Minimax Rates for Homology Inference

Sivaraman Balakrishnan[†] **Alessandro Rinaldo**[†] **Don Sheehy**[††]
Aarti Singh[†] **Larry Wasserman**[†]
[†]Carnegie Mellon University [††]INRIA

Abstract

Often, high dimensional data lie close to a low-dimensional submanifold and it is of interest to understand the geometry of these submanifolds. The homology groups of a manifold are important topological invariants that provide an algebraic summary of the manifold. These groups contain rich topological information, for instance, about the connected components, holes, tunnels and sometimes the dimension of the manifold. In this paper, we consider the statistical problem of estimating the homology of a manifold from noisy samples under several different noise models. We derive upper and lower bounds on the minimax risk for this problem. Our upper bounds are based on estimators which are constructed from a union of balls of appropriate radius around carefully selected points. In each case we establish complementary lower bounds using Le Cam's lemma.

1 Introduction

Let M be a d-dimensional manifold embedded in $\mathbb{R}^D$ where $d \leq D$. The *homology groups* $\mathcal{H}(M)$ of M [12] are an algebraic summary of the properties of M. The homology groups of a manifold describe its topological features such as its connected components, holes, tunnels, etc.

In machine learning, there is much focus on clustering. However, the clusters are only the zeroth order homology and hence only scratch the surface of the topological information in a dataset. Extracting information beyond clustering is known as topological data analysis. It is worth emphasizing that the homology groups are topological invariants of a manifold that can be *efficiently* computed [4, 5]. Examples of applications of homology inference have been growing rapidly in the last few years. Homology inference has found application in medical imaging and neuroscience [3, 22], sensor networks [6, 21], landmark-based shape data analyses [11], proteomics [20], microarray analysis [7] and cellular biology [15]. The books by [8, 19, 24] contain various case studies in applications in fields ranging from computational biology to geophysics.

In this paper we study the problem of estimating the homology of a manifold M from a noisy sample $Y_1, \ldots, Y_n$. Specifically, we bound the minimax risk

$$R_n \equiv \inf_{\widehat{\mathcal{H}}} \sup_{Q \in \mathcal{Q}} Q^n \Big(\widehat{\mathcal{H}} \neq \mathcal{H}(M) \Big) \qquad (1)$$

where the infimum is over *all* estimators $\widehat{\mathcal{H}}$ of the homology of M and the supremum is over appropriately defined classes of distributions $\mathcal{Q}$ for Y. Note that $0 \leq R_n \leq 1$ with $R_n = 1$ meaning that the problem is hopeless. Bounding the minimax risk is equivalent to bounding the *sample complexity* of the best possible estimator, defined by $n(\epsilon) = \min\big\{n : \ R_n \leq \epsilon\big\}$ where $0 < \epsilon < 1$.

1.1 Related Work

Other work on statistical homology includes that of Chazal et. al. [2] who show under certain conditions the homology estimate of a manifold from a sample is stable under noise perturbation that is small in a Wasserstein sense. Kahle [14] studies the homology of random geometric graphs and proves many threshold and central limit theorems for their homology. Adler et. al. [1] study the homology induced by the level sets of certain Gaussian random fields. There is also a large literature on manifold denoising that focuses on aspects of the manifold not related to homology; see for instance [13] and references therein.

Our upper bounds mainly generalize those in the work of Niyogi, Smale and Weinberger (henceforth NSW) [17, 18]. They establish a general result showing that when *all* the samples are dense in a thin region sur-

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

rounding the manifold, a union of appropriately sized balls around the samples can be used to construct an accurate estimate of the homology with high probability. Under a variety of different noise models we will show that even when *all* the samples are not close to the manifold it is possible to "clean" the samples (essentially removing those in regions of low-density) and be left with samples which are dense in a thin region around the manifold.

In the case of additive noise with general noise distributions however, we cannot expect too many samples to fall close to the manifold. We will show that when the noise distribution is known one can use a statistical deconvolution procedure to obtain a "deconvolved measure" concentrated around the manifold from which we can in turn draw a small number of samples and apply the cleaning procedure described above to them. Deconvolution has been extensively studied in the statistical literature (see [10] and references therein). Most related to our application is the work of Koltchinskii [16] who uses deconvolution to estimate the dimension and cluster tree of a distribution supported on a submanifold. We defer a detailed comparison to Section 5.4.1 after the necessary preliminaries have been introduced.

To the best of our knowledge, ours is the first paper to obtain lower and upper minimax bounds for the problem of inferring the homology of a manifold. There are a few existing results on upper bounds. A summary of previous results and the results in this paper are in Table 1.

Outline. In Section 2 we describe the statistical model. In Section 3 we give a brief description of homology. In Section 4 we give an overview of our techniques. We derive the minimax rates for the four noise settings in Section 5. Technical proofs are contained in the Appendix.

2 Statistical Model

We assume that the sample $\{Y_1, \ldots, Y_n\} \subset \mathbb{R}^D$ constitutes a set of "noisy" observations of an unknown d-dimensional manifold M, with $d < D$, whose homology we seek to estimate. The distribution of the sample depends on the properties of the manifold M as well as on the type of sampling noise, which we describe below by formulating various statistical models for sampling data from manifolds.

Notation. We let $B_r^k(x)$ denote a k-dimensional ball of radius r centered at x. When $k = D$, we write $B_r(x)$ instead of $B_r^D(x)$. For any set M and any $\sigma > 0$ define $\mathsf{tube}_\sigma(M) = \bigcup_{x \in M} B_\sigma(x)$. Let v_k denote the volume of the k-dimensional unit ball. Finally, for clarity we let $c_1, c_2, \ldots, C_1, C_2, \ldots$ denote various positive constants whose value can be different in different expressions. The constants will be specified in the corresponding proofs.

Manifold Assumptions. We assume that the unknown manifold M is a d-dimensional smooth compact Riemannian manifold without boundary embedded in the compact set $\mathcal{X} = [0,1]^D$. We further assume that the volume of the manifold is bounded from above by a constant which can depend on the dimensions d, D, i.e. we assume $\mathrm{vol}(M) \leq C_{D,d}$. We will also make the further assumption that $D > d$. The main regularity condition we impose on M is that its *condition number* be not too small. The *condition number* $\kappa(M)$ (see [17]) is the largest number τ such that the open normal bundle about M of radius r is imbedded in $\mathbb{R}^D$ for every $r < \tau$. For $\tau > 0$ let $\mathcal{M} \equiv \mathcal{M}(\tau)=\Big\{M : \kappa(M) \geq \tau\Big\}$ denote the set of all such manifolds with condition number no smaller than τ. A manifold with large condition number does not come too close to being self-intersecting. We consider the collection $\mathcal{P} \equiv \mathcal{P}(\mathcal{M}) \equiv \mathcal{P}(\mathcal{M}, a)$ of all probability distributions supported over manifolds M in $\mathcal{M}$ having densities p with respect to the volume form on M uniformly bounded from below by a constant $a > 0$, i.e. $0 < a \leq p(x) < \infty$ for all $x \in M$. For expositional clarity we treat a as a *fixed* constant although our upper and lower bounds match in their dependence on a.

The Noise Models. We consider four noise models and, for each of them, we specify a class $\mathcal{Q}$ of probability distributions for the sample.

Noiseless. We observe data $Y_1, \ldots, Y_n \sim P$ where $P \in \mathcal{P}$. In this case, $\mathcal{Q} = \mathcal{Q}(\tau) = \mathcal{P}$.

Clutter Noise. We observe data $Y_1, \ldots, Y_n$ from the mixture $Q = (1-\pi)U + \pi P$ where, $P \in \mathcal{P}$, $0 \leq \pi \leq 1$ and U is a uniform distribution on $\mathcal{X}$. The points drawn from U are called background clutter. Then $\mathcal{Q} = \mathcal{Q}(\pi, \tau) = \Big\{Q = (1-\pi)U + \pi P : \ P \in \mathcal{P}\Big\}$. Notice that $\pi = 1$ reduces to the noiseless case.

Tubular Noise. We observe $Y_1, \ldots, Y_n \sim Q_{M,\sigma}$ where $Q_{M,\sigma}$ is uniform on a tube of size σ around M. In this case $\mathcal{Q} = \mathcal{Q}(\sigma, \tau) = \Big\{Q_{M,\sigma} : \ M \in \mathcal{M}\Big\}$.

Additive Noise. The data are of the form $Y_i = X_i + \epsilon_i$, where $X_1, \ldots, X_n \sim P$, for some $P \in \mathcal{P}$, and $\epsilon_1, \ldots, \epsilon_n$ are a sample from a noise distribution Φ. Note that $Q = P \star \Phi$, that is, Q is the convolution of P and Φ. We consider two cases:

1. Φ is a D-dimensional Gaussian with mean $(0, \ldots, 0)$ and covariance $\sigma^2 I$, with $\sigma \ll \tau$. Define

	Noise Model				
	Noiseless	**Clutter**	**Tubular**	**Additive Gaussian**	**General additive (τ fixed)**
Upper Bound	NSW	This paper	NSW	This paper	This paper
Lower Bound	This paper	This paper	This paper	This paper	This paper

Table 1: Summary of our contributions

$\mathcal{Q} = \mathcal{Q}(\sigma, \tau) = \Big\{Q = P \star \Phi : \ P \in \mathcal{P}\Big\}$.

2. Φ is any known noise distribution whose Fourier transform is bounded away from 0 but with the added restriction that we only consider manifolds with τ being a fixed constant. Then $\mathcal{Q} = \mathcal{Q}(\Phi) = \Big\{Q = P \star \Phi : \ P \in \mathcal{P}_\tau\Big\}$. where $\mathcal{P}_\tau$ is the subset of $\mathcal{P}$ comprised of distributions supported on manifolds M with condition number at least as large as the *fixed* value τ.

The noise model used in [18] is to take the noise at any point to be only along the normal fibres; this seems unnatural and we will not consider that model here.

In almost all of the distribution classes considered we allow for τ to vanish as n gets bigger, which is equivalent to letting the difficulty of the statistical problem increase with the sample size. To this end, we will also analyze the quantity $\tau_n \equiv \tau_n(\epsilon) = \inf\{\tau : \ R_n \leq \epsilon\}$, which corresponds to the smallest condition number that permits accurate estimation. We call this the *resolution*.

3 Homology

Often in our paper we will use phrases like "the homology of the union of balls around samples". In this section we explain this usage and discuss briefly *simplicial homology* (see Hatcher (2001) for a detailed treatment) and its computation.

The homology $\mathcal{H}$ of a space M is a collection of groups that correspond to topological features of M. We will consider the case when M is a compact Riemannian manifold. In what follows, it might help the reader's intuition to imagine that we are starting with a dense sample of points U on the manifold and building a collection of simplices from these points. The union of balls $\bigcup_{y \in U} B_\epsilon(y)$ gives a geometric approximation to the underlying manifold. This is however a continuous (infinite) collection of points. To make computation tractable we need to be able to reduce the computation of homology from a continuous space to its discretization. The Čech complex (a particular *simplicial complex*, see Figure 3) which is described below gives a discrete representation of the union of balls. A classic result in topology called the Nerve Theorem [12] states that the homology of $\bigcup_{y \in U} B_\epsilon(y)$ is identical to the homology of the corresponding Čech complex.

We now describe a simplicial complex and its homology. A *simplicial complex* is a hereditary set system $\mathcal{K}$ over a vertex set V, i.e. $\sigma \subset \sigma' \in \mathcal{K}$ implies that $\sigma \in \mathcal{K}$. The *dimension* of a simplex σ is $|\sigma| - 1$; singletons are 0-simplices or vertices, pairs in $\mathcal{K}$ are 1-simplices or edges, triples are 2-simplices or triangles, etc. A p-chain is a formal sum of p-simplices. The coefficients are taken in $Z/2Z$, the integers mod 2.[1] Thus, chains may be viewed as subsets of simplices and addition (mod 2) as symmetric difference of sets. Addition of chains forms an abelian group called the *chain group* C_p with 0 denoting the empty chain.

A p-simplex $\sigma = \{v_0, \ldots, v_p\}$ has $p+1$ simplices of dimension $p-1$ on its boundary, denoted $\sigma_i = \sigma \setminus \{v_i\}$. The *boundary* of a simplex is $\partial_p \sigma = \sum_{i=0}^{p} \sigma_i$. The *boundary operator* $\partial_p : C_p \to C_{p-1}$ is the natural extension of the boundary of a simplex to the boundary of a chain: $\partial_p c = \sum_{\sigma \in c} \partial_p \sigma$.

The kernel and image of the boundary operator are two important subgroups of the chain group: *the cycle group*: $Z_p = \ker \partial_p = \{z \in C_p : \partial_p z = 0\}$, and *the boundary group*: $B_p = \mathrm{im}\ \partial_p = \{\partial_{p+1} c : c \in C_{p+1}\}$. The *cycles* Z_p are those chains that have boundary 0. The *boundary cycles* B_p are those p-chains that are the boundary of some $p+1$-chain. It is easy to check that $\partial_{p-1}\partial_p c = 0$ and thus $B_p \subset Z_p \subset C_p$. See Figure 1.

Two cycles $z_1, z_2 \in Z_p$ are *homologous* if $z_1 - z_2 \in B_p$, i.e. their difference is the boundary of a $p+1$-chain. The pth homology group H_p is defined as the quotient group Z_p/B_p. That is, the homology group is a collection of equivalence classes of cycles. The first homology group H_0 corresponds to connected components (clusters). The next homology group H_1 corresponds to non-bounding cycles (or loops). Higher order homology groups correspond to equivalence classes of higher dimensional cycles.[2] The homology of $\mathcal{K}$ is the collection $\mathcal{H}$ of all its homology groups.

The Čech complex is a specific simplicial complex defined as follows. Fix some $\epsilon > 0$ and a set of points $S \subset \mathbb{R}^D$. The *Čech complex* consists of all simplices σ such that $\bigcap_{x \in \sigma} B_\epsilon(x) \neq \emptyset$ where $B_\epsilon(x)$ is a ball of radius ϵ centered at x. See Figure 3.

[1] In general, homology may be defined over any ring, but we stick with $\mathbb{Z}_2$ for ease of exposition and computation.

[2] Intuitively, boundary cycles are "filled in" cycles and two cycles are homologous if one cycle can be deformed into the other cycle.

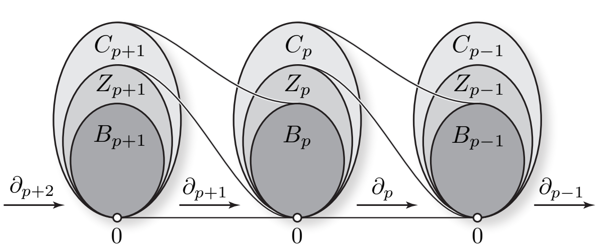

Figure 1: Relationship between chains C_p, cycles $Z_p = \ker \partial_p$ and boundaries $B_p = \mathrm{im}\ \partial_{p+1}$. The chains C_p are just collections of simplices. The chains in Z_p are the cycles. The cycles in B_p are the cycles that happen to be boundaries of chains in C_{p+1}.

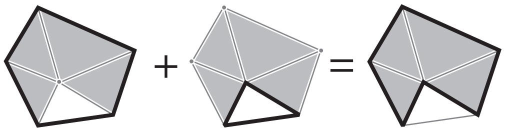

Figure 2: The sum of two 1-cycles is another 1-cycle. Here the cycles are homologous because their sum (in $\mathbb{Z}_2$)is the boundary of a 2-chain of triangles.

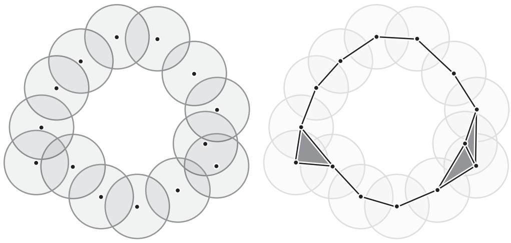

Figure 3: A union of balls and its corresponding Čech complex.

Since the coefficient ring is a field, the computations may be completely described by linear algebra. The groups C_p, Z_p, B_p, and H_p are vector spaces and the boundary operators are linear maps. It is possible to efficiently compute the homology groups of a simplicial complex in time polynomial in the size of the complex. The algorithm only involves row reduction on the matrix representations of ∂_p.

4 Techniques

4.1 Techniques for lower bounds

The *total variation distance* between two measures P and Q is defined by $\mathsf{TV}(P,Q) = \sup_A |P(A) - Q(A)|$ where the supremum is over all measurable sets. It can be shown that $\mathsf{TV}(P,Q) = P(G) - Q(G) = 1 - \int \min(P,Q)$ where $G = \{y :\ p(y) \geq q(y)\}$ and p and q are the densities of P and Q with respect to any measure μ that dominates both P and Q.

We shall make repeated use of Le Cam's lemma which we now state (see, e.g., Lemma 1 in [23]).

Lemma 1 (Le Cam). *Let $\mathcal{Q}$ be a set of distributions. Let $\theta(Q)$ take values in a metric space with metric ρ. Let $Q_1, Q_2 \in \mathcal{Q}$ be any pair of distributions in $\mathcal{Q}$. Let $Y_1, \ldots, Y_n$ be drawn iid from some $Q \in \mathcal{Q}$ and denote the corresponding product measure by Q^n. Then*

$$\inf_{\widehat{\theta}} \sup_{Q \in \mathcal{Q}} \quad \mathbb{E}_{Q^n}\left[\rho(\widehat{\theta}, \theta(Q))\right] \geq$$
$$\frac{1}{8}\rho(\theta(Q_1), \theta(Q_2))\ (1 - \mathsf{TV}(Q_1, Q_2))^{2n}$$

where the infimum is over all estimators.

Le Cam's lemma makes precise the intuition that if there are *distinct* members of the class $\mathcal{Q}$ for which the data generating distributions are close then the statistical problem is hard given a small sample.

When we apply Le Cam's lemma in this paper, Q_1 and Q_2 will be associated with two different manifolds M_1 and M_2. We will take $\theta(Q)$ to be the homology of the manifold and $\rho(\theta(Q_1), \theta(Q_2)) = 1$ if the homologies are the different and $\rho(\theta(Q_1), \theta(Q_2)) = 0$ if the homologies are the same. The subtlety of establishing *tight* lower bounds boils down to the task of finding a set of distributions in the class $\mathcal{Q}$ for which the homology of the underlying submanifolds are distinct but whose empirical distributions are hard to distinguish from a small number of samples.

We will use two representative manifolds M_1 and M_2 in the application of LeCam's lemma which we describe here. See Figure 4. The manifold M_1 is a pair of $1-\tau$ d-balls (shown in blue) embedded 2τ apart in $\mathbb{R}^D$ joined smoothly at their ends (shown in red). The manifold M_2 is a pair of d-annuli (shown in blue) embedded 2τ apart with outer radius $1-\tau$ and inner radius 4τ, smoothly joined at both the inner and outer ends (shown in red). It is clear from the construction that both these manifolds are d-dimensional compact, have no boundary and have condition number τ. It is also the case that $\mathcal{H}(M_1) \neq \mathcal{H}(M_2)$.

If there exist two manifolds M_1 and M_2 with corre-

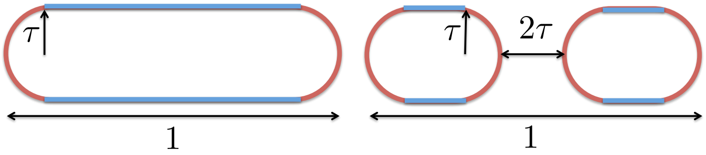

Figure 4: The two manifolds M_1 and M_2, with d = 1, D = 2

sponding distributions Q_1 and Q_2 in $\mathcal{Q}$ such that (i) $\mathcal{H}(M_1) \neq \mathcal{H}(M_2)$ and (ii) $Q_1 = Q_2$ then we say that the model $\mathcal{Q}$ is *non-identifiable*. In this case, recovering the homology is impossible and we write $R_n = 1$ and $n(\epsilon) = \infty$.

4.2 Techniques for upper bounds

To establish an upper bound we need to construct an estimator that achieves the upper bound. In the noiseless and tubular noise cases the samples are in a thin region around the manifold and our estimator is constructed from a union of balls (of a carefully chosen radius) around the sample points.

In the case of clutter noise and additive Gaussian noise samples are concentrated around the manifold but a few samples may be quite far away from the manifold. In these cases our upper bounds are obtained by analyzing the performance of the Algorithm 1 (CLEAN) with a carefully specified threshold and radius, which is used to remove points in regions of low density far away from the manifold. Our estimator is then constructed from a union of balls around the remaining points. In the case of additive noise with general known distribution the samples are not expected to be concentrated around the manifold. We will first use deconvolution to estimate a deconvolved measure $\widehat{P}_n$ which we will show is densely concentrated in a thin region around the manifold. We will then draw samples from this measure, clean them and construct a union of balls of appropriate radius around the remaining samples, and show that this set has the right homology with high probability.

Algorithm 1 CLEAN

```
• IN: $(X_i)_{i=1}^n$, threshold $t$, radius $r$
1. Construct a graph $G_r$ with nodes $\{X_i\}_{i=1}^n$. Include edge $(X_i, X_j)$, if $||X_i - X_j|| \leq r$.
2. Mark all vertices with degree $d_i \leq (n-1)t$.
• OUT: All unmarked vertices
```

We now briefly review statistical deconvolution. We refer the interested reader to the work of Fan [10] for more details and to [16] for an application related to ours. The procedure is similar to kernel density estimation with a kernel modified to account for the additive noise. For symmetric noise distributions Φ, we consider two kernels $\mathcal{K}$ and Ψ such that $\mathcal{K} \star \Phi = \Psi$, where $\star$ denotes convolution. The deconvolution estimator is $\widehat{P}_n(A) = 1/n \sum_{i=1}^n \mathcal{K}(Y_i - A)$. It is easy to verify that $E\widehat{P}_n = P \star \Psi$ similar to regular kernel density estimation with the kernel Ψ. In the noiseless case we can even take $\mathcal{K} = \Psi = \delta_0$ (a Dirac at 0) and get back the empirical distribution of the sample. More generally, we will be interested in Ψ that satisfies $\Psi\{x : |x| \geq \epsilon\} \leq \gamma$. for ϵ and γ that we will later specify.

In each of the above cases our final estimator is constructed from a union of balls around appropriate points, and our theorems will show that these have the correct homology with high probability. To compute the homology one would construct the corresponding Čech complex and compute its "boundary matrices" (as described in Section 3). Recovering the homology from these matrices consists of linear algebraic manipulation. There are several fast algorithms to compute the homology (either exactly [4] or approximately [5]) of the Čech complexes from large point sets in high dimensions.

5 Minimax Rates

We now derive the minimax rates for homology estimation under the four noise models described in section 2. There are three quantities of interest: the minimax risk R_n, the resolution τ_n and the sample complexity $n(\epsilon)$. We write $R_n \asymp a_n$ (similarly for $\tau_n \asymp a_n$) if there are positive constants c and C such that $c \leq R_n/a_n \leq C$ for all large n. Similarly, we write $n(\epsilon) \asymp a(\epsilon)$ if there are positive constants c and C such that $c \leq n(\epsilon)/a(\epsilon) \leq C$ for all small ϵ. Our analysis will show that the rates (as a function of n) are typically polynomial for the resolution and exponential for the risk. We will often match upper and lower bounds on sample complexity and resolution only up to logarithmic factors, and correspondingly those on the risk upto polynomial factors. In this case we will use the notation $R_n \asymp^* a_n$, $\tau_n \asymp^* a_n$, and $n(\epsilon) \asymp^* a(\epsilon)$.

It is worth emphasizing at this point that despite the fact that we use two specific manifolds in the application of Le Cam's lemma, the resulting lower bound holds for *all* manifolds in $\mathcal{M}$ and *all* distributions in $\mathcal{Q}$. Le Cam's lemma allows one to get a lower bound that holds for *any* estimator by using *two* carefully chosen distributions in $\mathcal{Q}$. The upper bounds are from specific estimators and they establish an upper bound on the number of samples to estimate the homology of any manifold in our class.

5.1 Noiseless Case

Theorem 1. *For all $\tau \leq \tau_0(a,d)$, in the noiseless case the minimax rate, $R_n \asymp^* e^{-n\tau^d}$, where $\tau_0(a,d)$ is a constant which depends on a and d. Also, $n(\epsilon) \asymp^* \tau^{-d}\log(1/\epsilon)$ and $\tau_n \asymp^* ((1/n)\log(1/\epsilon))^{1/d}$.*

We provide proof sketches for the lower and upper bounds on R_n separately.

Lower Bound: Proof Sketch

To obtain a lower bound on the minimax risk over the class $\mathcal{Q}(\tau)$ we will consider the two carefully chosen manifolds M_1 and M_2 described earlier.

We further need to specify the density on each of the manifolds, and we choose two densities from $\mathcal{P}$ so that the data distributions are as similar as possible while respecting the constraint $p(x) \geq a$. The construction is described in more detail in the Appendix A.1.1, but for now it suffices to notice that the two densities can be constructed to differ only on the sets $W_1 = M_1 \setminus M_2$ and $W_2 = M_2 \setminus M_1$ and can be made as low as a on one of these sets. A straightforward calculation shows that

$$\mathsf{TV}(p_1,p_2) \leq a \max(\mathrm{vol}(W_1), \mathrm{vol}(W_2)) \leq C_d a \tau^d$$

where the constant C_d depends on d. Now, we apply Le Cam's lemma to obtain that

$$R_n \geq \frac{1}{8}\left(1 - C_d a\tau^d\right)^n \geq \frac{1}{8} e^{-2C_d n a \tau^d}$$

for all $\tau \leq \tau_0(a,d)$. $\tau_0(a,d)$ is a constant depending on a and d. The lower bound of Theorem 1 follows.

Upper Bound: Proof Sketch

In the noiseless case the samples are densely concentrated around the manifold and our estimator is constructed from a union of balls of radius $\tau/2$ around the sample points. The upper bound on the minimax risk follows from a straightforward modification of the results of [17]. For completeness, we reproduce an adaptation of their main homology inference theorem (Theorem 3.1) here.

Lemma 2. *[NSW] Let $0 < \epsilon < \tau$ and let $U = \bigcup_{i=1}^n B_\epsilon(X_i)$. Let $\widehat{\mathcal{H}} = \mathcal{H}(U)$. Let $\zeta_1 = \frac{\mathsf{vol}(M)}{a\cos^d\theta_1 \mathsf{vol}(B^d_{\epsilon/4})}$, $\zeta_2 = \frac{\mathsf{vol}(M)}{a\cos^d\theta_2 \mathsf{vol}(B^d_{\epsilon/8})}$, $\theta_1 = \sin^{-1}\frac{\epsilon}{8\tau}$ and $\theta_2 = \sin^{-1}\frac{\epsilon}{16\tau}$. Then for all $n > \zeta_1\left(\log(\zeta_2) + \log\left(\frac{1}{\delta}\right)\right)$, $\mathbb{P}(\widehat{\mathcal{H}} \neq \mathcal{H}(M)) < \delta$.*

By assumption $\mathsf{vol}(M) \leq C_{D,d}$ for some constant $C_{D,d}$ depending on d and D. To obtain a sample complexity bound we simply choose $\epsilon = \tau/2$ and this gives us $n(\epsilon) \leq C_1/(a\tau^d)(C_2\log(1/(a\tau^d)) + \log(1/\epsilon))$ which matches the lower bound upto the factor of $\log(1/\tau)$. Further calculation (see Appendix A.1.1) then shows that as desired $R_n \leq C_1/\tau^d \exp(-C_2 n a \tau^d)$ for appropriate constants C_1, C_2, and $\tau_n \leq C(\frac{\log n \log(1/\epsilon)}{an})^{1/d}$. This establishes Theorem 1.

5.2 Clutter Noise

Theorem 2. *For all $\tau \leq \tau_0(a,d)$, in the clutter noise case, $R_n \asymp^* e^{-n\pi\tau^d}$, where $\tau_0(a,d)$ is a constant which depends on a and d. Also, $n(\epsilon) \asymp^* (1/(\pi\tau^d))\log(1/\epsilon)$ and $\tau_n \asymp^* (1/(n\pi)\log(1/\epsilon))^{1/d}$.*

Lower Bound: Proof Sketch

The lower bound for the class $\mathcal{Q}(\pi,\tau)$ follows via the same construction as in the noiseless case. In the calculation of the total variation distance (see Appendix A.1.2) we have instead

$$\mathsf{TV}(q_1,q_2) \leq \pi a \max(\mathrm{vol}(W_1), \mathrm{vol}(W_2)) \leq C_d \pi a \tau^d$$

where C_d depends on d. As before the lower bound follows from the application of Le Cam's lemma.

Upper Bound: Proof Sketch

As a preliminary step we clean the data samples to eliminate points that are far away from, while retaining those close to, the manifold. Our analysis shows that Algorithm 1 will achieve this, with high probability for a carefully chosen threshold and radius. We then show that taking a union of balls of the appropriate radius around the remaining points will give us the correct homology, with high probability. We give an outline here and defer details to Appendix A.1.2.

1. We define two regions $A = \mathsf{tube}_r(M)$ and $B = \mathbb{R}^D \setminus \mathsf{tube}_{2r}(M)$ where $r < \frac{(\sqrt{9}-\sqrt{8})\tau}{2}$.
2. We then invoke Algorithm CLEAN on the data with threshold $t = \left(\frac{v_D s^D(1-\pi)}{\mathrm{vol(Box)}} + \frac{\pi a v_d r^d \cos^d\theta}{2}\right)$ and radius $2r$. Let I be the set of vertices returned.
3. Through careful analysis we show that with high probability I contains *all* the vertices from the region A and *none* of the points in region B.
4. We further show that the retained points form a thin dense cover of the manifold M, i.e. $\left\{M \subset \bigcup_{i\in I} B_{2r}(X_i)\right\}$.

5. Using a straightforward corollary of Lemma 2 we show that this thin dense cover can be used to recover the homology of M with high probability.

Formally, in Appendix A.1.2 we prove the following lemma,

Lemma 3. *If $n > \max(N_1, N_2)$, and $r < (\sqrt{9} - \sqrt{8})\frac{\tau}{2}$ where $N_1 = 4\kappa \log(\kappa)$*

$$\begin{aligned} \text{with } \kappa &= \max\left(1 + \frac{200}{3\zeta}\log\left(\frac{2}{\delta}\right), 4\right) \\ \text{and } N_2 &= \frac{1}{\zeta}\left(\log\left(\frac{\mathrm{vol}(M)}{\cos^d(\theta) v_d r^d}\right) + \log\left(\frac{2}{\delta}\right)\right) \end{aligned}$$

where $\zeta = \pi a v_d r^d \cos^d(\theta)$ and $\theta = \sin^{-1}(r/2\tau)$, then after cleaning the points $\{X_i : i \in I\}$ are all in $\mathsf{tube}_{2r}(M)$ and are $2r$ dense in M. Let $U = \bigcup_{i\in I} B_w(X_i)$ with $w = r + \frac{\tau}{2}$ and let $\widehat{\mathcal{H}} = \mathcal{H}(U)$. We have that $\widehat{\mathcal{H}} = \mathcal{H}(\mathcal{M})$ with probability at least $1 - \delta$.

Taking $r = (\sqrt{9} - \sqrt{8})\tau/4$, we obtain the sample complexity bound, $n(\epsilon) \leq \frac{C_1}{\pi\tau^d}(\log\frac{C_2}{\tau^d} + \log(C_3/\epsilon))$. Given this sample complexity upper bound, the upper bounds on minimax risk and resolution follow identical arguments to the noiseless case (Appendix A.1.1).

5.3 Tubular Noise

Under this noise model we get samples uniformly from a tubular region of width σ around the manifold. This model highlights an important phenomenon in high-dimensions. Although, we receive samples *uniformly* from a full D dimensional shape these samples concentrate tightly around a d dimensional manifold. We show that with some care we can still reconstruct the homology at a rate independent of D.

Theorem 3. *Under the tubular noise model we establish the following cases.*

1. *If $\sigma \geq 2\tau$ then the model is non-identifiable and hence, $R_n = 1$ and $n(\epsilon) > \infty$.*
2. *If $\sigma \leq C_0\tau$, with C_0 small and $\tau \leq \tau_0(a, d)$, then $R_n \asymp^* e^{-n\tau^d}$, where $\tau_0(a, d)$ is a constant which depends on a and d. Also, $n(\epsilon) \asymp^* 1/\tau^d$ and $\tau_n \asymp^* (\frac{1}{n}\log(1/\epsilon))^{1/d}$.*

Remark 1. *The case when σ is very close to τ is significantly more involved since it involves the* exact *calculation of the volume of the tubular region and establishing tight upper and lower bounds here is an open problem we are attempting to address in current work.*

Lower bound: Proof Sketch

1. When $d < D$ and $\sigma \geq 2\tau$ the two manifolds M_1 and M_2 that we have considered thus far are still identifiable because even when $\sigma \geq \tau$ M_2 has a "dimple" along the co-dimensions that M_1 does not. To show that the class $\mathcal{Q}$ is still not identifiable we require a different construction. Consider the manifolds M_1 and M_2 with two points placed above and below the manifold at a distance 2τ above their centers along each of the co-dimensions. Denote these new manifolds M_1' and M_2'. It is clear that $\mathcal{H}(M_1') \neq \mathcal{H}(M_2')$, however $Q_1' = Q_2'$ since the extra points hide the "dimple" and the two manifolds cannot be distinguished.
2. When $d < D$, and $\sigma \leq C_0\tau$ we return to our old constructions of M_1 and M_2. There is however an important difference in that the two manifolds differ on full D-dimensional sets, and one might suspect that $TV(q_1, q_2) = O(\tau^D)$ or perhaps $O(\sigma^{D-d}\tau^d)$. As we show in Appendix A.1.3 however, $TV(q_1, q_2)$ is still $O(\tau^d)$, and we recover an identical lower bound to the noiseless case.

Upper bound: Proof Sketch

We are interested in case when $\sigma \leq C_0\tau$ (in particular $\sigma < \tau/24$ will suffice). Our proof will involve two main steps which we sketch here.

1. We first show that if we consider balls of sufficiently large radius ϵ (compared to σ) then the probability mass in these balls is $O(\epsilon^d)$. This is a manifestation of the phenomenon alluded to earlier: inside large enough balls the mass is concentrated around the lower dimensional manifold. Precisely, define $k_\epsilon = \inf_{p\in M} Q(B_\epsilon(p))$. In Lemma 9 in the Appendix, we show that, if $\epsilon \gg \sigma$ is large, k_ϵ is of order $\Omega(\epsilon^d)$.
2. There is however a disadvantage to considering balls that are too large. The homology of the union of balls around the samples may no longer have the right homology. Using tools from NSW, we show in the Appendix that we can balance these two considerations for manifolds with high condition number, i.e. provided $\sigma < \tau/24$, we can choose balls that are both large relative to σ and whose union still has the correct homology.

We will prove the following main lemma in the Appendix.

Lemma 4. *Let N_ϵ be the ϵ-covering number of the submanifold M. Let $U = \bigcup_{i=1}^n B_{\epsilon+\tau/2}(X_i)$. Let $\widehat{\mathcal{H}} = \mathcal{H}(U)$. Then if $n > \frac{1}{k_\epsilon}(\log(N_\epsilon) + \log(1/\delta))$, $\mathbb{P}(\widehat{\mathcal{H}} \neq \mathcal{H}(M)) < \delta$ as long as $\sigma \leq \epsilon/2$ and $\epsilon < \frac{(\sqrt{9}-\sqrt{8})\tau}{2}$.*

Notice, that we require $\sigma < \frac{(\sqrt{9}-\sqrt{8})\tau}{4}$ which is satisfied if $\sigma < \tau/24$ (for instance). To obtain the upper bound set $\epsilon = 2\sigma$, and observe that $N_\epsilon = O(1/\epsilon^d) = O(1/\tau^d)$

and $k_\epsilon = O(\epsilon^d) = O(\tau^d)$. This gives us that if $n \geq \frac{C_1}{\tau^d}(\log(\frac{C_2}{\tau^d}) + \log(\frac{1}{\delta}))$ we recover the right homology with probability at least $1-\delta$. The upper bound on minimax risk and resolution follows from similar arguments to those made previously.

5.4 Additive Noise

For additive noise we consider two cases. In the first case, we derive the minimax rates for additive *Gaussian* noise under the somewhat restrictive assumption that $C\sqrt{D}\sigma < \tau$. This problem is related of the problem of separating mixtures of Gaussians (which corresponds to the case where the manifold is a collection of points and 2τ is the distance between the closest pair). In this case have the following theorem.

Theorem 4. *For all $\tau \leq \tau_0(a,d)$ and $8\sqrt{D}\sigma < \tau$, $R_n \asymp^* e^{-n\tau^d}$, where $\tau_0(a,d)$ is a constant which depends on a and d. Also, $n(\epsilon) \asymp^* (1/\tau^d)\log(1/\epsilon)$ and $\tau_n \asymp^* ((1/n)\log(1/\epsilon))^{1/d}$.*

As in the clutter noise case we need to first clean the data and then take a union of balls around the points which survive. We analyze this procedure in the Appendix.

5.4.1 Deconvolution

Here we consider more general *known* noise distributions but work over the class of distributions $\mathcal{Q}(\Phi)$ over manifolds with τ fixed. We first use deconvolution to estimate a deconvolved measure $\widehat{P}_n$ which is concentrated around the manifold. We then draw samples from this measure, clean them and construct a union of balls H around these samples, and show that H has the right homology with high probability. The class of noise distributions we will consider satisfy the following assumption on its density.

Assumption 1. *Denote $\rho(R) = \inf_{|t|_\infty \leq R} |\Phi^\star(t)|$, where $R > 0, |t|_\infty = \max_{1\leq j\leq m}|t_j|$ and $\Phi^\star(t)$ is the Fourier transform of the symmetric noise density Φ. We assume $\rho(R) > 0$.*

This is a standard assumption in the literature on deconvolution (see [10, 16]), since as described deconvolution requires us to divide by the Fourier transform of the noise which needs to be bounded away from 0 for the procedure to be well behaved. The assumption is satisfied by a variety of noise distributions including Gaussian noise. Our main result says that for this broad class of noise distributions the deconvolution procedure described above will achieve an optimal rate of convergence.

Theorem 5. *In the additive noise case with τ fixed for Φ satisfying Assumption 1. $R_n \asymp e^{-n}$. Hence, $n(\epsilon) \asymp \log(1/\epsilon)$.*

Lower Bound: Proof Sketch To obtain the lower bound one can consider the same construction from the previous subsection with additive Gaussian noise. If τ is taken to be fixed we obtain the desired bound.

Upper Bound: Proof Sketch Our proof of the upper bound follows similar lines to that of Koltchinskii [16]. We deviate in two significant aspects. Koltchinskii only assumes an upper bound on the density, which he shows is sufficient to estimate weak geometric characteristics like the dimension of the manifold. To show that we can accurately reconstruct its homology we require both an upper and lower bound and our methods are quite different. Koltchinskii uses an epsilon net of the *entire* compact set containing the manifold critically in his construction and his procedure is thus not implementable/practical. Our algorithm instead draws a small number of samples from the deconvolved measure and uses those to estimate the homology resulting in a practical procedure. We prove the following upper bound in the Appendix.

Lemma 5. *Given n samples from $\mathcal{Q}(\Phi)$ with Φ satisfying Assumption 1, there exist $C_1, C_2, c_1 > 0$ such that $P(\mathcal{H}(H) \neq \mathcal{H}(M)) \leq C_1 e^{-c_1 n}$, where H is a union of balls of radius $\frac{5\epsilon+\tau}{2}$ centered around $m \geq C_2 n$ samples drawn from the deconvolved measure $\widehat{P}_n$ with a kernel Ψ with parameters γ, ϵ (specified in the proof). The samples are cleaned using the deconvolved measure by considering balls of radius 4ϵ at a threshold 2γ.*

Remark 2. *The cleaning procedure we use here is different from the Algorithm* CLEAN. *We remove points around which a ball of appropriate radius has low probability mass under the deconvolved measure. This is equivalent to using the deconvolved measure in place of the k-NN density estimate implicitly constructed by the* CLEAN *procedure.*

Simple calculations show that this lemma together with the lower bound give the exponential minimax rate described in Theorem 5.

6 Conclusion

We have given the first minimax bounds for homology inference. These bounds give insight into the intrinsic difficulty of the problem under various assumptions. Our bounds show that it is often possible to estimate the homology of a manifold at fast rates independent of the ambient dimension.

Actual implementation of homology inference has become tractable thanks to advances in computational topology. However, as our proofs reveal, recovering the homology requires the careful selection of several tuning parameters. In current work, we are developing methods for choosing these parameters in a statistically sound, data-driven way.

References

[1] Robert J. Adler, Omer Bobrowski, Matthew S. Borman, Eliran Subag, and Shmuel Weinberger. Persistent homology for random fields and complexes. In James O. Berger, Tony Cai, and Iain M. Johnstone, editors, *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, pages 124–143. Institute of Mathematical Statistics, 2010.

[2] Frederic Chazal, David Cohen-Steiner, and Quentin Merigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 2011. to appear.

[3] Moo K. Chung, Peter Bubenik, and Peter T. Kim. Persistence diagrams of cortical surface data. In *Proceedings of the 21st International Conference on Information Processing in Medical Imaging*, IPMI '09, pages 386–397, Berlin, Heidelberg, 2009. Springer-Verlag.

[4] Vin de Silva. *PLEX: Simplicial complexes in MATLAB.*

[5] Vin de Silva and Gunnar Carlsson. Topological estimation using witness complexes. In M. Alexa and S. Rusinkiewicz, editors, *Eurographics Symposium on Point-Based Graphics.* The Eurographics Association, 2004.

[6] Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358, 2007.

[7] Mary-Lee Dequeant, Sebastian Ahnert, Herbert Edelsbrunner, Thomas M. A. Fink, Earl F. Glynn, Gaye Hattem, Andrzej Kudlicki, Yuriy Mileyko, Jason Morton, Arcady R. Mushegian, Lior Pachter, Maga Rowicka, Anne Shiu, Bernd Sturmfels, and Olivier Pourqui. Comparison of pattern detection methods in microarray time series of the segmentation clock. *PLoS ONE*, 3(8):e2856, 08 2008.

[8] H. Edelsbrunner and J.H. Harer. *Computational topology.* American mathematical society, 2009.

[9] Herbert Edelsbrunner. *Computational Topology.* American Mathematical Society, 2009.

[10] Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, 1991.

[11] Jennifer Gamble and Giseon Heo. Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *Journal of Multivariate Analysis*, 101(9):2184–2199, October 2010.

[12] Allen Hatcher. *Algebraic Topology.* Cambridge University Press, 2002.

[13] Matthias Hein and Markus Maier. Manifold denoising. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *NIPS*, pages 561–568. MIT Press, 2006.

[14] Matthew Kahle. Topology of random clique complexes. *Discrete Mathematics*, 309(6):1658 – 1671, 2009.

[15] P. M. Kasson, A. Zomorodian, S. Park, N. Singhal, L. J. Guibas, and V. S. Pande. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759, 2007.

[16] V. I. Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *Ann. Statist.*, 28(2):591–629, 2000.

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

[18] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. A topological view of unsupervised learning and clustering. *SIAM J. Comput.*, 40(3):646–663, 2011.

[19] Valerio Pascucci, Xavier Tricoche, Hans Hagen, and Julien Tierny. *Topological methods in Data Analysis and Visualization: Theory, Algorithms and Applications.* Springer, 2001.

[20] Ahmet Sacan, Ozgur Ozturk, Hakan Ferhatosmanoglu, and Yusu Wang. Lfm-pro: a tool for detecting significant local structural sites in proteins. *Bioinformatics*, 23:709–716, February 2007.

[21] Vin De Silva and Robert Ghrist. Homological sensor networks. *Notices of the American Mathematical Society*, 54:2007, 2007.

[22] Gurjeet Singh, Facundo Memoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L. Ringach. Topological analysis of population activity in visual cortex. *J. Vis.*, 8(8):1–18, 6 2008.

[23] Bin Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[24] Afra Zomorodian. *Topology for Computing.* Cambridge University Press, 2005.