token embeddings. Our analysis reveals that the latent spaces of contextual models are not uniformly structured but instead exhibit meaningful variation in local dimensionality across tokens, inputs, and datasets. Despite this heterogeneity, the mean local dimension emerges as a stable and interpretable summary statistic that provides actionable insights into model behavior.

Across a range of experimental settings, including fine-tuning, grokking, and overfitting scenarios, we found that sustained decreases in mean local dimension con-

7. Local Geometry and Large Language Models

sistently correlate with improved generalization. These reductions reflect a compressive transformation of the latent space, wherein the model learns to represent task-relevant information more efficiently. Conversely, rising local dimensions are indicative of specialization and memorization, often accompanying overfitting. These findings suggest that local dimension can serve as an unsupervised diagnostic tool capable of detecting transitions in learning dynamics, even in the absence of labeled validation data.

This unsupervised nature is particularly valuable in low-resource or semi-supervised settings, where validation metrics may be unreliable or unavailable. By offering a training-time signal rooted in the geometry of the model’s internal representations, our approach opens a new avenue for interpretability and model selection based on structural indicators rather than downstream task performance alone. Overall, our results demonstrate the utility of geometric descriptors for tracking, understanding, and potentially guiding the learning process in large language models.

To the best of our knowledge, this is the first comprehensive study of local dimensional shifts during the training and fine-tuning of contextual language models across diverse downstream tasks. Our results contribute to the broader discourse on interpretability, generalization, and representational efficiency in modern NLP systems. In doing so, we establish a methodological foundation for leveraging intrinsic geometric properties as first-class tools for model analysis and control.

There are several promising directions for future research. First, the inherently local nature of our dimension estimates opens the door to more fine-grained interpretability: one could investigate how different types of tokens, such as entities, syntactic markers, or rare words, are associated with higher or lower intrinsic dimension, potentially shedding light on semantic and structural encoding in LLMs. Second, the ability to detect dataset-specific compression may enable new approaches to data provenance and model auditing. For instance, one could explore whether local dimension patterns can be used to infer which portions of a dataset were used during training, or to identify distributional shifts in deployment scenarios.

More broadly, our framework could inform the development of regularization strategies or architecture designs that encourage or exploit low-dimensional representations. Embedding local dimension as a soft constraint or auxiliary signal during training might provide a novel means of improving generalization or controlling model complexity. In this way, the geometric insights introduced here not only deepen our understanding of LLM behavior but may also inspire practical innovations in how we train, evaluate, and deploy them.

8. Conclusion and Future Work

This thesis explored the interplay between topology, geometry, and representation learning, proposing novel methodologies to better understand and leverage the structural properties of data across a range of machine learning settings. By developing tools that probe local manifold structure, topological-geometric invariants, and latent space geometry, we intend to have advanced the theoretical foundations and practical applicability of topological and geometric data analysis in modern ML workflows.

We introduced TARDIS, a framework for the localized estimation of intrinsic dimension and deviation from Euclidean geometry via a new statistic called *Euclidicity*. This method enables fine-grained analysis of data geometry, particularly in singular or non-manifold regions where traditional techniques fall short. We showed its relevance for anomaly detection, unsupervised learning, and model robustness, and we outlined pathways to integrate Euclidicity into training pipelines or as a regularizer in generative modeling.

In parallel, we developed the *Local Euler Characteristic Transform* (ℓ -ECT), a novel topological representation tailored for graph-structured data. Unlike message-passing neural networks, ℓ -ECT avoids lossy feature aggregation, offering expressive, interpretable embeddings that are robust to heterophily. The model-agnostic nature of ℓ -ECT paves the way for hybrid architectures and applications where explainability is critical. Further research on scalable algorithms and deeper integration into GNNs promises to unlock broader adoption in graph learning.

Building on different geometric and topological insights, we also introduced a new Bayesian inference framework that incorporates topological losses into posterior estimation. This approach is particularly suited to simulation-based settings where classical likelihoods are inaccessible, and it extends the scope of uncertainty quantification and parameter recovery for complex models. Our work thus connects topological data analysis not only with classical machine learning but also with fundamental aspects of probabilistic modeling.

Finally, we studied how local intrinsic dimension evolves during the training and fine-tuning of large language models. By treating local dimension as a geometric diagnostic signal, we uncovered unsupervised indicators of generalization, grokking, and overfitting. These findings not only advance the interpretability of LLMs but also point to novel training objectives and auditing tools rooted in the geometry of learned representations.

Across all these contributions, a unifying theme emerges: geometric and topological

8. Conclusion and Future Work

signals offer powerful, underutilized lenses through which to analyze, interpret, and improve machine learning models. Whether as representations, priors, or diagnostics, these signals bridge statistical learning with structural insight, enabling new levels of control and understanding.

Future Outlook

Looking ahead, several promising research directions arise:

Representation Regularization and Control: Embedding Euclidicity or local dimension estimates as constraints in loss functions offers a path toward models that not only generalize better but are structurally aligned with the geometry of the data. This could be especially useful in domains with irregular or singular input spaces.

Scalable Topological Computation: For both ℓ -ECT and Euclidicity, improving computational efficiency is critical. New sampling strategies, filtration methods (e.g., witness complexes), and parallel algorithms will be central to applying these tools at scale.

Interpretability and Auditing Tools: Local intrinsic dimension and topological descriptors hold potential as interpretability primitives. They can support model auditing, data provenance tracing, and the identification of spurious correlations—all without requiring labeled validation data.

Hybrid Models: Integrating local topological and geometric information into deep learning architectures, be it through positional encodings, structural regularization, or attention priors, could lead to new architectures that better balance expressivity and transparency.

Application to Simulation-Based Inference: As models become increasingly simulation-driven (e.g., in climate, epidemiology, or neuroscience), topological loss functions can offer domain-informed guidance where gradients are unavailable and likelihoods are implicit.

In sum, this thesis contributes to recent advances in a broader topological-geometric perspective on learning systems, one that respects the complex structure of data while enabling new forms of inference, explanation, and control. The fusion of topology, geometry, and machine learning is still in its infancy; the results herein suggest that its full potential is far from realized.

A. Appendix

A.1. Pseudocode

We provide brief pseudocode implementations of the algorithms discussed in Section 4.3. In the following, we use $\# \text{Bar}_i(\mathbb{X})$ to denote the number of i -dimensional persistent barcodes of $\mathbb{X}$ (w.r.t. the Vietoris–Rips filtration, but any other choice of filtration affords the same description). Algorithm 4 explains the calculation of *persistent intrinsic dimension*. For the subsequent algorithms, we assume that the estimated dimension of the intrinsic dimension of the data is n . We impose no additional requirements on this number; it can, in fact, be obtained by any choice of intrinsic dimension estimation method. As a short-hand notation, for $p_i = \text{PH}_{n-1}(\mathcal{V}(\text{EB}_\bullet^\bullet(x), \bullet))$ w.r.t. some sample of $\{y \in \mathbb{R}^n \mid r \leq d(x, y) \leq s\}$, we denote by $p_i^{r,s} = \text{PH}_{n-1}(\mathcal{V}(\text{EB}_r^s(x), \bullet))$ the respective fibred persistent local homology barcode (calculated w.r.t. the same sample). Algorithm 5 then shows how to calculate the *Euclidity* values, following Eq. (4.5) and one of its potential implementations, given in Eq. (4.6).

A.2. Stability of Euclidity Estimates

Fig. A.1 shows that Euclidity is robust under sampling; repeating the calculations for smaller batches of the ‘pinched torus’ data set (500 points each) still lets us detect the singularity and its neighbours reliably. This robustness is an important property in practice where we are dealing with samples from an unknown data set whose shape properties we want to capture. Euclidity enables us to perform these calculations in a robust manner. Following the brief discussion in Section 4.4.1, we show the results of varying $s_{\max}$, the outer radius of the local annulus, for the ‘pinched torus’ data set. Fig. A.2 depicts point clouds of 1000 samples; we observe that the singularity, i.e. the ‘pinch point,’ is always detected. For larger radii, however, this detection becomes progressively more *global*, incorporating larger parts of the point cloud. Fig. A.3 depicts the corresponding histograms; we observe the same shift in probability mass towards the tail end of the distribution. For extremely large annuli, we estimate that we lose a clear distinction between singular values and non-singular values. Our data-driven parameter selection procedure is thus to be preferred in practice since it incorporates data density. However, even for markedly varying densities, we find that Euclidity leads to reasonably stable results, as can be seen in Fig. A.4. Here, we calculated Euclidity scores for the pinched torus for varying data sample densities of 50000, 75000 and 100000 samples. Each diagram shows the histogram of Euclidity values for 1000 random samples in the respective

A. Appendix

Algorithm 4 An Algorithm for Calculating the *Persistent Intrinsic Dimension* (PID)

Require: $x \in \mathcal{X}$, $s_{\max}$, ℓ

Ensure: $i_x(s)$ for each $s \in \Gamma$

```

1: for  $s \in \Gamma$  do                                     ▷ Iterate over the parameter grid
2:    $i_x(s) \leftarrow 0$ 
3:   for  $r < s \in \Gamma$  do
4:     for  $i \leftarrow 1$  to  $N - 1$  do
5:       Compute  $\# \text{Bar}_i(B_r^s(x))$ 
6:       if  $\# \text{Bar}_i(B_r^s(x)) > 0$  then
7:          $i_x(s) \leftarrow i + 1$ 
8:       end if
9:     end for
10:  end for
11:  return  $i_x(s)$ 
12: end for

```

Algorithm 5 An Algorithm for Calculating the *Euclidicity Values* δ_{jk}

Require: $x \in \mathcal{X}$, $s_{\max}$, ℓ , n , $\{p_1, \dots, p_m\}$

Ensure: δ_{jk} values

```

1: for  $j \leftarrow 1$  to  $m$  do
2:   for  $k \leftarrow j + 1$  to  $m$  do
3:     for  $s \in \Gamma$  do
4:       for  $r \in \Gamma$  such that  $r < s$  do
5:         Compute  $d_B(p_j^{r,s}, p_k^{r,s})$            ▷ Bottleneck distance
6:       end for
7:     end for
8:     Compute  $D(p_j, p_k)$            ▷ Evaluate corresponding distance
9:   end for
10: end for
11: return  $\delta_{jk}$ 

```

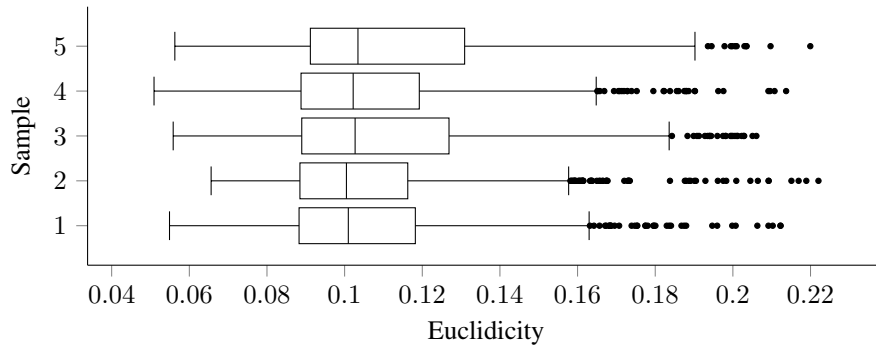


Figure A.1.: Boxplots of the Euclidicity values of different random samples of the ‘pinched torus’ data set. While each sample invariably exhibits some degree of geometric variation, we are able to reliably identify the singularity and its neighbourhood.

A.2. Stability of Euclidity Estimates

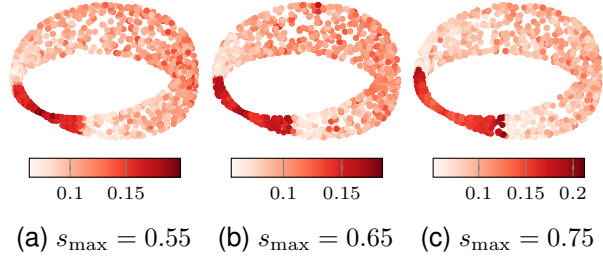


Figure A.2.: Modifying the outer radius $s_{\max}$ still enables us to detect the singularity of the ‘pinched torus.’ Larger radii, however, progressively increase the field of influence of our method, thus starting to assign high Euclidity values to larger regions of the point cloud.

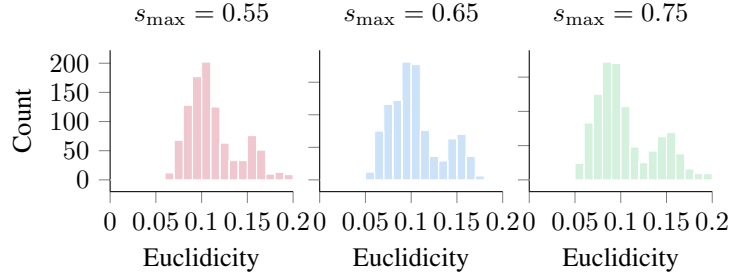


Figure A.3.: Histograms of the Euclidity values for the point clouds shown in Fig. A.2. Larger radii result in the distribution accumulating more probability mass at higher Euclidity values, making the singularity detection procedure less local (but still succeeding in detecting the singularity and its environs).

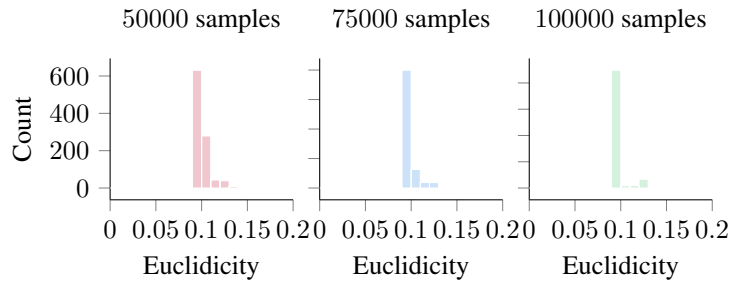


Figure A.4.: Histograms of the Euclidity values for varying point sample densities. Each diagram represents Euclidity values of 1000 random samples in the respective data space.

A. Appendix

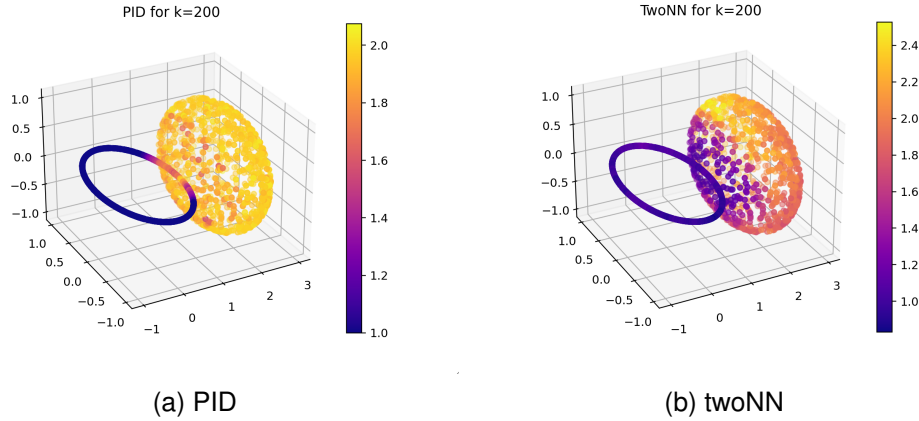


Figure A.5.: Even for large values of k , PID still does not overestimate the local dimensionality of the data, exhibiting a clear distinction between the circle and the sphere, respectively.

data space. Although the histograms differ, the distributions exhibit the same overall characteristics.

A.3. Comparison of PID With Other Dimension Estimates

In order to assess the quality of PID, we decided to test its performance on a space that is both singular and has non-constant dimension. The data space we chose consists of 2000 samples of $S^1 \vee S^2$, i.e. a 1-sphere glued together with a 2-sphere at a certain concatenation point. We then applied the PID procedure for a maximum locality scale that was given by the k nearest neighbour distances, for $k \in \{25, 50, 75, 100, 125, 150, 175, 200\}$. We assigned to each point the average of the PID scores at the respective scales that are less than or equal to the k nearest neighbour bound. Subsequently, we compared the results with other local dimension estimates for the respective number of neighbours. The methods that were chosen for comparison include `lpca`, `twoNN`, `KNN`, and `DANCo`; we used the respective implementation from the `scikit-dimension` Python package.¹

Fig. A.5a shows the PID results for a maximum locality scale of 200 neighbours, with colours showing the estimated dimension values for each point. Overall, the correct intrinsic dimension is detected for most of the points. However, points that lie close to the singular point show a PID value between 1 and 2. Similarly to what we already discussed for Euclidicity, PID should therefore also be interpreted as a measure that incorporates the intrinsic dimension of a point on *multiple scales* of locality. For real-world data, the dimension will generally change when changing the locality scale. However, since there is no canonical choice of scale, we believe that any such scale provides valuable information about the intrinsic dimension that is worth being measured. We therefore argue that a multi-scale approach like ours

¹<https://scikit-dimension.readthedocs.io/en/latest/>

A.3. Comparison of PID With Other Dimension Estimates

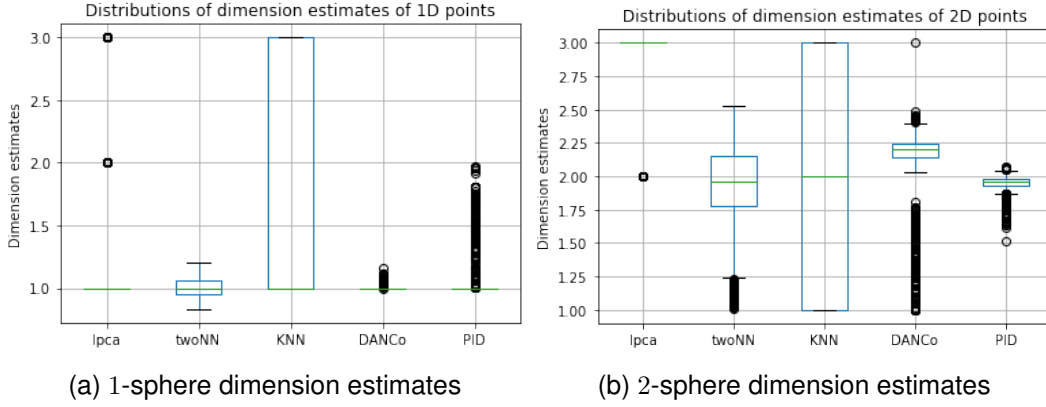


Figure A.6.: Estimates of the local intrinsic dimension for points that are close to the 1D-sphere, i.e. the circle, or the 2D-sphere, respectively.

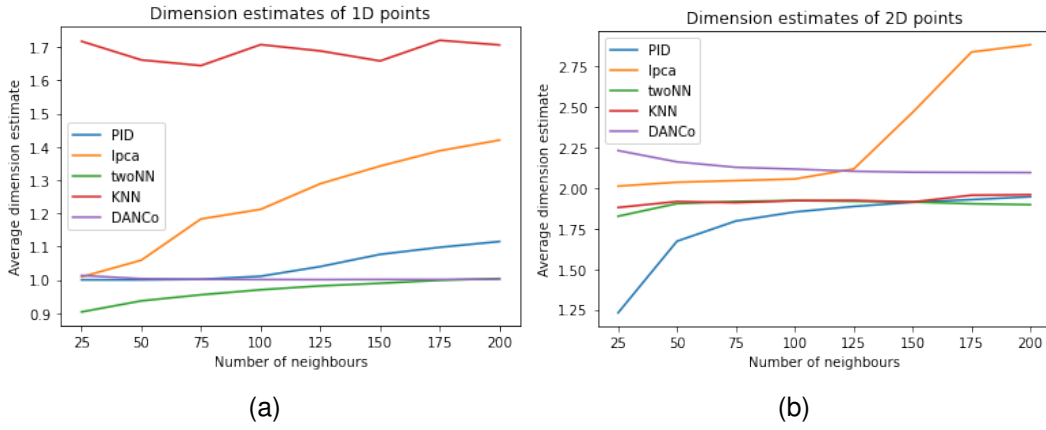


Figure A.7.: Dimension estimates of the 1D-sphere and the 2D-sphere for different methods, plotted as a function of the number of neighbours k .

is appropriate in practice, especially in a regime that is agnostic with respect to the underlying intrinsic dimension. By contrast, Fig. A.5b shows the corresponding dimension estimates for twoNN, where we observe less stable and reliable results across the dataset.

Fig. A.6a shows boxplots of the distributions of the dimension estimates, for all points that lie on the 1D-sphere. We see that for PID, the mass is concentrated at a value of 1. Although there are outliers present, these correspond to points that are close to the singularity, as it was expected. We note that other methods like KNN and lpca might highly overestimate the dimension, and that the interquartile range is significantly higher for twoNN and KNN. Fig. A.6b shows the same distributions for the points that lie on the 2D-sphere. Again, lpca highly overestimates the dimension since the median lies at a value of 3. Again, the interquartile range of PID is the tightest, and the estimates are closest to the ground truth. Moreover, the lower-value outliers again correspond to points that are close to the singular gluing point.

Fig. A.7a and Fig. A.7b show average dimension estimate scores of all investigated

A. Appendix

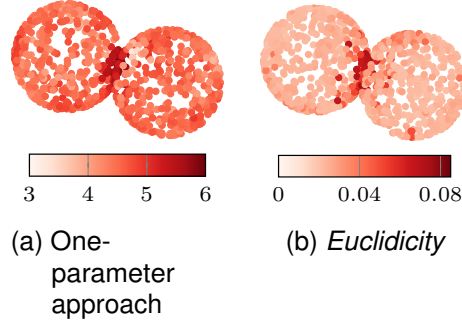


Figure A.8.: In 2D, *Euclidicity* [b](#) results in a clearly-delineated singular region (also shown in the main paper), when compared to a single-parameter score [a](#).

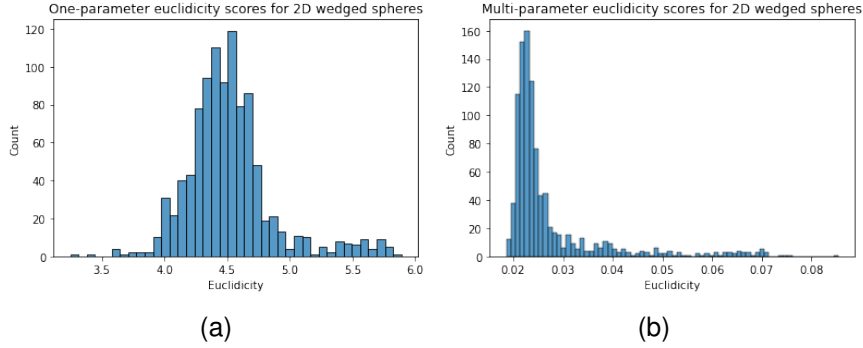


Figure A.9.: A comparison of Euclidicity values of a one-parameter approach (left) and our multi-parameter approach (right) demonstrates that multiple scales are necessary to adequately capture singularities.

methods for varying values of k , both for points on the 1-sphere and the 2-sphere. We note that on average, only `twoNN` and `DANCo` lead to results which are comparable with the reliability of our method. However, as we already saw in Fig. [A.6a](#) and Fig. [A.6b](#), the variance of the scores of our method is significantly lower, leading to more reliable outputs for each of the points.

A.4. Euclidicity is More Expressive than Single-Parameter Approaches

Fig. [A.8](#) depicts a comparison between Euclidicity, an inherently multi-parametric approach, and a single-parameter approach, which assumes the *same* locality properties to hold for all points. This method is similar to the one described by Stolz et al. [173]. We observe that both methods are capable of detecting the isolated singularity of $S^2 \vee S^2$, with Euclidicity leading to a score that more clearly delineates the region around the singularity. Notice that this behaviour will become even more crucial in real-world data, which may exhibit strong sparsity patterns and varying density.

A.5. Euclidity of MNIST and FASHIONMNIST

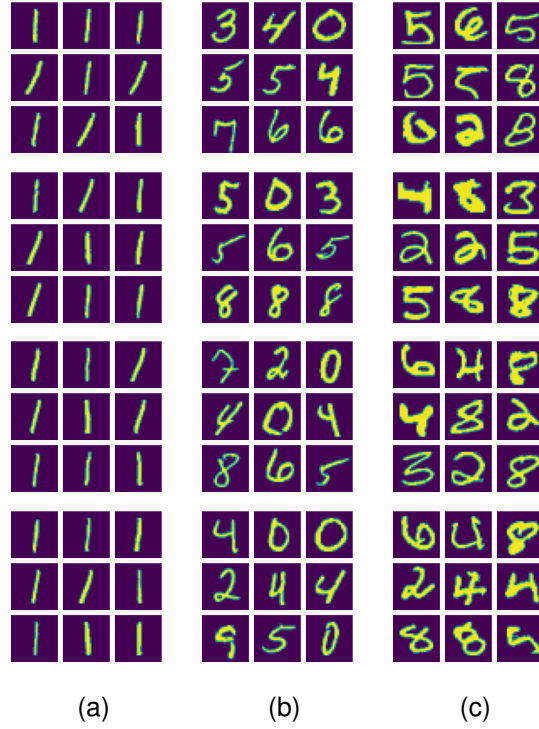


Figure A.10.: From left to right: more examples of low Euclidity values, median Euclidity values, and high Euclidity values for the MNIST data set.

Further underscoring these findings, Fig. A.9 shows the empirical distributions of Euclidity scores for fixed locality parameters (left) and for our proposed multi-scale locality approach (right). We see that the variance is *significantly lower* in the multi-scale regime, indicating more stable and robust results. Moreover, the ratio of maximum and mean is higher in the multi-parameter setting, where high Euclidity scores correspond to data points that lie close to the singularity, resulting in more reliable outcomes.

A.5. Euclidity of MNIST and FASHIONMNIST

Fig. A.10 and Fig. A.11 show the Euclidity results for the 4 additional runs on both the MNIST and FASHIONMNIST data sets. Again, we depicted the 9 images with lowest (left), medium (middle), and highest (right) Euclidity scores for the two datasets. Moving from left to right, the images exhibit increases in the complexity of the local geometry, giving evidence for the reproducibility of the observation we remarked in Section 4.4.6.

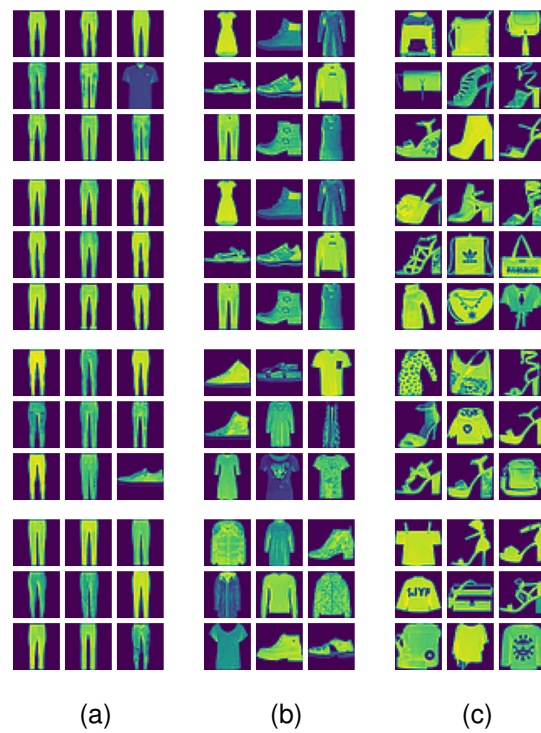
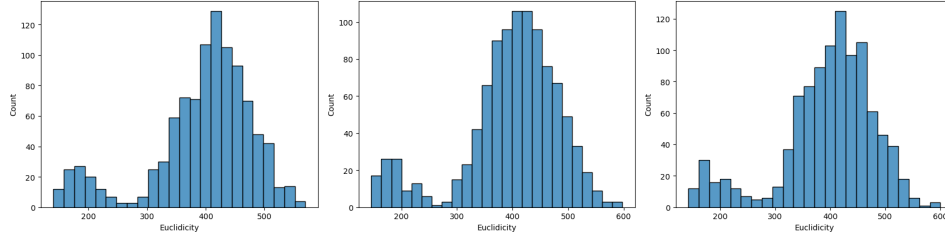


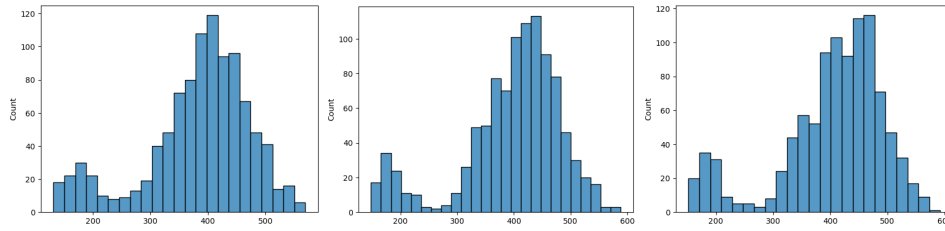
Figure A.11.: From left to right: more examples of low Euclidity values, median Euclidity values, and high Euclidity values for the FASHIONMNIST data set.

A.5. Euclidity of MNIST and FASHIONMNIST



(a) MNIST Euclidity scores for different choices of the intrinsic dimension.

Figure A.12.: Left to right: distribution of MNIST Euclidity scores for intrinsic dimension of 5, 10, and 15, respectively. As a consequence of the stability of our method, the distributions remain similar.



(a) MNIST Euclidity scores for different choices of the nearest neighbour parameter k .

Figure A.13.: Left to right: distribution of MNIST Euclidity scores for $k \in \{25, 50, 75\}$, respectively. As a consequence of the stability of our method, the distributions remain similar.

Moreover, as Fig. A.16 shows, the empirical distributions of the calculated Euclidity scores differ significantly for the MNIST and FASHIONMNIST data sets, with the distribution for MNIST exhibiting a bimodal behaviour, whereas the FASHIONMNIST Euclidity value distribution is unimodal. We hypothesise that this corresponds to regions of simple complexity—and locally linear structures—in the MNIST data set, which are absent in the FASHIONMNIST data set. Finally, we observe that Euclidity values for MNIST are surprisingly stable with respect to specific choices of the intrinsic dimension, as well as the nearest neighbour parameter k , which can be seen in Figs. A.12 and A.13. Here, we calculated Euclidity scores for MNIST for intrinsic dimensions of 5, 10 and 15, respectively, and $k \in \{25, 50, 75\}$. Although the scale of locality for which Euclidity is computed varies significantly for different k , the respective distributions of Euclidity values appear to be almost indistinguishable. We surmise that this is a consequence of the multi-scale nature of our approach.

Euclidity and anomaly scores. Fig. A.14 shows a comparison of Euclidity scores and the scores of *IsolationForest*, an anomaly detection algorithm, for a UMAP projection of MNIST. Here, we used the implementation for *IsolationForest*

A. Appendix

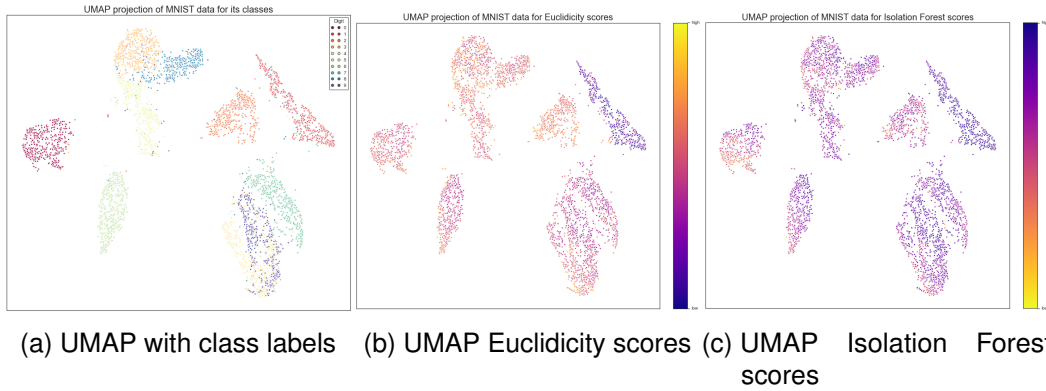


Figure A.14.: Euclidity correlates with Isolation Forest scores for MNIST samples. Notice that we switched the colour maps because *low* Isolation Forest scores indicate anomalies.

from the `scikit-learn` Python package.² The clusters correspond to the different classes of MNIST, see Fig. A.14a. Although the original purpose of Euclidity is the detection of non-Euclidean neighbourhoods of points, we observe a correlation of low Euclidity scores (i.e. points that are closer to admitting a Euclidean neighbourhood) and high `IsolationForest` scores (i.e. points that are less likely to be considered anomalies). This provides further evidence that Euclidity could be expected to serve as a powerful anomaly detection method, on top of its ability to capture local geometric complexity. However, in our experiments we observe a (linear) correlation of only 0.5–0.7, indicating that Euclidity detects *different* anomalous behaviour than just plain outliers. We plan on investigating this in future work.

Details on neural network misclassification analysis. The neural network that we trained for the analysis in Section 4.4.6 consists of an input layer of 784 nodes, one dense hidden layer of 5 nodes and an output layer possessing 10 nodes, respectively. The same architecture was used for both MNIST and FASHIONMNIST, resulting in 3985 trainable parameters. Accuracy scores are 0.777 for MNIST, and 0.714 for FASHIONMNIST. Neither these scores nor the accuracy itself are supposed to be competitive; we merely use this network as an example to outline how Euclidity could be potentially used to inform network training and highlight potentially problematic regions in data. Fig. A.15 shows the respective results of Euclidity scores for misclassified versus correctly-classified samples. To compare the statistical significance of the respective means for correctly and misclassified samples, we performed Welch’s t-test, as the samples are approximately normally distributed. We obtained p-values of 2.8×10^{-5} in the case of MNIST, and 5.7×10^{-3} in the case of FASHIONMNIST. With respect to the common rejection thresholds in the literature, this leads to a clear rejection of the null hypothesis, which assumes the population means to be equal. However, we only claim a *correlation* between the likelihood for samples to be correctly classified and their corresponding Euclidity scores. A statement in terms of *causality* is to be investigated in future work.

²<https://scikit-learn.org/stable/>

A.5. Euclidity of MNIST and FASHIONMNIST

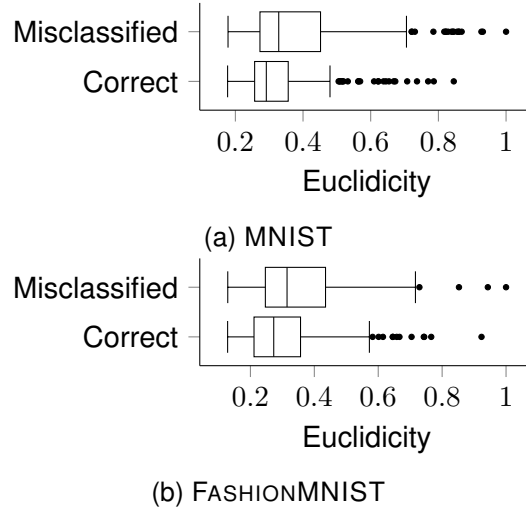


Figure A.15.: A comparison of Euclidity scores for misclassified and correctly classified samples in two image data sets.

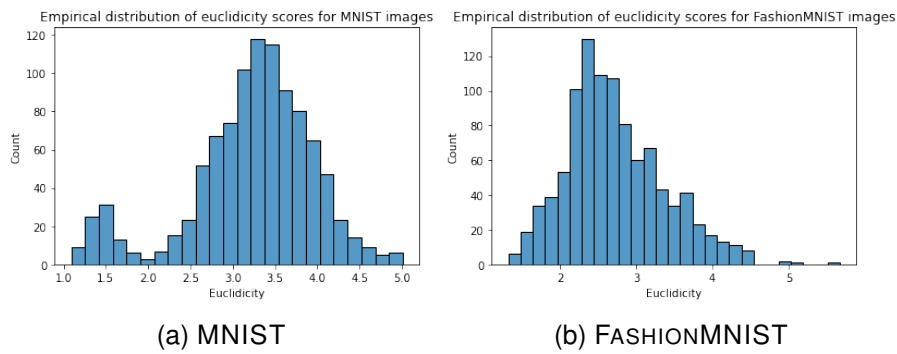


Figure A.16.: Both MNIST and FASHIONMNIST exhibit markedly different distributions in terms of Euclidity: MNIST Euclidity values are bimodal, whereas FASHIONMNIST Euclidity values are unimodal.

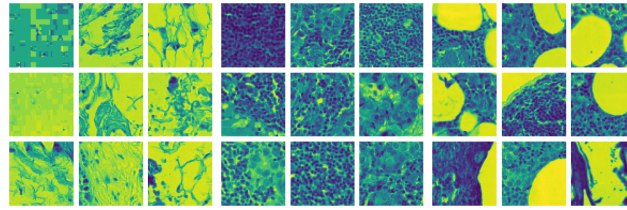


Figure A.17.: Left to right: sample histology images exhibiting low, median, and high Euclidity, respectively. All samples are taken from the ‘PatchCamelyon’ benchmark data set.

A.6. Euclidity Captures Geometric Complexity in Histology Data

Finally, we applied Euclidity to a benchmark histology data set.³ The 96×96 pixel images were first scaled down to 60×60 pixels, to speed up the computation. Moreover, since the images are inherently coloured, we transformed them to greyscale images in order to naturally obtain a high-dimensional representation in Euclidean space, by flattening them. We then randomly picked 1000 of the training images and calculated Euclidity scores of these samples (with respect to the whole space of training images). The intrinsic dimension was chosen to be 35, as suggested by `twoNN`. All the other hyperparameters were chosen as in the setting of Section A.5. The results of images admitting different scores are depicted in Fig. A.17. Similarly to the observations in Section A.5, we see very different structural behaviour inside of the images for different Euclidity values: while low scores correspond to images that admit large voids, medium-score images correspond to images with finer local geometric structure. Images of high Euclidity scores, however, seem to admit *both* large voids *and* regions of fine-grained geometric structure. We therefore hypothesise the potential utility of Euclidity for biological data, for instance to ensure data quality of (histology or biomedical) image datasets.

³<https://github.com/basveeling/pcam>

B. Appendix

B.1. Ablation on directions and interpretability

Coming back to our approximation of $\text{ECT}(X)$ via

$$\overline{\text{ECT}}(X)_{(m,l)} := \text{ECT}(X)|_{\{v_1, \dots, v_m\} \times \{t_1, \dots, t_l\}} \quad (\text{B.1})$$

for uniformly distributed directions $v_1, \dots, v_m \in S^{n-1}$ and filtration steps $t_1, \dots, t_l \in \mathbb{R}$, we notice that the $(l \cdot (j-1) + 1)$ -th till $(l \cdot j)$ -th entries of $\overline{\text{ECT}}(X)_{(m,l)}$ correspond to the direction v_j . The latter gives us the opportunity to get a deeper spatial understanding of how the model predicts its outcome, by analysing its feature importances (which are available for tree-based algorithms like e.g. XGBoost). Therefore, our approach enables us to analyse which features (i.e. directions) of the underlying ECT vector are most important. In practice, we often observe that a small number of features admits high feature importance with respect to the corresponding model, see Fig.B.1. This raises the question if we may use only a smaller random collection of features and still get reasonable results. We therefore ran experiments for a collection of datasets for a varying number of randomly sampled entries of the ℓ -ECT₁ vector see Tab.B.1 for the results. Here, 4096 corresponds to the whole vector. We observe that for certain tasks like e.g. Coauthor CS, Coauthor Physics and Amazon Ratings the performance of the model only slightly changes when using only a small portion of the ℓ -ECT₁ vector. In light of the results in [48], this observation is not entirely surprising: one main claim therein is that the ECT can be determined using only a small number of directions.

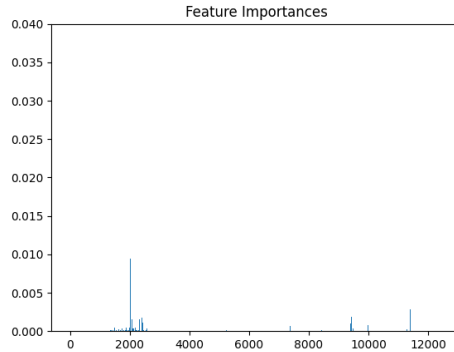


Figure B.1.: Feature importances of an XGBoost model for the Coauthor Physics dataset (using ℓ -ECT₁). Only a small number of features admit high importance scores.

B. Appendix

Table B.1.: Mean accuracy results (5 runs each) for different node classification tasks, and a varying number of randomly sampled entries of the respective ℓ -ECT₁ vectors.

Dataset	50	100	500	1000	4096
WikiCS	69.2%	70.5%	71.3%	72.7%	74.6%
Coauthor CS	92.3%	92.4%	92.5%	92.6%	92.6%
Coauthor Physics	95.6%	95.6%	95.8%	95.9%	96.1%
Roman Empire	73.7%	75.8%	78.3%	79.7%	80.4%
Amazon Ratings	47.9%	48.2%	48.4%	48.2%	48.4%

B.2. Spatial Alignment of High-Dimensional Data

Following our previous observations that d_{ECT} enables us to align two spaces, we now use it to investigate its effect on high-dimensional data. We start this discussion with the well-known MNIST benchmark dataset. We first represent each (gray-scale) image in the dataset as a 784-dimensional vector, by flattening the image. In this way, we obtain a high-dimensional point cloud corresponding to the dataset. Subsequently, we sample 300 points of digits of ‘1’ and calculate the pairwise distances of their respective ℓ -ECT (with respect to the whole point cloud), for $k = 10$. Finally, we calculate the pairwise distances of the respective aligned ℓ -ECTs (by using the approach of Eq. 5.7 and for $k = 10$). The results are shown in Fig. B.2. We see that the aligned ℓ -ECTs have a significantly lower squared L2 distance (with a median of ≈ 112) than the non-aligned ones (with a median of ≈ 224), showcasing that rotations cause dissimilarity between small neighborhoods of points, in many cases.

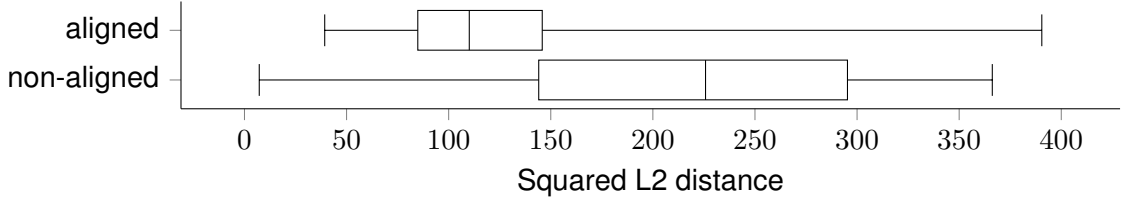


Figure B.2.: A comparison of the squared L2 distances of ℓ -ECTs of aligned and non-aligned MNIST digits of ‘1’, respectively.

B.3. Post-hoc evaluation of node classification experiments

A critical difference diagram arranges the average ranks of multiple models across a set of datasets in order to facilitate overall performance comparisons between the model performances. Fig. B.3 shows the results for all node classification results given in Sec. 6.3 which include both homophilic and heterophilic graph datasets. We see that the ℓ -ECT-based approaches outperform standard methods and the heterophily-specific architecture H2GCN by far, when averaged over all datasets

B.4. Spatial alignment of wedged spheres

¹. The best performing method ℓ -ECT₁ + ℓ -ECT₂ exhibits an average rank of 2, while the worst performing method is GIN with an average rank of 5.7. Even the worst performing ℓ -ECT-based method (ℓ -ECT₂) performs better than the best non- ℓ -ECT-based method GAT.

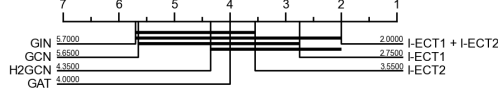


Figure B.3.: Critical difference diagram for ranks of different models across all node classification tasks in Sec.5.4. Even the worst performing ℓ -ECT-based approach (ℓ -ECT₂) works better than all other methods, when averaged over all tasks.

To evaluate the performance of our methods in comparison to those reported in the literature, we included a comparison with the results presented in [139], utilizing the ranks of the respective models as the basis for evaluation. The results are shown in Tab.B.3, where the table presents the ranking of various models on heterophilic node classification tasks. Among the listed methods, several, such as H2GCN, CPGNN, and GPR-GNN, are explicitly designed for heterophilic graph settings, leveraging specialized architectures to handle the challenges posed by such data. In contrast, our ℓ -ECT₁ + ℓ -ECT₂ method, despite being a general-purpose approach not tailored specifically for heterophilic settings, achieves a competitive rank of 11. This performance is on par with other top-performing heterophily-specific models, such as GloGNN, and outperforms well-established architectures like GT and GAT by a significant margin. These results highlight the robustness and adaptability of our method, demonstrating its ability to handle diverse graph structures effectively without requiring customization for heterophilic scenarios. In consideration of the results given in Fig.B.3, this makes ℓ -ECT-based approaches a versatile general-purpose solution for various node classification tasks.

B.4. Spatial alignment of wedged spheres

We approximate the optimization problem in Eq. 5.7 to show that we can learn a spatial alignment of two given data spaces, while the distance between ECTs of non-aligned spaces that only differ up to a rotation will in fact generally be high. We start with a so-called wedged sphere, meaning two 2-dimensional spheres which are concatenated at a gluing point (see Fig. B.5). We use 2000 uniformly sampled points from such a wedged sphere, and compare the squared L2 loss between the ECTs of this sample and a rotation of the same data space. We repeat this procedure 500 times, where at each step both the sample of the wedged sphere and the rotation matrix which yields the rotated version of the same space are randomly sampled. We notice that the L2 losses between the non-aligned spaces are high (with a median of around 19), whereas the L2 losses of the non-aligned spaces are

¹We used <https://github.com/hfawaz/cd-diagram> for the creation of the critical difference diagram.

B. Appendix

Model	Rank
H2GCN	18.250
CPGNN	16.750
GPR-GNN	15.250
ResNet	13.750
I-ECT1	12.375
I-ECT2	12.375
GAT	12.250
GT	11.000
I-ECT1 + I-ECT	11.000
GloGNN	11.000
ResNet+SGC	10.750
FAGCN	10.000
JacobiConv	9.750
GCN	9.625
GBK-GNN	9.000
ResNet+adj	7.250
SAGE	5.875
GAT-sep	5.500
GT-sep	5.250
FSGNN	3.000

Table B.2.: Ranks of the various models in [139] across the heterophilic datasets therein, in comparison to our methods.

significantly lower, with a median loss close to zero (see Fig. B.4 in the appendix for more details). Moreover, we see that the ECT of the same space significantly changes when the coordinate system is changed, which undermines the necessity of a rotation-invariant metric for a comparison of ECTs that we introduced in Sec. 6.3. We conclude that an alignment of the ECTs of the two underlying data spaces in fact leads to an alignment of the data spaces itself, as promised by the theoretical results in Sec. 6.3.

Robustness with respect to outliers and noise Fig. B.6 and Fig. B.7 in the appendix show that the spatial alignment of wedged spheres still works in the presence of outliers and noise. This property is an important feature when dealing with real-world data, which is often noisy, and enables us to align spaces that only approximately differ up to a rotation. By contrast, the Hausdorff distance which is a widely used metric between point clouds is (by definition) highly sensitive to outliers. We therefore conclude that the proposed metric based on ECTs is a robust metric to compare point clouds of potentially different cardinalities.

Fig. B.4 shows that the L2 losses between the non-aligned spaces are high (with a median of around 19), whereas the L2 losses of the non-aligned spaces are significantly lower, with a median loss close to zero. We therefore see that the ECT

B.4. Spatial alignment of wedged spheres

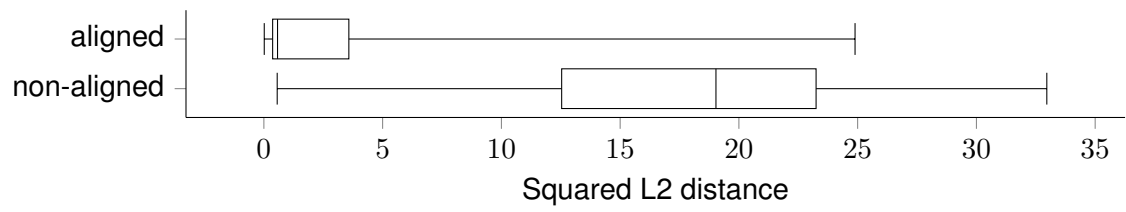


Figure B.4.: A comparison of the squared L2 distances of the respective ECTs of aligned and non-aligned wedged spheres, respectively.

of the same space significantly changes when the coordinate system is changed, which undermines the necessity of a rotation-invariant metric for a comparison of ECTs that we introduced in Sec. 6.3.

B. Appendix

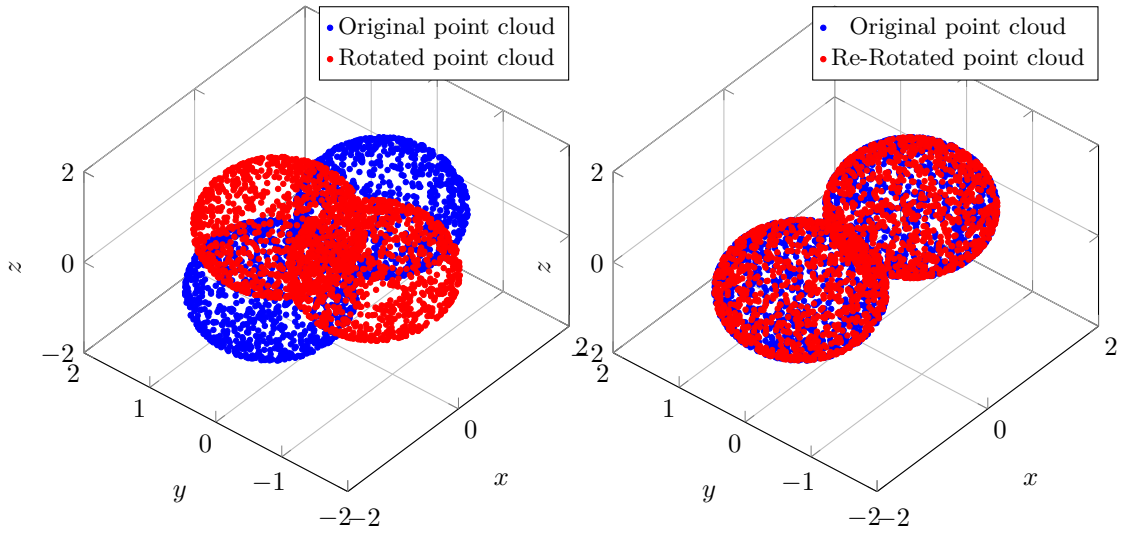


Figure B.5.: A comparison of two wedged spheres with one being rotated around the wedge point and the points being perturbed by Gaussian noise (left) and the learned re-rotated sphere that is aligned with the original data (right).

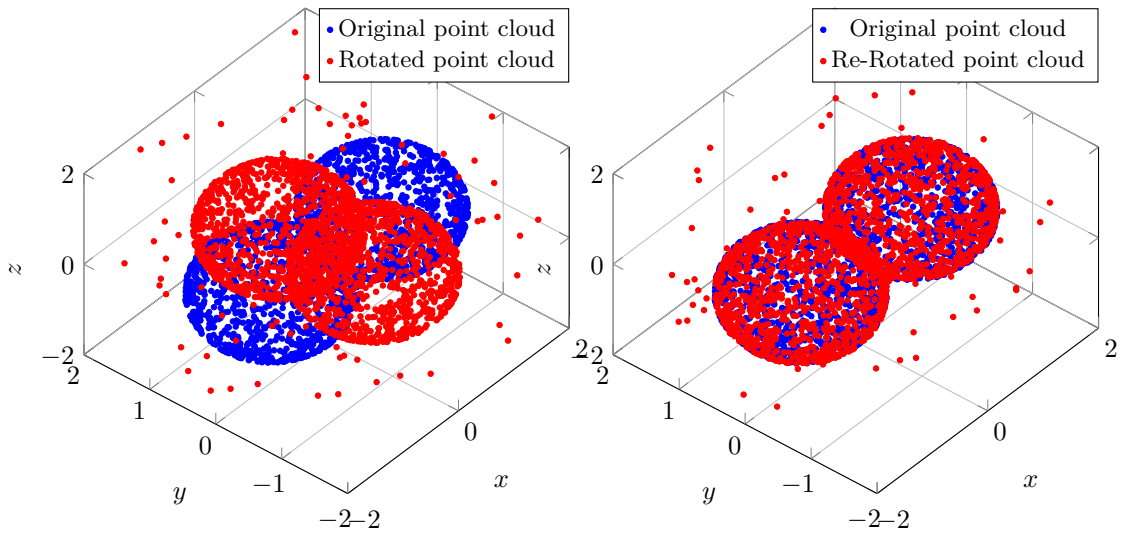


Figure B.6.: A comparison of two wedged spheres with one being rotated around the wedge point and added 200 outliers (left) and the learned re-rotated sphere that is aligned with the original data (right).

B.4. Spatial alignment of wedged spheres

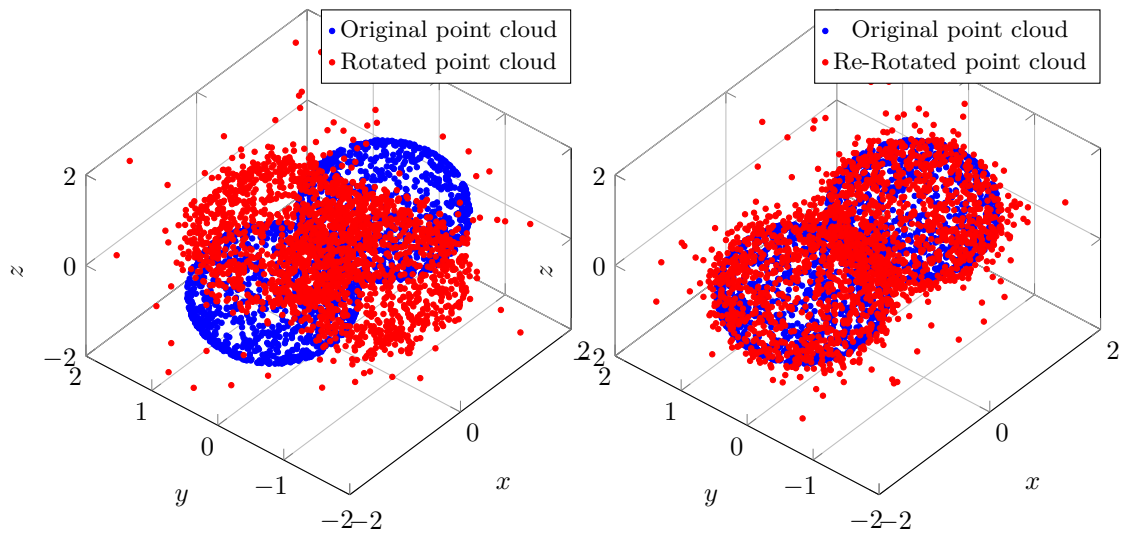


Figure B.7.: A comparison of two wedged spheres with one being rotated around the wedge point and the points being perturbed by Gaussian noise (left) and the learned re-rotated sphere that is aligned with the original data (right).

C. Appendix

C.1. Synthetic data parameter estimates.

As a consistency check and illustrative example, we use our proposed methods for inferring the radius parameters of two objects, namely the 2-sphere and the 2-torus. The 2-sphere is the surface of a closed 3-dimensional ball of radius r , while the 2-torus is constructed using an *inner radius* r and an *outer radius* R . The point clouds were sampled uniformly from the respective surface of known parameter.

Table C.1 depicts the results of the inference process. We observe that both types of losses are capable of inferring the right parameters of these simple geometric objects, with our topology-based loss ℓ_T providing slightly more reliable estimates in the case of a torus. Table C.1 shows the raw values of learning radius parameters for synthetic data sets (2D spheres and 2D tori). This experiment primarily shows that for such simple shapes, geometry-based and topology-based losses perform equivalently. For more complex data sets, however, we find that the improved robustness of our topology-based loss term helps in inferring the ground truth parameters.

C.2. Vicsek model

Swarm behaviour of the Vicsek Model. Fig. C.1 shows simulations of the Vicsek model for different noise parameters η . The higher η , the more ‘chaotic’ the behaviour of the resulting complex system.

Table C.1.: Estimating the radius parameters of a 2D sphere (r) and a 2D torus (r, R). For the more complex torus parameter estimation, we observe that our topological loss ℓ_T provides more reliable estimates than the geometrical loss ℓ_G . The results show the mean of five runs for each parameter choice.

Loss	Sampling method	Sphere radius r			Torus radii (r, R)				
		1.00	5.00	10.00	(1, 2)	(3, 5)	(5, 10)		
ℓ_G	Importance sampling	0.97	4.94	9.91	0.94	1.96	2.99	4.95	4.98
	MCMC	1.00	4.99	10.00	0.99	2.04	2.77	4.96	5.01
ℓ_T	Importance sampling	1.01	5.02	9.98	1.02	1.96	3.01	5.05	4.95
	MCMC	0.99	4.99	10.00	1.04	1.89	2.99	4.97	5.01

C. Appendix

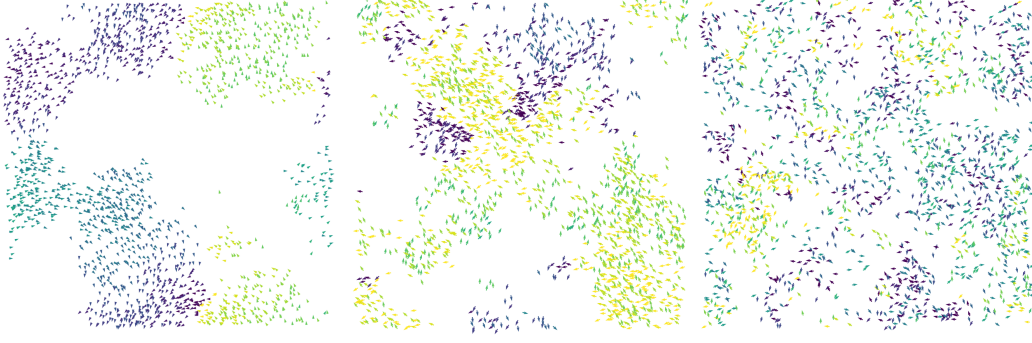


Figure C.1.: Left to right: simulations of the Vicsek model for $\eta = 0.15, 0.3$ and 0.6 , respectively. For lower values of η swarming behaviour emerges, whereas for higher values of η swarms tend to merge and the model becomes more ergodic.

Vicsek model parameter estimates. Accompanying Fig. 6.2, we show the raw parameter estimates of η , the noise parameter of the model in Table C.2. Moreover, we show the results of a *rejection sampling* procedure with summary statistics (mean and standard deviation, respectively), in Fig. C.2. In the latter setting we used the L_2 distance between the respective summary statistics as the loss that determines the results of our method. We observe that there is no statistic that clearly discriminates the other one, and that the performance of the results highly depends on the parameter and time step that is considered. We find that the overall accuracy and reliability of the respective experiments by using topological losses is significantly higher; see Fig. 6.2 for a comparison.

Table C.2.: Noise parameter estimation after $n = \{5, 10, 50\}$ iterations of the Vicsek model. Estimates of η based on a topological loss are always closer to the ground truth value than estimates obtained via a geometrical loss.

	<i>Loss</i>	<i>Sampling method</i>	True parameter η		
			0.15	0.30	0.60
$n = 5$	Hausdorff distance	Importance sampling	0.17	0.35	0.67
		MCMC	0.13	0.36	1.15
	Topological distance	Importance sampling	0.14	0.29	0.59
		MCMC	0.12	0.33	0.63
$n = 10$	Hausdorff distance	Importance sampling	0.11	0.39	0.70
		MCMC	0.18	0.41	1.95
	Topological distance	Importance sampling	0.18	0.27	0.65
		MCMC	0.13	0.34	0.62
$n = 50$	Hausdorff distance	Importance sampling	0.12	0.29	0.72
		MCMC	0.28	1.40	2.28
	Topological distance	Importance sampling	0.16	0.31	0.61
		MCMC estimate	0.21	0.24	0.59

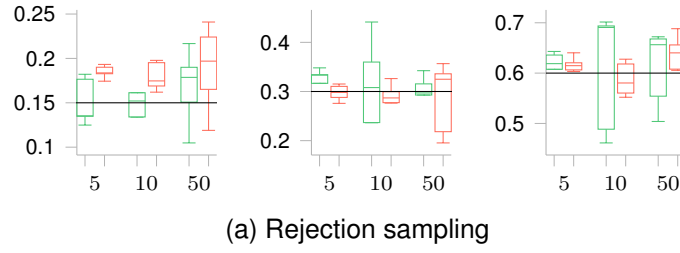


Figure C.2.: Summary of parameter estimates for the noise parameter η of the Vicsek model with respect to summary statistics and a standard rejection sampling procedure, using 5 repetitions of the inference procedure. Each subplot depicts estimates for $\eta \in \{0.15, 0.30, 0.60\}$, respectively. The true parameter is shown as a solid black line. The horizontal axis shows the number of iterations of the Vicsek model. Estimates based on the **mean statistic L_2 loss** and **standard deviation statistic L_2 loss** both are significantly less accurate than the respective topological loss, see Fig. 6.2b.

C.3. Lattice Boltzmann model

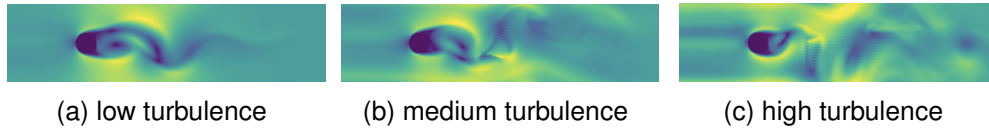


Figure C.3.: Left to right: simulations of the fluid model after 3000 iterations, for $\eta = 0.2, 0.3$ and 0.4 , respectively. For lower values of η the fluid flow is less turbulent than for higher values of η .

The degree of randomness in the local propagation and collision steps in the lattice Boltzmann method (LBM) model controls the turbulent behaviour of the global system, as is illustrated in Fig. C.3. This is the parameter η that we estimate in our experiments. The results are shown in Fig. C.4 and Fig. C.5, respectively. Using importance sampling, the topological loss outperforms the geometrical loss in terms of accuracy, as can be seen from the median and interquartile range of the boxplots in Fig. C.4. Moreover, the topological loss in combination with importance sampling also outperforms the geometrical loss with MCMC sampling, as evidenced from Table 6.1. The topological loss in combination with MCMC sampling, however, is far off in terms of accuracy. The latter indicates that the convergence rates of MCMC sampling and importance sampling may be very different, and depend on the given setting. Finally, we note that rejection sampling with both the mean and standard deviation summary statistic does not perform reliably, and its accuracy highly depends on the true parameter in the underlying observation.

C. Appendix

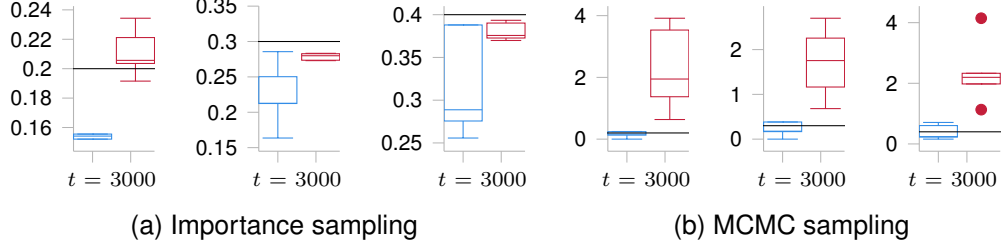


Figure C.4.: Summary of parameter estimates for the parameter η of the fluid model, using 5 repetitions of the inference procedure. Each subplot depicts estimates for $\eta \in \{0.20, 0.30, 0.40\}$, respectively. The true parameter is shown as a solid black line. Estimates based on the **geometrical loss ℓ_G** are less accurate than the **topological loss ℓ_T** .

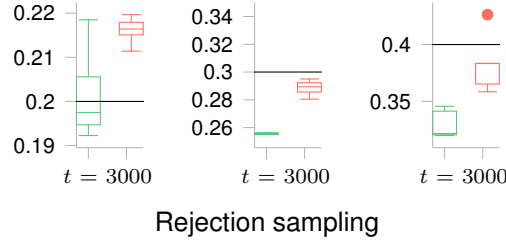


Figure C.5.: Summary of parameter estimates for the parameter η of the fluid dynamics model with respect to summary statistics and a standard rejection sampling procedure, using 5 repetitions of the inference procedure. Each subplot depicts estimates for $\eta \in \{0.20, 0.30, 0.40\}$, respectively. The true parameter is shown as a solid black line. Estimates based on the **mean statistic L_2 loss** and on the **standard deviation statistic L_2 loss** tend to be less reliable than estimates based on the topological loss, see Fig. C.4.

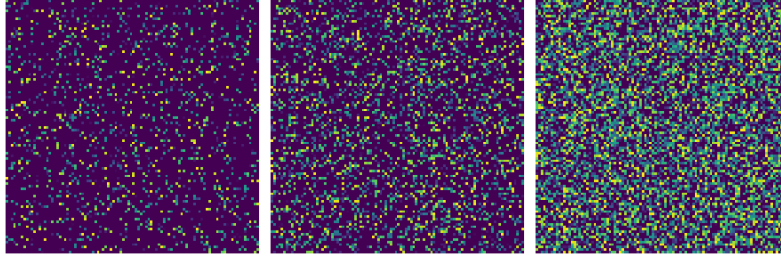


Figure C.6.: Left to right: samples of the proposed percolation model for $p = 0.15, 0.3$ and 0.6 , respectively. All samples admit a fixed maximum greyscale value of 50. Although the geometric distance between two samples of fixed p can be large due to the uniform sampling of pixels, the overall topological structure (which is captured by persistent homology) for a given p is more stable with respect to sampling.

C.4. Multi-scale structures in percolation models

Fig. C.6 shows an illustration of our proposed percolation model. The higher the value of p , the more likely it is for a pixel to be non-zero. Consequently, there are more non-zero pixels for higher values of p . As discussed in the main paper, the topological distance outperforms the SCC distance significantly, with both importance sampling and MCMC sampling. However, the rejection sampling procedure together with the mean statistic performs even more accurate in this experiment, as can be seen by comparing Fig. C.7 and Fig. C.8. This is not surprising: once the pixel size converges to infinity, the mean of the pixel values will converge to the expected pixel value, where the latter only depends on p (since the maximum greyscale value is fixed in our experiments, and therefore does not have any discriminative power for inference). Therefore, at least in a large-pixel regime we can infer the true parameter of the observation from the mean of its pixel values. Note that this is due to the additional structure in this experiment, and does not hold for complex systems that cannot be described in such a simple way, as has been seen in the previous experiments. Finally, rejection sampling with the standard deviation statistic leads to highly erroneous results, as can also be seen from Fig. C.8. For complex systems it is therefore very difficult to determine the ‘right’ summary statistic which contains a sufficient amount of information to infer the underlying parameter. This choice is not necessary when using topological losses, which makes the latter an appropriate generic choice, in many applications.

C.5. Description of Algorithm 2

Inputs:

- **Observed Data y :** The actual data you have collected.

C. Appendix

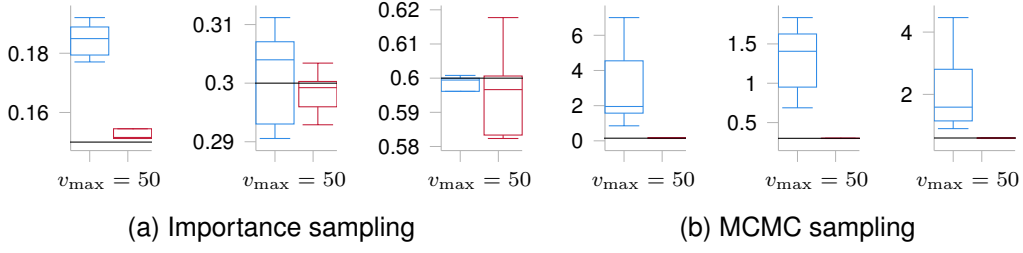


Figure C.7.: Summary of parameter estimates for the parameter p of the Percolation model, using 5 repetitions of the inference procedure. Each subplot depicts estimates for $p \in \{0.15, 0.30, 0.60\}$, respectively. The true parameter is shown as a solid black line. Estimates based on the **geometrical loss** ℓ_G are less accurate than the **topological loss** ℓ_T , in particular for larger number of iterations.

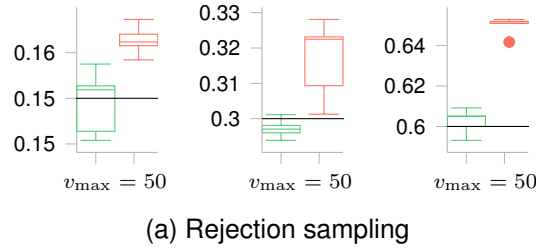


Figure C.8.: Summary of parameter estimates for the parameter p of the Percolation model with respect to summary statistics and a standard rejection sampling procedure, using 5 repetitions of the inference procedure. Each subplot depicts estimates for $p \in \{0.15, 0.30, 0.60\}$, respectively. The true parameter is shown as a solid black line. Estimates based on the **mean statistic L_2 loss** are more accurate than the **standard deviation statistic L_2 loss**.

Table C.3.: For 500 simulation points: Estimating the randomness parameter of the lattice Boltzmann model after 3000 simulation iterations, using 5 repetitions of the inference procedure. The topological loss together with importance sampling outperforms all of the other configurations.

Loss	Sampling method	Randomness parameter η		
		0.20	0.30	0.40
ℓ_G	Importance sampling	0.14 ± 0.02	0.22 ± 0.02	0.26 ± 0.03
	MCMC	0.18 ± 0.07	0.21 ± 0.03	0.22 ± 0.07
ℓ_T	Importance sampling	0.21 ± 0.01	0.28 ± 0.01	0.38 ± 0.05
	MCMC	2.78 ± 1.82	2.91 ± 2.11	2.35 ± 0.80

- **Start Value θ_0 :** An initial guess for the parameter θ , whose posterior distribution one intends to estimate.
- **Proposal Distribution $q(\vartheta | \theta)$:** A distribution used to propose new values for the parameter θ , based on its current value.

Steps:

Initialize: Start by sampling X_0 from the distribution $p(x | \theta_0)$, where $p(x | \theta)$ represents the likelihood of the data x given a parameter θ . This initializes the first sample of the latent variable X_0 .

Iterative Sampling (Loop from $i = 1$ to K): For each iteration, repeat the following steps to generate a sequence of samples that approximate the distribution of θ :

- **Generate Random Number U :** Sample $U \sim \text{Unif}[0, 1]$. This will be used for the acceptance/rejection step.
- **Propose a New Parameter Value θ' :** Sample a new candidate value θ' from the proposal distribution $q(\cdot | \theta_{i-1})$, which suggests a new possible value for the parameter based on the current value θ_{i-1} .
- **Sample a New Latent Variable X' :** Sample a new value for the latent variable X' from the distribution $p(x | \theta')$, corresponding to the newly proposed parameter θ' .
- **Compute the Acceptance Ratio $a(\theta_{i-1}, \theta')$:** Calculate the acceptance ratio, which compares the likelihood of the data given the new parameter θ' and the current parameter θ_{i-1} , adjusted by their prior probabilities and proposal distributions. This ratio helps decide whether the new proposed value θ' should be accepted or not.
- **Acceptance/Rejection Step:** Compare U (the random number sampled earlier) with the acceptance ratio $a(\theta_{i-1}, \theta')$:
 - If $U \leq a(\theta_{i-1}, \theta')$, accept the new proposed parameter θ' and set $\theta_i = \theta'$, and also update the latent variable to $X_i = X'$.
 - Otherwise, reject the proposed value and retain the current parameter and latent variable, i.e., set $\theta_i = \theta_{i-1}$ and $X_i = X_{i-1}$.

Repeat: Continue this process for a total of K iterations.

Output: After completing the iterations, the algorithm outputs a sequence of sampled parameter values $\theta = (\theta_1, \theta_2, \dots, \theta_K)$, which approximates the posterior distribution of the parameter given the observed data.

C.6. Pseudo-marginal MCMC in a nutshell

Pseudo-marginal Markov Chain Monte Carlo (MCMC) methods are a class of algorithms used for sampling from a posterior distribution when the likelihood function is intractable or expensive to compute. In such scenarios, the exact likelihood is replaced by an unbiased estimator, which introduces randomness into the acceptance ratio of the Metropolis-Hastings (MH) algorithm. Despite this randomness, pseudo-marginal MCMC methods can still produce samples from the correct target distribution. In Bayesian inference, one often needs to sample from a posterior distribution $\pi(\theta | y) \propto \pi(\theta)p(y | \theta)$, where $\pi(\theta)$ is the prior, and $p(y | \theta)$ is the likelihood function. In many practical cases, such as when dealing with complex models, the likelihood $p(y | \theta)$ may not be analytically tractable or computationally feasible to evaluate directly. Pseudo-marginal MCMC provides a solution by replacing the likelihood with an unbiased estimator $\hat{p}(y | \theta)$. This approach retains the correct target distribution $\pi(\theta | y)$ because the pseudo-marginal method treats the noisy estimate $\hat{p}(y | \theta)$ as if it were the exact likelihood within the Metropolis-Hastings framework. The pseudo-marginal MCMC algorithm proceeds similarly to the standard Metropolis-Hastings algorithm, with the key difference being that the likelihood ratio in the acceptance probability is replaced by a ratio of likelihood estimators. Specifically, given a current state θ_t , a new state θ' is proposed, and the acceptance probability is calculated as:

$$\alpha(\theta_t, \theta') = \min \left(1, \frac{\hat{p}(y | \theta') \pi(\theta') q(\theta_t | \theta')}{\hat{p}(y | \theta_t) \pi(\theta_t) q(\theta' | \theta_t)} \right),$$

where $\hat{p}(y | \theta')$ and $\hat{p}(y | \theta_t)$ are unbiased estimates of the likelihood, and $q(\theta' | \theta_t)$ is the proposal distribution. Crucially, the unbiasedness of the likelihood estimator ensures that, despite the randomness introduced, the chain remains reversible and converges to the correct posterior distribution $\pi(\theta | y)$. See Andrieu and Roberts [5] for a thorough introduction to pseudo-marginal MCMC.

D. Appendix

D.1. Sensitivity Analysis

The following analysis gives insights into the stability of the sampling process in our pipeline of 3. In particular, we discuss the influence of the following variations on our final mean local estimate:

- In Section D.1.1, we show that with increasing sample and neighborhood sizes, the TwoNN estimates tend to converge in our setting, and already a small portion of samples suffices for reasonable reliability.
- In Section D.1.2 we show that the mean local estimates are stable under applying noise to the embedding vectors.
- In Section D.1.3, we show that for masked language models, both masked and regular token embeddings lead to similar mean estimates.

All embeddings in the following analyses are computed using the regular token embeddings from the last layer of the RoBERTa base model.

D.1.1. Dependence on hyperparameter choices

Dataset sequence sub-sampling (dependence on M) Text corpora data splits usually contain different numbers of text sequences, since, for example, the training dataset is larger than the validation or test dataset. Here, we present experiments to show that our method can be made stable in the size of text corpus sequence sub-sample $M \in \mathbb{N}$, and how it depends on the seed of the sequence sub-sample.

Note that in our sub-sampling, we shuffle the dataset split using a given seed, and then truncate to a beginning segment of the specified size. Thus, when comparing different data sequence sub-sampling sizes for a fixed seed, the smaller sequence sub-samples are subsets of the larger ones. In all the experiments in this section, the size of token sub-sample is set to $N = 60\,000$, and the local neighborhood size parameter $L = 128$. Note that for the smallest sequence sub-sample size $M = 2000$, not enough non-padding tokens are necessarily available to reach a token sample size of $N = 60\,000$, but with more sequences, we get this common token space size.

We take sequence sub-samples of each given size with 5 different seeds and select a random token sub-sample for each of these, on which the local TwoNN estimates are based. For the results, see Figure D.1 and Figure D.2.

D. Appendix

For a given sequence sample size, the average of the mean local estimates, calculated across different sampling seeds, remains consistent between the various dataset splits of the same dataset. This observation indicates that the sampling process is robust to differences in how the data is divided. Furthermore, the estimates stabilize as the sequence sample size increases, suggesting convergence of the local mean estimates with larger samples. Based on these observations, we set M to the values 7000 or 10 000 for the natural language-based tasks in Sections 7.4.1, 7.4.3 and 7.4.4, because it represents a practical compromise between computation effort and stability of the value at that scale. This value provides sufficient sampling density to capture the underlying data distribution comparably across datasets while maintaining computational efficiency.

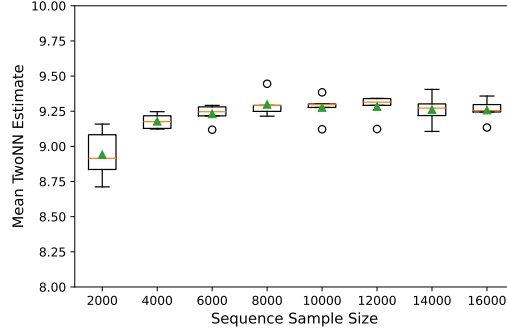
Token sub-sampling (dependence on N) Here, we want to investigate the influence of the token sub-sample size on the local TwoNN estimate. We conduct our analysis on the MultiWOZ dataset. Since here we are interested in the effect of the second sub-sampling process, we want to fix the first sampling step for the sequences: We take the sub-sample of the first $M = 10\,000$ sequences from the MultiWOZ training, validation, and test split. From the resulting token embedding spaces, we then sample 100 000 points (ignoring padding and end-of-sequence (EOS) tokens). Finally, we compute the local estimates based on the first $N = 10\,000, 20\,000, \dots, 90\,000, 100\,000$ points in the given sub-sample. Note that we are left with slightly fewer points in the local estimation step for the largest sub-sample size because of our de-duplication step in 3. This means that in a given token sub-sample, the beginning segment of the first 10 000 points is the same even when increasing the sample size, and we compare the mean of the local estimates of this beginning segment.

The analysis of the datasets reveals a clear trend in the behavior of the truncated mean values and their associated standard deviations as the sample size increases, see Figure D.3. Here, the mean stabilizes quickly, and the standard deviation is comparatively small. This justifies our choice of $N = 60\,000$ in the natural language dataset settings.

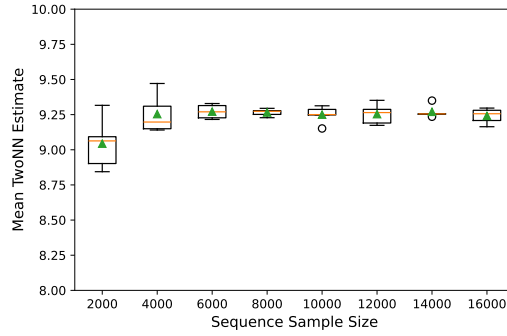
Neighborhood size for local estimates (dependence on L) Having fixed sample sizes M and N , Figure D.4 shows the distribution of TwoNN dimension estimates across varying locality scale parameters, represented by the number of neighbors L . A key observation is the relative stability of the estimates as the number of neighbors increases, with smaller locality scales (fewer neighbors) yielding more diverse estimate distributions. This indicates that local properties of the data significantly influence dimension estimates.

The increasing stability with larger neighborhoods suggests that the embedding space exhibits more homogeneity at broader locality scales. However, the variability at smaller scales aligns with the observation in [23] that in many cases, numerical data is not confined to a single manifold of uniform dimensionality but rather consists

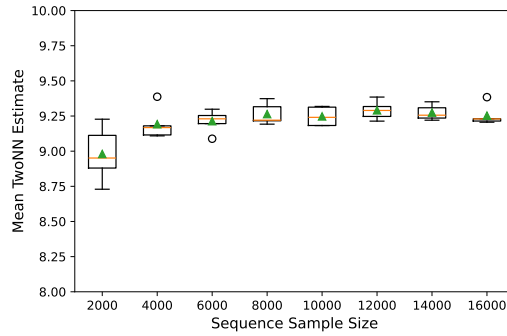
D.1. Sensitivity Analysis



(a) MultiWOZ Training split



(b) MultiWOZ Validation split



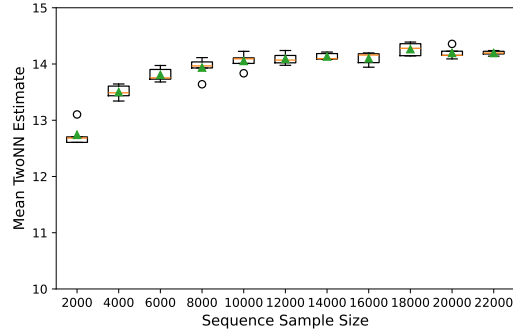
(c) MultiWOZ Test split

Figure D.1.: Boxplots of the mean local TwoNN estimates for different sequence sample sizes M ranging from 2000 to 16 000. We compute the mean estimates for 5 different sequence sub-sampling seeds for the three splits of the MultiWOZ dataset. Here and in all subsequent boxplots, the average of the mean estimates is depicted as green triangles, their median is the orange line; outliers are shown as circles.

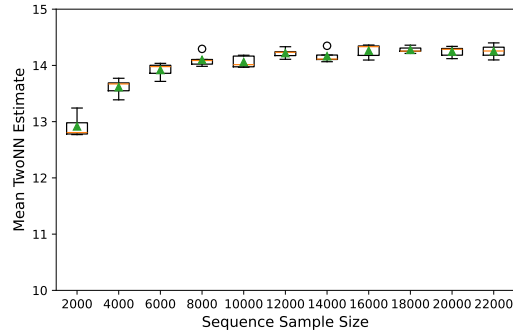
of multiple manifolds or regions with varying local dimensions.

Our findings provide evidence that token embedding spaces of this nature do not conform to a single, unique dimension. Instead, they exhibit a complex structure, where local estimates diverge, reflecting the diverse geometric and topological properties within the space. This emphasizes the importance of considering locality

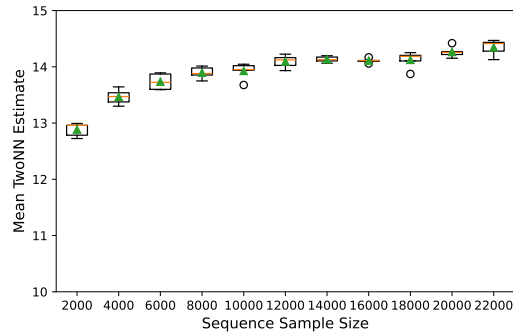
D. Appendix



(a) Reddit Training split



(b) Reddit Validation split



(c) Reddit Test split

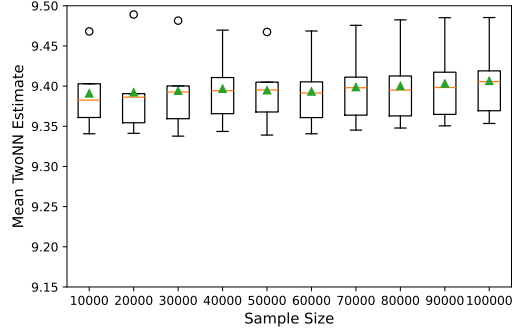
Figure D.2.: Mean local estimates for different sequence sample sizes M for splits of the Reddit dataset, shown for 5 sampling seeds.

when evaluating dimension estimates in such embedding spaces.

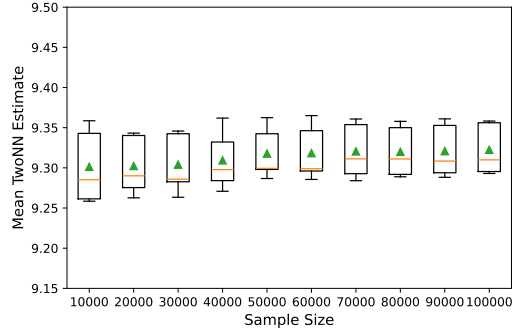
D.1.2. Noise Analysis

The following experiments investigate the stability of the TwoNN dimension estimates under small perturbations of point clouds in a realistic setting of high-dimensional language model embedding spaces. The original work in [67] studies the influence of artificial noise on the global TwoNN estimator for specific toy datasets (points sampled from a 2-dimensional plane in 20-dimensional space, and points sampled

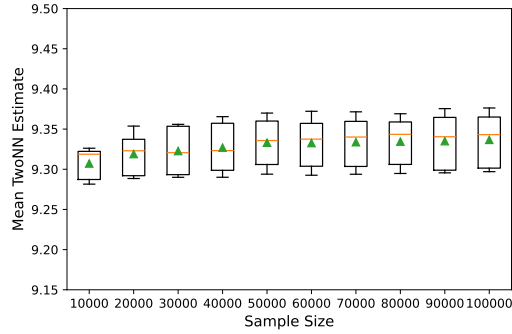
D.1. Sensitivity Analysis



(a) MultiWOZ training data.



(b) MultiWOZ validation data.



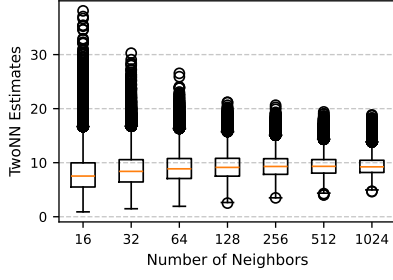
(c) MultiWOZ test data.

Figure D.3.: Distribution of the mean local TwoNN estimates for different cardinalities of token sub-samples. We compute the mean estimates for 5 different token sub-sampling seeds and show their distribution for each fixed subsampling cardinality.

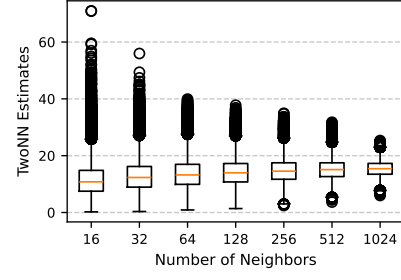
from a 2-dimensional Gaussian distribution wrapped around a Swiss roll). Here, we confirm the stability under noise for our mean local estimates and in the realistic setting of high-dimensional contextual language model embeddings.

We analyze the effect of Gaussian noise on the embeddings from the MultiWOZ test set using a fixed set of parameters and varying noise distortions. Specifically, we use $M = 10\,000$ random sequences from the MultiWOZ test set and extract the masked token embeddings from the last layer of the RoBERTa base model, which

D. Appendix

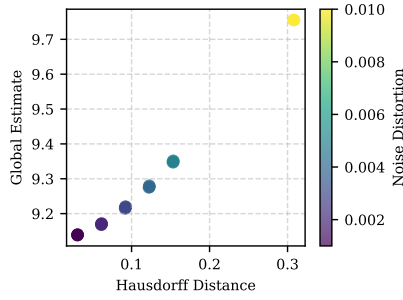


(a) MultiWOZ validation data.

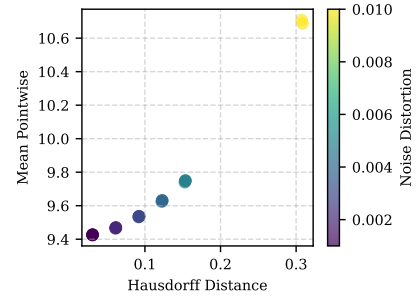


(b) Reddit validation data.

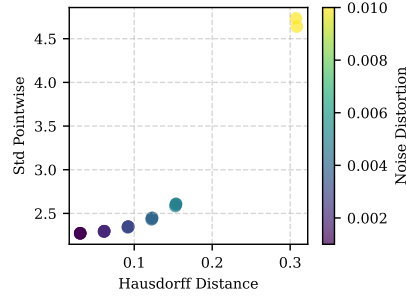
Figure D.4.: Distribution of the local token-wise TwoNN estimates over different locality scales L (number of neighbors). Here shown are the mean and quartiles of the resulting measure varying over the different tokens in a fixed subsample. The more global the estimates, the smaller the standard deviation in the resulting distribution.



(a) Global TwoNN estimate.



(b) Mean of local TwoNN estimates.



(c) Standard deviation of local TwoNN estimates.

Figure D.5.: Effect of artificial Gaussian noise on masked token embeddings. On the x -axis we mark the Hausdorff distance of the noisy embeddings to the clean embeddings, with color denoting the distortion σ in the artificial Gaussian noise.

produces 768-dimensional vectors. To compute local estimates, we sub-sample $N = 60\,000$ points from the embeddings and set the neighborhood size to $L = 128$.

Gaussian noise is applied to all dimensions of the embeddings with varying distortion

D.2. Additional details on datasets used

parameters, $\sigma \in \{0.001; 0.002; 0.003; 0.004; 0.01\}$, and multiple random seeds to introduce variability. For each noise level, we compute the approximate Hausdorff distance between the point cloud of the 60 000 clean embedding vectors and the 60 000 noisy embedding vectors, using the Euclidean distance as the distance metric.

The results are visualized in Figure D.5, where we plot the Hausdorff distance against three measures: the global TwoNN-estimate of the noisy point cloud, the mean of the local TwoNN estimates, and the standard deviation of the local TwoNN estimates.

From the plots, we observe the following trends. The Hausdorff distances between the clean and noisy embeddings and the global and the mean local TwoNN estimates increase with larger noise levels, and there is a clear correlation between the two. Notably, for a distortion parameter of $\sigma = 0.01$, the difference in the dimension estimates exceeds one. As expected, the standard deviation of the local TwoNN estimates also increases with higher noise levels, reflecting the increased variability introduced by the noise. Nevertheless, under small noise levels, the mean local estimate remains relatively stable and has lower variability than the global TwoNN estimate.

D.1.3. Masked Token Embeddings versus Regular Token Embeddings

The plots in Figure D.6 compare the TwoNN estimates for regular tokens (left) and masked tokens (right) in the `MultiWOZ` dataset at the last layer. Each box-plot represents the distribution of token-wise TwoNN estimates.

The estimates for regular tokens have an interquartile range from approximately 5.5 to 8, with a median near 7. This distribution closely resembles that of masked tokens, which exhibit a similar range and median. However, the broader spread for regular tokens suggests distinct token-level geometric behaviors influenced by their role in the learned representation space. In any case, this justifies the computation of the mean local estimates for the regular token embeddings, which are cheaper to obtain: Every sequence token must be masked individually to retrieve masked embeddings with a corresponding model forward pass. In contrast, for the regular embeddings, a single forward pass of the entire sequence automatically produces embeddings for every input token.

D.2. Additional details on datasets used

See Table D.1 for details on the dataset cards and corresponding split sizes.

D. Appendix

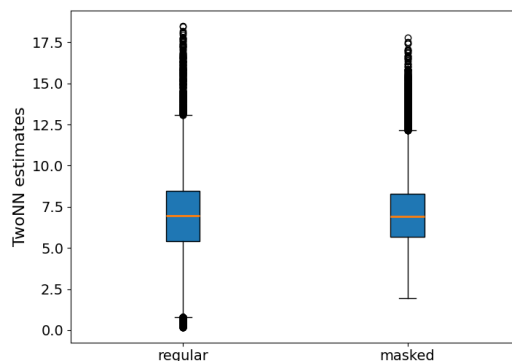


Figure D.6.: TwoNN estimates for regular and masked tokens in the MultiWOZ dataset at the last layer. Both token types show similar distributions.

Table D.1.: Dataset sizes for various datasets used in the experiments. Some datasets use pre-determined splits, while others involve random splits.

Dataset	Training Size	Validation Size	Test Size
MultiWOZ2.1 https://huggingface.co/datasets/ConvLab/multiwoz21	113 556	14 748	14 744
EmoWOZ https://paperswithcode.com/dataset/emowoz-1	66 474	8 509	8 634
SGD https://huggingface.co/datasets/google-research-datasets/schema_guided_dstc8	329 964	48 726	84 594
Wikipedia (random 80%-10%-10% split) https://huggingface.co/datasets/Salesforce/wikitext/viewer/wikitext-103-v1	1 441 080	180 135	180 135
Reddit (random 80%-10%-10% split) https://huggingface.co/datasets/SocialGrep/one-year-of-tsla-on-reddit	174 996	21 874	21 874

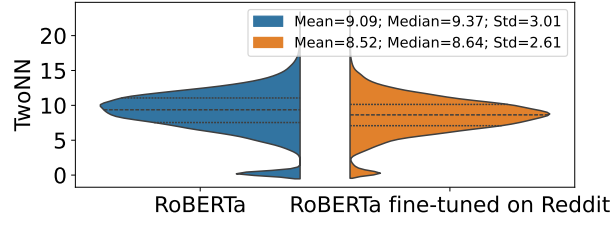
D.3. Additional Experimental Results

D.3.1. Additional fine-tuning results

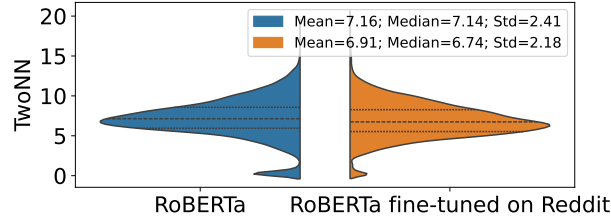
Fine-tuning on unseen data Continuing the discussion in Section 7.4.1, in Figures D.7 to D.9, we provide additional distribution plots of the local estimates. Figure D.7 is interesting because here the RoBERTa model is fine-tuned on data that it has not seen during pre-training.

Fine-tuning an autoregressive language model Moreover, in Figures D.8 and D.9 we show the distribution of local estimates for the auto-regressive GPT-2-medium model in a natural language modeling task, and their shift under fine-tuning. The GPT-2-medium fine-tuning here is performed with standard parameters on the first 10 000 sequences of the given dataset’s train portion, and we show the local estimates at the checkpoint after 1200 batches.

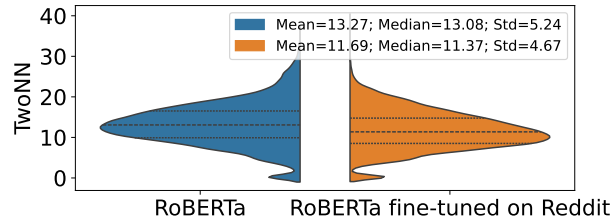
D.3. Additional Experimental Results



(a) TwoNN estimates on MultiWOZ validation embeddings.



(b) TwoNN estimates on SGD validation embeddings.



(c) TwoNN estimates on Reddit validation embeddings.

Figure D.7.: Comparison of the intrinsic dimension (ID) across three data modalities for both models. The ID of embeddings originating from the fine-tuning distribution (Reddit) differs markedly between models, whereas the IDs for the out-of-distribution corpora (SGD, MultiWOZ) are almost indistinguishable.

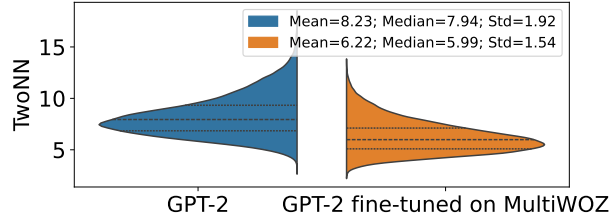
D.3.2. Layerwise computation of local estimates

Our setup from the methods' section and 3 naturally applies to embeddings derived from arbitrary layers of the language model. In particular, the mean local estimates can be compared layer-wise between different checkpoints of a model, as we show in Figure D.10. The main observations here are that the drop in mean local dimension on the dataset used for fine-tuning is visible over all the model layers. It appears to be most pronounced in the intermediate and last layers of the model.

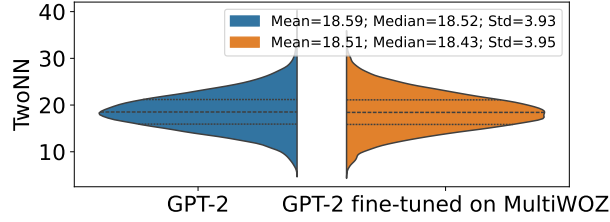
D.3.3. Additional Grokking results

This section supplements the grokking experiments in Section 7.4.2. See Figure D.11 for a plot comparing model performance and mean local estimates on the training

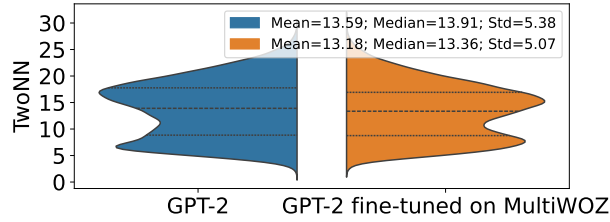
D. Appendix



(a) TwoNN estimates on MultiWOZ validation embeddings.



(b) TwoNN estimates on Wikipedia validation embeddings.



(c) TwoNN estimates on Reddit validation embeddings.

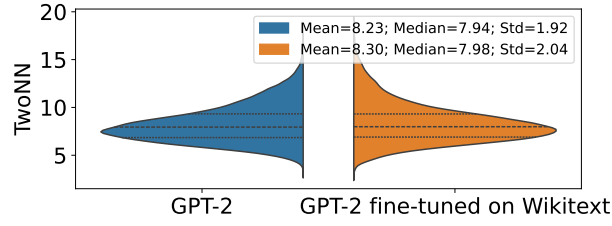
Figure D.8.: Comparison of the intrinsic dimension (ID) across three data modalities for both models. The ID of embeddings originating from the fine-tuning distribution (MultiWOZ) differs markedly between models, whereas the IDs for the out-of-distribution corpora (Wikipedia, Reddit) are almost indistinguishable.

set, shown separately for the training run seeds. This representation is useful, as the timing of the onset of grokking can be very dependent on the model and dataset shuffling seed. It remains true that for individual seeds, the drop in mean local dimension on the training set coincides with the increase in validation accuracy.

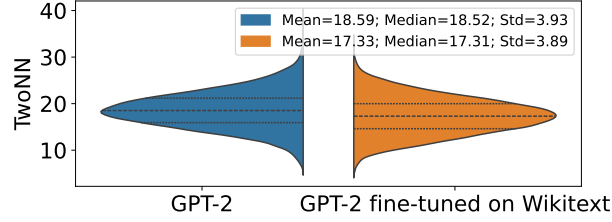
D.3.4. Additional Trippy-R dialogue state tracking results

We here show additional results for Section 7.4.3, where we discuss exhausting model training capabilities. See Figure D.12c for longer TripPy-R dialog state tracking runs over 50 epochs. After about 20 training epochs, the mean local estimates stabilize and predict the convergence of the model performance on the downstream task.

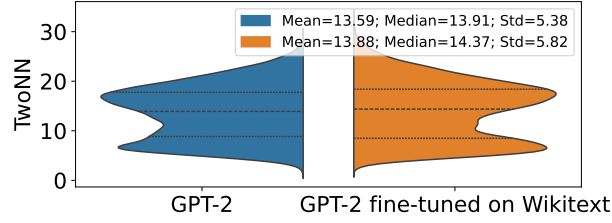
D.3. Additional Experimental Results



(a) TwoNN estimates on MultiWOZ validation embeddings.



(b) TwoNN estimates on Wikipedia validation embeddings.



(c) TwoNN estimates on Reddit validation embeddings.

Figure D.9.: Comparison of the intrinsic dimension (ID) across three data modalities for both models. The ID of embeddings originating from the fine-tuning distribution (Wikipedia) is markedly lower, whereas the IDs for the out-of-distribution corpora (MultiWOZ, Reddit) are even higher for the fine-tuned variants.

D. Appendix

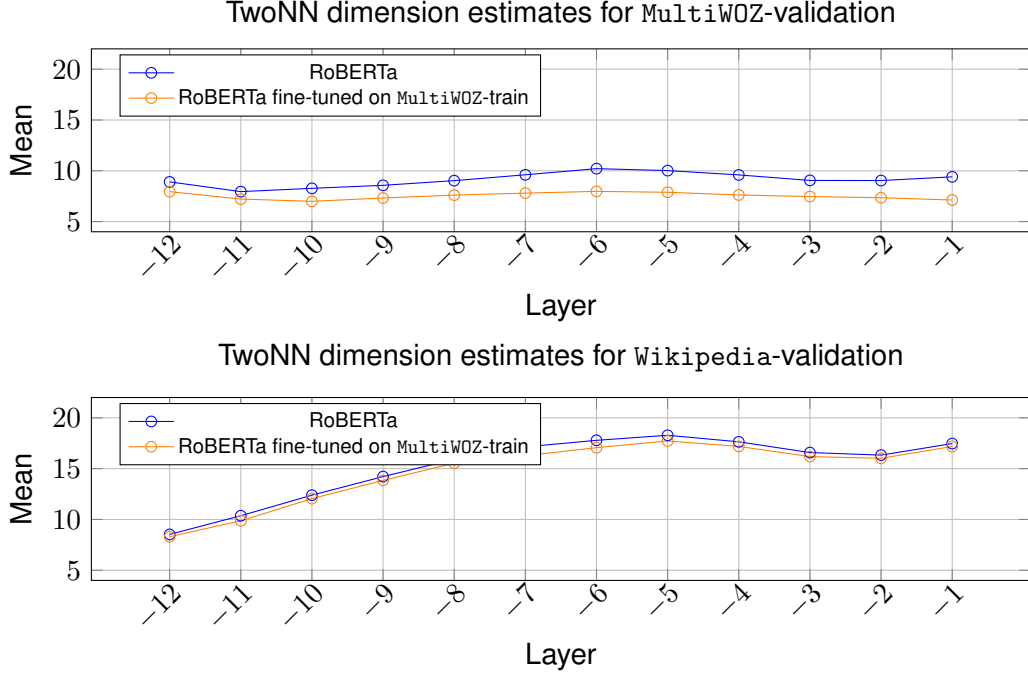


Figure D.10.: Comparison of the development of the ID through the different layers of the two models. The ID of embeddings from the distribution that has been used for fine-tuning differs significantly between the two models, whereas the ID of other data distributions are indistinguishable.

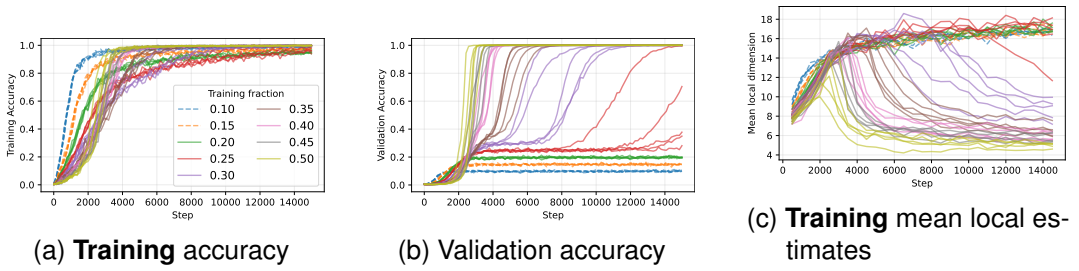


Figure D.11.: Training a model on addition mod $p = 197$ with different training data fraction selected from $\{0.1; 0.15; 0.2; 0.25; 0.3; 0.4; 0.5\}$. Five seeds per configuration are shown separately, demonstrating that the onset of grokking and the drop in mean local dimension vary greatly depending on the specific training run. The plots show the development over 15 000 batches. The mean local estimates are computed on the **training** split for $N = 3000$; $L = 64$.

D.3. Additional Experimental Results

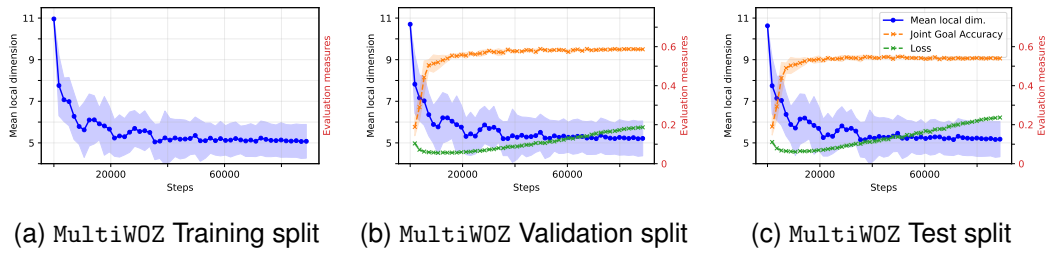


Figure D.12.: Longer Trippy-R training runs over 50 epochs with a linear learning rate schedule. Shown are the mean and standard deviation of the measures over 4 seeds of the training runs.

Bibliography

- [1] Henry Adams et al. “Persistence Images: A Stable Vector Representation of Persistent Homology”. In: *Journal of Machine Learning Research* 18.8 (2017), pp. 1–35.
- [2] Henry Adams et al. “Persistence images: A stable vector representation of persistent homology”. In: *Journal of Machine Learning Research* 18.8 (2017), pp. 1–35.
- [3] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. “Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 7319–7328. DOI: [10.18653/v1/2021.acl-long.568](https://doi.org/10.18653/v1/2021.acl-long.568). URL: <https://aclanthology.org/2021.acl-long.568/>.
- [4] Hirokazu Anai et al. “DTM-Based Filtrations”. In: *Topological Data Analysis*. Ed. by N. A. Baas et al. 2020, pp. 33–66.
- [5] Christophe Andrieu and Gareth O. Roberts. “The pseudo-marginal approach for efficient Monte Carlo computations”. In: *The Annals of Statistics* 37.2 (2009), pp. 697–725. DOI: [10.1214/07-AOS574](https://doi.org/10.1214/07-AOS574).
- [6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).
- [7] Jonathan Bac et al. “Scikit-Dimension: A Python Package for Intrinsic Dimension Estimation”. In: *Entropy* 23.10 (2021). ISSN: 1099-4300. DOI: [10.3390/e23101368](https://doi.org/10.3390/e23101368). URL: <https://www.mdpi.com/1099-4300/23/10/1368>.
- [8] Markus Banagl. *Topological invariants of stratified spaces*. Springer Science & Business Media, 2007.
- [9] Ulrich Bauer. “Ripser: efficient computation of Vietoris–Rips persistence barcodes”. In: *Journal of Applied and Computational Topology* 5.3 (2021), pp. 391–423.
- [10] Mark A Beaumont, Wenyang Zhang, and David J Balding. “Approximate Bayesian computation in population genetics”. In: *Genetics* 162.4 (2002), pp. 2025–2035.
- [11] József Beck. “Irregularities of distribution. I”. In: *Acta Mathematica* 159 (1987), pp. 1–49. DOI: [10.1007/BF02392553](https://doi.org/10.1007/BF02392553).

Bibliography

- [12] Mikhail Belkin and Partha Niyogi. “Laplacian eigenmaps for dimensionality reduction and data representation”. In: *Neural computation* 15.6 (2003), pp. 1373–1396.
- [13] Paul Bendich. “Analyzing stratified spaces using persistent versions of intersection and local homology”. PhD thesis. Duke University, 2008.
- [14] Paul Bendich and John Harer. “Persistent Intersection Homology”. In: *Foundations of Computational Mathematics* 11.3 (2011), pp. 305–336.
- [15] Paul Bendich et al. “Inferring Local Homology from Sampled Stratified Spaces”. In: *IEEE Symposium on Foundations of Computer Science (FOCS)*. 2007, pp. 536–546.
- [16] Paul Bendich et al. “Persistent homology analysis of brain artery trees”. In: *Annals of Applied Statistics* 10.1 (2016), pp. 198–218.
- [17] Guillermo Bernárdez et al. “ICML Topological Deep Learning Challenge 2024: Beyond the Graph Domain”. In: *arXiv preprint arXiv:2409.05211* (2024).
- [18] Espen Bernton et al. “Approximate Bayesian computation with the Wasserstein distance”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.2 (2019), pp. 235–269.
- [19] Dhananjay Bhaskar et al. “Diffusion Curvature for Estimating Local Curvature in High Dimensional Data”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: <https://openreview.net/forum?id=AYII8AkvD1e>.
- [20] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. “A general framework for updating belief distributions”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5 (2016), p. 1103.
- [21] Cristian Bodnar et al. “Weisfeiler and lehman go topological: Message passing simplicial networks”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1026–1037.
- [22] Antonio Briola et al. *Homological Convolutional Neural Networks*. 2023. arXiv: 2308.13816 [cs.LG]. URL: <https://arxiv.org/abs/2308.13816>.
- [23] Bradley C.A. Brown et al. “Verifying the Union of Manifolds Hypothesis for Image Data”. In: *International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=Rvee9CAX4fi>.
- [24] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [25] Peter Bubenik et al. “Statistical topological data analysis using persistence landscapes.” In: *J. Mach. Learn. Res.* 16.1 (2015), pp. 77–102.
- [26] Peter Bubenik et al. “Persistent homology detects curvature”. In: *Inverse Problems* 36.2 (2020), p. 025008.
- [27] Xingyu Cai et al. “Isotropy in the Contextual Embedding Space: Clusters and Manifolds”. In: *International Conference on Learning Representations (ICLR)*. 2021. URL: <https://openreview.net/forum?id=xYGN0860WDH>.

- [28] Francesco Camastra and Antonino Staiano. “Intrinsic dimension estimation: Advances and open problems”. In: *Information Sciences* 328 (2016), pp. 26–41.
- [29] Gunnar Carlsson. “Topological pattern recognition for point cloud data”. In: *Acta Numerica* 23 (2014), pp. 289–368.
- [30] Gunnar Carlsson. “Topology and data”. In: *Bulletin of the American Mathematical Society* 46.2 (2009), pp. 255–308.
- [31] Gunnar Carlsson et al. “On the local behavior of spaces of natural images”. In: *International journal of computer vision* 76 (2008), pp. 1–12.
- [32] Mathieu Carrière et al. “PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures”. In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2786–2796.
- [33] Mathieu Carrière et al. “Perslay: A neural network layer for persistence diagrams and new graph topological signatures”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2786–2796.
- [34] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.
- [35] Tyler Chang, Zhuowen Tu, and Benjamin Bergen. “The Geometry of Multilingual Language Model Representations”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 119–136. DOI: [10.18653/v1/2022.emnlp-main.9](https://doi.org/10.18653/v1/2022.emnlp-main.9). URL: <https://aclanthology.org/2022.emnlp-main.9>.
- [36] Frédéric Chazal, Vin De Silva, and Steve Oudot. “Persistence stability for geometric complexes”. In: *Geometriae Dedicata* 173.1 (2014), pp. 193–214.
- [37] Frédéric Chazal et al. “Persistence-Based Clustering in Riemannian Manifolds”. In: *Journal of the ACM* 60.6 (2013).
- [38] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PmlR. 2020, pp. 1597–1607.
- [39] Yuzhou Chen, Jose Frias, and Yulia R Gel. “Topogcl: Topological graph contrastive learning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. 10. 2024, pp. 11453–11461.
- [40] Zhengdao Chen et al. “Can graph neural networks count substructures?” In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10383–10395.
- [41] Badr-Eddine Chérif-Abdellatif and Pierre Alquier. “MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy”. In: *Symposium on Advances in Approximate Bayesian Inference*. PMLR. 2020, pp. 1–21.

Bibliography

- [42] James R Clough et al. “A topological loss function for deep-learning based image segmentation using persistent homology”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.12 (2020), pp. 8766–8778.
- [43] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of persistence diagrams”. In: *Proceedings of the twenty-first annual symposium on Computational geometry*. 2005, pp. 263–271.
- [44] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of persistence diagrams”. In: *Proceedings of the twenty-first annual symposium on Computational geometry*. 2005, pp. 263–271.
- [45] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of persistence diagrams”. In: *Discrete & Computational Geometry* 37.1 (Jan. 2007), pp. 103–120. DOI: [10.1007/s00454-006-1276-5](https://doi.org/10.1007/s00454-006-1276-5).
- [46] Richard T Cox. “Probability, frequency and reasonable expectation”. In: *American journal of physics* 14.1 (1946), pp. 1–13.
- [47] Imre Csiszár. “I-divergence geometry of probability distributions and minimization problems”. In: *The annals of probability* (1975), pp. 146–158.
- [48] Justin Curry, Sayan Mukherjee, and Katharine Turner. “How many directions determine a shape and other sufficiency results for two topological transforms”. In: *Transactions of the American Mathematical Society, Series B* 9.32 (2022), pp. 1006–1043.
- [49] Justin Michael Curry. *Sheaves, cosheaves and applications*. University of Pennsylvania, 2014.
- [50] Bruno De Finetti, Henry E Kyburg, and Howard E Smokler. “Foresight: Its logical laws, its subjective sources”. In: *Breakthroughs in statistics* 1 (1937), pp. 134–174.
- [51] Francesco Denti et al. “Distributional results for model-based intrinsic dimension estimators”. In: *arXiv preprint arXiv:2104.13832* (2021).
- [52] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Tamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423/>.
- [53] Tamal K. Dey, Fengtao Fan, and Yusu Wang. “Dimension detection with local homology”. In: *Canadian Conference on Computational Geometry (CCCG)*. 2014.
- [54] Francesco Di Giovanni et al. “On over-squashing in message passing neural networks: The impact of width, depth, and topology”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 7865–7885.
- [55] Abdessalam Ed-dib, Zhanibek Datbayev, and Amine Mohamed Aboussalah. “GeLoRA: Geometric Adaptive Ranks For Efficient LoRA Fine-tuning”. In: *arXiv preprint arXiv:2412.09250* (2024). arXiv: [2412.09250](https://arxiv.org/abs/2412.09250) [cs.LG]. URL: <https://arxiv.org/abs/2412.09250>.

- [56] Monroe D Donsker and SR Srinivasa Varadhan. “Asymptotic evaluation of certain Markov process expectations for large time, I”. In: *Communications on Pure and Applied Mathematics* 28.1 (1975), pp. 1–47.
- [57] Arnaud Doucet, Adam M Johansen, et al. “A tutorial on particle filtering and smoothing: Fifteen years later”. In: *Handbook of nonlinear filtering* 12.656-704 (2009), p. 3.
- [58] Nadir Durrani et al. “On the Transformation of Latent Space in Fine-Tuned NLP Models”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1495–1516. DOI: [10.18653/v1/2022.emnlp-main.97](https://doi.org/10.18653/v1/2022.emnlp-main.97). URL: <https://aclanthology.org/2022.emnlp-main.97>.
- [59] Stefania Ebli, Michaël Defferrard, and Gard Spreemann. “Simplicial neural networks”. In: *arXiv preprint arXiv:2010.03633* (2020).
- [60] Herbert Edelsbrunner and John Harer. *Computational topology: An introduction*. Providence, RI, USA: American Mathematical Society, 2010.
- [61] Herbert Edelsbrunner, John Harer, et al. “Persistent homology-a survey”. In: *Contemporary mathematics* 453.26 (2008), pp. 257–282.
- [62] Herbert Edelsbrunner, David Letscher, and Afra J. Zomorodian. “Topological persistence and simplification”. In: *Discrete & Computational Geometry* 28.4 (Nov. 2002), pp. 511–533. DOI: [10.1007/s00454-002-2885-2](https://doi.org/10.1007/s00454-002-2885-2).
- [63] Edelsbrunner, Letscher, and Zomorodian. “Topological persistence and simplification”. In: *Discrete & computational geometry* 28 (2002), pp. 511–533.
- [64] Mihail Eric et al. “MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines”. eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 422–428. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.53/>.
- [65] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.
- [66] Kawin Ethayarajh. “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65. DOI: [10.18653/v1/D19-1006](https://doi.org/10.18653/v1/D19-1006). URL: <https://aclanthology.org/D19-1006>.
- [67] Elena Facco et al. “Estimating the intrinsic dimension of datasets by a minimal neighborhood information”. In: *Scientific reports* 7.1 (2017), p. 12140.

Bibliography

- [68] Wenqi Fan et al. “A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models”. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD ’24. Barcelona, Spain: Association for Computing Machinery, 2024, pp. 6491–6501. ISBN: 9798400704901. DOI: [10.1145/3637528.3671470](https://doi.org/10.1145/3637528.3671470). URL: <https://doi.org/10.1145/3637528.3671470>.
- [69] Brittany Terese Fasy and Bei Wang. “Exploring persistent local homology in topological data analysis”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 6430–6434.
- [70] Shutong Feng et al. “EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 4096–4113. URL: <https://aclanthology.org/2022.lrec-1.436/>.
- [71] Shutong Feng et al. “From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue”. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Svetlana Stoyanchev et al. Prague, Czechia: Association for Computational Linguistics, Sept. 2023, pp. 85–103. DOI: [10.18653/v1/2023.sigdial-1.8](https://doi.org/10.18653/v1/2023.sigdial-1.8). URL: <https://aclanthology.org/2023.sigdial-1.8/>.
- [72] Javier Ferrando et al. “A Primer on the Inner Workings of Transformer-based Language Models”. In: *arXiv preprint arXiv:2405.00208* (2024). arXiv: [2405.00208](https://arxiv.org/abs/2405.00208) [cs.CL]. URL: <https://arxiv.org/abs/2405.00208>.
- [73] Rémi Flamary et al. “POT: Python Optimal Transport”. In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.
- [74] David T Frazier, Christian Robert, and Judith Rousseau. “Model Misspecification in ABC: Consequences and Diagnostics”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2019).
- [75] David T Frazier et al. “Asymptotic properties of approximate Bayesian computation”. In: *Biometrika* 105.3 (2018), pp. 593–607.
- [76] Ulderico Fugacci et al. “Persistent Homology: a Step-by-step Introduction for Newcomers.” In: *STAG*. 2016, pp. 1–10.
- [77] Robert Ghrist. “Barcodes: the persistent topology of data”. In: *Bulletin of the American Mathematical Society* 45.1 (2008), pp. 61–75.
- [78] Robert Ghrist, Rachel Levanger, and Huy Mai. “Persistent homology and Euler integral transforms”. In: *Journal of Applied and Computational Topology* 2 (2018), pp. 55–60.
- [79] Saumya Gupta, Dimitris Samaras, and Chao Chen. “Topodiffusionnet: A topology-aware diffusion model”. In: *arXiv preprint arXiv:2410.16646* (2024).
- [80] Will Hamilton, Zhitao Ying, and Jure Leskovec. “Inductive representation learning on large graphs”. In: *Advances in Neural Information Processing Systems* 30 (2017).

- [81] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [82] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.
- [83] Michael Heck et al. “Robust Dialogue State Tracking with Weak Supervision and Sparse Data”. In: *Transactions of the Association for Computational Linguistics* 10 (2022). Ed. by Brian Roark and Ani Nenkova, pp. 1175–1192. DOI: [10.1162/tac1_a_00513](https://doi.org/10.1162/tac1_a_00513). URL: <https://aclanthology.org/2022.tac1-1.68/>.
- [84] Felix Hensel, Michael Moor, and Bastian Rieck. “A Survey of Topological Machine Learning Methods”. In: *Frontiers in Artificial Intelligence* 4 (2021).
- [85] Felix Hensel, Michael Moor, and Bastian Rieck. “A survey of topological machine learning methods”. In: *Frontiers in Artificial Intelligence* 4 (2021), p. 681108.
- [86] Christoph Hofer et al. “Deep Learning with Topological Signatures”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2017).
- [87] Christoph Hofer et al. “Graph filtration learning”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4314–4323.
- [88] Christoph D Hofer, Roland Kwitt, and Marc Niethammer. “Learning representations of persistence barcodes”. In: *Journal of Machine Learning Research* 20.126 (2019), pp. 1–45.
- [89] Max Horn et al. “Topological Graph Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. 2022. arXiv: [2102.07835](https://arxiv.org/abs/2102.07835) [cs.LG].
- [90] Max Horn et al. “Topological Graph Neural Networks”. In: *International Conference on Learning Representations*. 2022. arXiv: [2102.07835](https://arxiv.org/abs/2102.07835) [cs.LG]. URL: <https://openreview.net/forum?id=oxxUMeFwEHd>.
- [91] Edward J Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [92] Jiangbei Hu et al. “Topology-aware latent diffusion for 3d shape generation”. In: *arXiv preprint arXiv:2401.17603* (2024).
- [93] Xiaoling Hu et al. “Topology-Preserving Deep Image Segmentation”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.
- [94] Alexander Jakubowski, Milica Gašić, and Marcus Zibrowius. “Topology of Word Embeddings: Singularities Reflect Polysemy”. In: *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*. 2020, pp. 103–113.

Bibliography

- [95] Uyeong Jang, Susmit Jha, and Somesh Jha. “On the Need for Topology-Aware Generative Models for Manifold-Based Defenses”. In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=r1lF_CeYwS.
- [96] Albert Q Jiang et al. “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825* (2023).
- [97] Jeff Johnson, Matthijs Douze, and Hervé Jégou. “Billion-scale similarity search with GPUs”. In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.
- [98] Chaitanya K Joshi et al. “On the Expressive Power of Geometric Graph Neural Networks. 2023”. In: *International Conference on Learning Representations*. 2023.
- [99] Tikeng Notsawo Pascal Junior et al. “Predicting Grokking Long Before it Happens: A look into the loss landscape of models which grok”. In: *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*. 2024. URL: <https://openreview.net/forum?id=AUjCSjxhGj>.
- [100] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 12, 2025. Online manuscript, 2025. URL: <https://web.stanford.edu/~jurafsky/slp3/>.
- [101] Diederik P. Kingma and Max Welling. “An Introduction to Variational Autoencoders”. In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392.
- [102] Thomas N Kipf and Max Welling. “Semi-supervised classification with graph convolutional networks”. In: *arXiv preprint arXiv:1609.02907* (2016).
- [103] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. “An Optimization-centric View on Bayes’ Rule: Reviewing and Generalizing Variational Inference”. In: *Journal of Machine Learning Research* 23.132 (2022), pp. 1–109.
- [104] Jan J Koenderink and Andrea J van Doorn. “Dynamic shape”. In: *Biological Cybernetics* 53.6 (1986), pp. 383–396.
- [105] Timm Krüger et al. “The lattice Boltzmann method”. In: *Springer International Publishing* 10.978-3 (2017), pp. 4–15.
- [106] Laida Kushnareva et al. “Artificial Text Detection via Examining the Topology of Attention Maps”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 635–649. DOI: [10.18653/v1/2021.emnlp-main.50](https://doi.org/10.18653/v1/2021.emnlp-main.50). URL: <https://aclanthology.org/2021.emnlp-main.50>.
- [107] Roland Kwitt et al. “Statistical Topological Data Analysis - A Kernel Perspective”. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 3070–3078.

- [108] Michael Lesnick. “The Theory of the Interleaving Distance on Multidimensional Persistence Modules”. In: *Foundations of Computational Mathematics* 15.3 (2015), pp. 613–650.
- [109] Tony Lindeberg. “Scale-space theory: A basic tool for analyzing structures at different scales”. In: *Journal of Applied Statistics* 21.1-2 (1994), pp. 225–270.
- [110] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint arXiv:1907.11692* (2019).
- [111] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).
- [112] Jan-Matthis Lueckmann et al. “Benchmarking simulation-based inference”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 343–351.
- [113] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.
- [114] Vasileios Maroulas, Cassie Putman Micucci, and Farzana Nasrin. “Bayesian Topological Learning for Classifying the Structure of Biological Networks”. In: *Bayesian Analysis* 17.3 (2022), pp. 711–736. DOI: [10.1214/21-BA1270](https://doi.org/10.1214/21-BA1270).
- [115] Lewis Marsh et al. “Detecting temporal shape changes with the euler characteristic transform”. In: *Transactions of Mathematics and Its Applications* 8.2 (2024).
- [116] Leland McInnes, John Healy, and James Melville. “UMAP: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv:1802.03426* (2018).
- [117] Amil Merchant et al. “What happens to BERT embeddings during fine-tuning?” In: *arXiv preprint arXiv:2004.14448* (2020).
- [118] Xiaoyue Mi et al. “Topology-preserving Adversarial Training for Alleviating Natural Accuracy Degradation”. In: *arXiv preprint arXiv:2311.17607* (2023).
- [119] Jeffrey W Miller and David B Dunson. “Robust Bayesian inference via coarsening”. In: *Journal of the American Statistical Association* (2018).
- [120] Kevin R. Moon et al. “Visualizing structure and transitions in high-dimensional biological data”. In: *Nature Biotechnology* 37.12 (2019), pp. 1482–1492.
- [121] Michael Moor et al. “Challenging Euclidean Topological Autoencoders”. In: *‘Topological Data Analysis and Beyond’ Workshop at NeurIPS*. 2020. URL: <https://openreview.net/forum?id=P3dZu0UnyEY>.
- [122] Michael Moor et al. “Topological Autoencoders”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Ed. by Hal Daumé III and Aarti Singh. Proceedings of Machine Learning Research 119. PMLR, 2020, pp. 7045–7054. arXiv: [1906.00722](https://arxiv.org/abs/1906.00722) [cs.LG].
- [123] Michael Moor et al. “Topological autoencoders”. In: *International conference on machine learning*. PMLR. 2020, pp. 7045–7054.
- [124] Michael Moor et al. “Topological autoencoders”. In: *International conference on machine learning*. PMLR. 2020, pp. 7045–7054.

Bibliography

- [125] Christopher Morris et al. “Weisfeiler and Leman go Machine Learning: The Story so far”. In: *Journal of Machine Learning Research* 24.333 (2023), pp. 1–59. arXiv: 2112.09992 [cs.LG].
- [126] Elizabeth Munch et al. “Probabilistic Fréchet means for time varying persistence diagrams”. In: *Electronic Journal of Statistics* 9.1 (2015), pp. 1173–1204.
- [127] Vishal S Ngairangbam et al. “Enhancing anomaly detection with topology-aware autoencoders”. In: *arXiv preprint arXiv:2502.10163* (2025).
- [128] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. “Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival”. In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 7265–7270.
- [129] Christopher A. Oballe. “Bayesian Topological Machine Learning”. PhD thesis. University of Tennessee, Knoxville, 2020.
- [130] OpenAI. *ChatGPT: GPT-4*. <https://chat.openai.com>. Accessed: 2024-08-28. 2024.
- [131] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc., 2022. ISBN: 9781713871088.
- [132] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013. URL: <https://artowen.su.domains/mc/>.
- [133] Theodore Papamarkou et al. “Position: Topological deep learning is the new frontier for relational learning”. In: *arXiv preprint arXiv:2402.08871* (2024).
- [134] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. “K2-ABC: approximate Bayesian computation with kernel embeddings”. In: *arXiv preprint arXiv:1502.02558* (2015).
- [135] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.
- [136] Hongbin Pei et al. “Geom-GCN: Geometric Graph Convolutional Networks”. In: *International Conference on Learning Representations*. 2020. URL: <https://openreview.net/forum?id=S1e2agrFvS>.
- [137] Ilan Perez and Raphael Reinauer. “The topological BERT: Transforming attention into topology for natural language processing”. In: *arXiv preprint arXiv:2206.15195* (2022).
- [138] Giovanni Petri et al. “Homological scaffolds of brain functional networks”. In: *Journal of The Royal Society Interface* 11.101 (2014), p. 20140873.
- [139] Oleg Platonov et al. “A critical look at the evaluation of GNNs under heterophily: Are we really making progress?” In: *arXiv preprint arXiv:2302.11640* (2023).
- [140] Phillip Pope et al. “The Intrinsic Dimension of Images and Its Impact on Learning”. In: *International Conference on Learning Representations*. 2021.

- [141] Alethea Power et al. “Grokking: Generalization beyond overfitting on small algorithmic datasets”. In: *arXiv preprint arXiv:2201.02177* (2022).
- [142] Jonathan K Pritchard et al. “Population growth of human Y chromosomes: a study of Y chromosome microsatellites”. In: *Molecular biology and evolution* 16.12 (1999), pp. 1791–1798.
- [143] Alec Radford et al. “Improving Language Understanding by Generative Pre-Training”. In: *blog* (June 2018). URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [144] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* (2019).
- [145] Abhinav Rastogi et al. “Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34.05. 2020, pp. 8689–8696.
- [146] Nalini Ravishanker and Renjie Chen. “An introduction to persistent homology for time series”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 13.3 (2021), e1548.
- [147] Jan Reininghaus et al. “A stable multi-scale kernel for topological machine learning”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Red Hook, NY, USA: Curran Associates, Inc., June 2015, pp. 4741–4748. DOI: [10.1109/CVPR.2015.7299106](https://doi.org/10.1109/CVPR.2015.7299106).
- [148] Jan Reininghaus et al. “A stable multi-scale kernel for topological machine learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4741–4748.
- [149] Bastian Rieck and Heike Leitte. “Agreement Analysis of Quality Measures for Dimensionality Reduction. Theory, Algorithms, and Applications”. In: *Topological Methods in Data Analysis and Visualization IV*. Ed. by Hamish Carr, Christoph Garth, and Tino Weinkauff. 2017, pp. 103–117.
- [150] Bastian Rieck and Heike Leitte. “Exploring and Comparing Clusterings of Multivariate Data Sets Using Persistent Homology”. In: *Computer Graphics Forum* 35.3 (2016), pp. 81–90.
- [151] Bastian Rieck and Heike Leitte. “Persistent Homology for the Evaluation of Dimensionality Reduction Schemes”. In: *Computer Graphics Forum* 34.3 (2015), pp. 431–440.
- [152] Bastian Rieck, Filip Sadlo, and Heike Leitte. “Topological Machine Learning with Persistence Indicator Functions”. In: *Topological Methods in Data Analysis and Visualization V*. Ed. by Hamish Carr et al. Cham, Switzerland: Springer, 2020, pp. 87–101. DOI: [10.1007/978-3-030-43036-8_6](https://doi.org/10.1007/978-3-030-43036-8_6). arXiv: [1907.13496](https://arxiv.org/abs/1907.13496) [math.AT].
- [153] Bastian Rieck et al. “Persistent Intersection Homology for the Analysis of Discrete Data”. In: *Topological Methods in Data Analysis and Visualization V*. Ed. by H. Carr et al. 2020, pp. 37–51. eprint: [1907.13485](https://arxiv.org/abs/1907.13485).

Bibliography

- [154] Bastian Rieck et al. “Uncovering the Topology of Time-Varying fMRI Data using Cubical Persistence”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 6900–6912.
- [155] Ernst Röell and Bastian Rieck. “Differentiable Euler Characteristic Transforms for Shape Classification”. In: *International Conference on Learning Representations*. 2024. arXiv: 2310.07630 [cs.LG]. URL: <https://openreview.net/forum?id=M0632iPq3I>.
- [156] Julius von Rohrscheidt and Bastian Rieck. *Diss-I-ECT: Dissecting Graph Data with local Euler Characteristic Transforms*. 2024. arXiv: 2410.02622 [cs.LG]. URL: <https://arxiv.org/abs/2410.02622>.
- [157] Julius von Rohrscheidt and Bastian Rieck. “Diss-I-ECT: Dissecting Graph Data with local Euler Characteristic Transforms”. In: *CoRR* abs/2410.02622 (2024). DOI: 10.48550/ARXIV.2410.02622. arXiv: 2410.02622. URL: <https://doi.org/10.48550/arXiv.2410.02622>.
- [158] Julius von Rohrscheidt, Bastian Rieck, and Sebastian M Schmon. “Bayesian Computation Meets Topology”. In: *Transactions on Machine Learning Research* (2024).
- [159] Sam T Roweis and Lawrence K Saul. “Nonlinear dimensionality reduction by locally linear embedding”. In: *science* 290.5500 (2000), pp. 2323–2326.
- [160] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. “Multi-scale attributed node embedding”. In: *Journal of Complex Networks* 9.2 (2021), cnab014.
- [161] Benjamin Matthias Ruppik et al. *Less is More: Local Intrinsic Dimensions of Contextual Language Models*. 2025. arXiv: 2506.01034 [cs.CL]. URL: <https://arxiv.org/abs/2506.01034>.
- [162] Benjamin Matthias Ruppik et al. “Local Topology Measures of Contextual Language Model Latent Spaces with Applications to Dialogue Term Extraction”. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara et al. Kyoto, Japan: Association for Computational Linguistics, Sept. 2024, pp. 344–356. DOI: 10.18653/v1/2024.sigdial-1.31. URL: <https://aclanthology.org/2024.sigdial-1.31/>.
- [163] T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. “A survey on oversmoothing in graph neural networks”. In: *arXiv preprint arXiv:2303.10993* (2023).
- [164] Julian Salazar et al. “Masked Language Model Scoring”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 2699–2712. DOI: 10.18653/v1/2020.acl-main.240. URL: <https://aclanthology.org/2020.acl-main.240>.
- [165] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. “Semi-parametric topological memory for navigation”. In: *arXiv preprint arXiv:1803.00653* (2018).

- [166] Sebastian M Schmon, Patrick W Cannon, and Jeremias Knoblauch. “Generalized Posteriors in Approximate Bayesian Computation”. In: *Third Symposium on Advances in Approximate Bayesian Inference*. 2021. URL: <https://openreview.net/forum?id=tKrg5DAyeWq>.
- [167] Luis Scoccola and Jose A. Perea. “FibeRed: Fiberwise Dimensionality Reduction of Topologically Complex Data with Vector Bundles”. In: *International Symposium on Computational Geometry (SoCG)*. Ed. by Erin W. Chambers and Joachim Gudmundsson. Vol. 258. 2023, 56:1–56:18.
- [168] Oleksandr Shchur et al. “Pitfalls of graph neural network evaluation”. In: *arXiv preprint arXiv:1811.05868* (2018).
- [169] Donald R. Sheehy. “Linear-Size Approximations to the Vietoris–Rips Filtration”. In: *Discrete & Computational Geometry* 49.4 (2013), pp. 778–796.
- [170] Reece Shuttleworth et al. “LoRA vs Full Fine-tuning: An Illusion of Equivalence”. In: *arXiv preprint arXiv:2410.21228* (2024).
- [171] Vin de Silva and Gunnar Carlsson. “Topological estimation using witness complexes”. In: *Symposium on Point-Based Graphics*. Ed. by M. Gross et al. The Eurographics Association, 2004.
- [172] Primož Skraba and Katharine Turner. *Wasserstein Stability for Persistence Diagrams*. 2022. arXiv: [2006.16824](https://arxiv.org/abs/2006.16824) [math.AT].
- [173] Bernadette J. Stolz et al. “Geometric anomaly detection in data”. In: *Proceedings of the National Academy of Sciences* 117.33 (2020), pp. 19664–19669.
- [174] Guillaume Tauzin et al. *giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration*. 2020. arXiv: [2004.02551](https://arxiv.org/abs/2004.02551) [cs.LG].
- [175] Guillaume Tauzin et al. “giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration”. In: *arXiv:2004.02551* (2020). arXiv: [2004.02551](https://arxiv.org/abs/2004.02551) [cs.LG].
- [176] Simon Tavaré et al. “Inferring coalescence times from DNA sequence data”. In: *Genetics* 145.2 (1997), pp. 505–518.
- [177] The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015. URL: <http://gudhi.gforge.inria.fr/doc/latest/>.
- [178] Thomas Thorne, Paul D W Kirk, and Heather A. Harrington. “Topological approximate Bayesian computation for parameter inference of an angiogenesis model”. In: *Bioinformatics* 38.9 (2022), pp. 2529–2535. DOI: [10.1093/bioinformatics/btac118](https://doi.org/10.1093/bioinformatics/btac118).
- [179] Chad M Topaz, Lori Ziegelmeier, and Tom Halverson. “Topological data analysis of biological aggregation models”. In: *PloS one* 10.5 (2015), e0126383.
- [180] Jake Topping et al. “Understanding over-squashing and bottlenecks on graphs via curvature”. In: *International Conference on Learning Representations*. 2022.
- [181] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

Bibliography

- [182] John W. Tukey. “Comparing Individual Means in the Analysis of Variance”. In: *Biometrics* 5.2 (1949), pp. 99–114.
- [183] Eduard Tulchinskii et al. “Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [184] Renata Turkeš, Guido Montúfar, and Nina Otter. “On the Effectiveness of Persistent Homology”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: <https://openreview.net/forum?id=DRjUkfExCix>.
- [185] Katharine Turner, Sayan Mukherjee, and Doug M. Boyer. “Persistent homology transform for modeling shapes and surfaces”. In: *Information and Inference: A Journal of the IMA* 3.4 (2014), pp. 310–344. DOI: [10.1093/imaiai/iau011](https://doi.org/10.1093/imaiai/iau011).
- [186] Adaku Uchendu, Thai Le, and Dongwon Lee. “TOPFORMER: Topology-Aware Authorship Attribution of Deepfake Texts with Diverse Writing Styles”. In: *arXiv preprint arXiv:2309.12934* (2023).
- [187] Lucrezia Valeriani et al. “The geometry of hidden representations of large transformer models”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [188] Petar Veličković. “Everything is connected: Graph neural networks”. In: *Current Opinion in Structural Biology* 79 (2023), p. 102538. DOI: [10.1016/j.sbi.2023.102538](https://doi.org/10.1016/j.sbi.2023.102538).
- [189] Petar Veličković et al. “Graph Attention Networks”. In: *International Conference on Learning Representations*. 2018. URL: <https://openreview.net/forum?id=rJXMpikCZ>.
- [190] Yogesh Verma, Amauri H Souza, and Vikas Garg. “Topological Neural Networks go Persistent, Equivariant, and Continuous”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 49388–49407. URL: <https://proceedings.mlr.press/v235/verma24a.html>.
- [191] Tamás Vicsek et al. “Novel type of phase transition in a system of self-driven particles”. In: *Physical Review Letters* 75.6 (1995), p. 1226.
- [192] Karthik Viswanathan et al. “The Geometry of Tokens in Internal Representations of Large Language Models”. In: *arXiv preprint arXiv:2501.10573* (2025). arXiv: [2501.10573](https://arxiv.org/abs/2501.10573) [cs.CL]. URL: <https://arxiv.org/abs/2501.10573>.
- [193] Julius Von Rohrscheidt and Bastian Rieck. “Topological singularity detection at multiple scales”. In: *International Conference on Machine Learning*. PMLR, 2023, pp. 35175–35197.
- [194] Hubert Wagner, Chao Chen, and Erald Vuçini. “Efficient Computation of Persistent Homology for Cubical Data”. In: *Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications*. Ed. by Ronald Peikert et al. Heidelberg, Germany: Springer, 2012, pp. 91–106. DOI: [10.1007/978-3-642-23175-9_7](https://doi.org/10.1007/978-3-642-23175-9_7).

- [195] Dominik J. E. Waibel et al. "Capturing Shape Information with Multi-Scale Topological Loss Terms for 3D Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Ed. by Linwei Wang et al. Cham, Switzerland: Springer, 2022, pp. 150–159. DOI: [10.1007/978-3-031-16440-8_15](https://doi.org/10.1007/978-3-031-16440-8_15). arXiv: [2203.01703](https://arxiv.org/abs/2203.01703) [cs.CV].
- [196] Bei Wang et al. "Branching and Circular Features in High Dimensional Data". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 1902–1911.
- [197] Yuanrong Wang, Antonio Briola, and Tomaso Aste. *Homological Neural Networks: A Sparse Architecture for Multivariate Complexity*. 2023. arXiv: [2306.15337](https://arxiv.org/abs/2306.15337) [cs.LG]. URL: <https://arxiv.org/abs/2306.15337>.
- [198] Keyulu Xu et al. "How powerful are graph neural networks?" In: *arXiv preprint arXiv:1810.00826* (2018).
- [199] Zuoyu Yan et al. "Neural Approximation of Graph Topological Features". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 33357–33370.
- [200] Maosheng Yang, Elvin Isufi, and Geert Leus. "Simplicial convolutional neural networks". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 8847–8851.
- [201] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. "Revisiting semi-supervised learning with graph embeddings". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 40–48.
- [202] Steve Young et al. "The hidden information state model: A practical framework for POMDP-based spoken dialogue management". In: *Computer Speech & Language* 24.2 (2010), pp. 150–174.
- [203] Arnold Zellner. "Optimal information processing and Bayes's theorem". In: *The American Statistician* 42.4 (1988), pp. 278–280.
- [204] Sebastian Zeng et al. "Topological Attention for Time Series Forecasting". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 24871–24882.
- [205] Xu Zhang et al. "A comprehensive review of the oversmoothing in graph neural networks". In: *CCF Conference on Computer Supported Cooperative Work and Social Computing*. Springer. 2023, pp. 451–465.
- [206] Zhenyue Zhang and Hongyuan Zha. "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment". In: *SIAM journal on scientific computing* 26.1 (2004), pp. 313–338.
- [207] Qi Zhao and Yusu Wang. "Learning metrics for persistence-based summaries and applications for graph classification". In: *Advances in neural information processing systems* 32 (2019).
- [208] Qi Zhao and Yusu Wang. "Learning metrics for persistence-based summaries and applications for graph classification". In: *Advances in neural information processing systems* 32 (2019).

Bibliography

- [209] Qi Zhao et al. “Persistence Enhanced Graph Neural Network”. In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2896–2906.
- [210] Qi Zhao et al. “Persistence enhanced graph neural network”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2896–2906.
- [211] Jie Zhou, Daniel L Civco, and John A Silander. “A wavelet transform method to merge Landsat TM and SPOT panchromatic data”. In: *International Journal of Remote Sensing* 19.4 (1998), pp. 743–757.
- [212] Jiong Zhu et al. “Beyond homophily in graph neural networks: Current limitations and effective designs”. In: *Advances in neural information processing systems* 33 (2020), pp. 7793–7804.
- [213] Ali Zia et al. “Topological deep learning: a review of an emerging paradigm”. In: *Artificial Intelligence Review* 57.4 (2024), p. 77.
- [214] Afra Zomorodian and Gunnar Carlsson. “Computing persistent homology”. In: *Discrete & Computational Geometry* 33.2 (2005), pp. 249–274.
- [215] Eli R. Zunder et al. “A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry”. In: *Cell Stem Cell* 16.3 (2015), pp. 323–337.