

TIGHT LOWER BOUNDS FOR HOMOLOGY INFERENCE

BY SIVARAMAN BALAKRISHNAN, ALESSANDRO RINALDO, AARTI SINGH AND LARRY WASSERMAN

*School of Computer Science and Statistics Department
Carnegie Mellon University*

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 earlier work [1], we have considered the statistical problem of estimating the homology of a manifold from noiseless samples and from noisy samples under several different noise models. We derived upper and lower bounds on the minimax risk for this problem. In this note we revisit the noiseless case. In [1], we used Le Cam's lemma to establish the lower bound ¹

$$R_n = \Omega(\exp(-n\tau^d))$$

for $d \geq 1$ and $D > d$. In the noiseless case the upper bound follows from the work of [2], who show that

$$R_n = O\left(\frac{1}{\tau^d} \exp(-n\tau^d)\right).$$

In this note we use a different construction based on the direct analysis of the likelihood ratio test to show that

$$R_n = \Omega\left(\frac{1}{\tau^d} \exp(-n\tau^d)\right),$$

as $n \rightarrow \infty$ thus establishing rate optimal asymptotic minimax bounds for the problem. The techniques we use here extend in a straightforward way to the noisy settings considered in [1]. Although, we do not consider the extension here non-asymptotic bounds are also straightforward.

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 (see [3]), 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 this note we study the problem of estimating the homology of a manifold M from a sample $\mathbf{X} = \{X_1, \dots, X_n\}$. Specifically, we bound the minimax risk

$$(1) \quad R_n \equiv \inf_{\hat{\mathcal{H}}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{\mathcal{H}} \neq \mathcal{H}(M))$$

where the infimum is over *all* estimators $\hat{\mathcal{H}}$ 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 $R_n = 1$

¹The asymptotic notation in both the upper and lower bounds hide constants that could depend on the dimensions d and D .

meaning that the problem is hopeless. Bounding the minimax risk is equivalent to bounding the *sample complexity* of the best possible estimator, defined by $n(\epsilon) = \min\{n : R_n \leq \epsilon\}$ where $0 < \epsilon < 1$.

We assume that the sample $\mathbf{X} \subset \mathbb{R}^D$ constitutes a set of 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 distribution of points on M . 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 $x \in M$.

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 $\text{vol}(M) \leq C_{D,d}$. 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 large. The *condition number* $\kappa(M)$ (see [2]) is $1/\tau$, where τ is the largest number 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

$$\mathcal{M} \equiv \mathcal{M}(\tau) = \left\{ M : \kappa(M) \geq \tau \right\}$$

denote the set of all such manifolds with condition number no smaller than τ . A manifold with small condition number does not come too close to being self-intersecting.

1.1. *Lower bounding the minimax risk.* In this note we will lower bound the minimax risk by considering a related *testing* problem.

Before describing the hypotheses we describe the null and alternate manifolds. The null manifold M_0 is a collection of m , d -spheres of radius τ , denoted S_1, \dots, S_m , with centers on one face of the unit hypercube in $d+1$ dimensions (M_0 is embedded in a space of dimension D which is of dimension at least $d+1$), with spacing between adjacent centers $= 4\tau$. It is easy to see that

$$m = O\left(\frac{1}{(4\tau)^d}\right)$$

because the manifold must be completely in $[0, 1]^D$, and that the manifold has condition number at least $1/\tau$. We will use

$$m = \Theta\left(\frac{1}{(4\tau)^d}\right)$$

in this note. Let P_0 denote the uniform distribution on M_0 .

The alternate manifolds are a collection $\{M_{1i} : i \in \{1, \dots, m\}\}$, where M_{1i} is M_0 with S_i removed. Let π denote the uniform distribution on $\{1, \dots, m\}$, and P_{1i} denote the uniform distribution on M_{1i} .

We need to ensure that the density p is lower bounded by a constant. Note that the total d -dimensional volume of M_0 is $v_d \tau^d m$, and so

$$p(x) \geq \frac{1}{v_d \tau^d m}$$

where v_d is the volume of the d -dimensional unit ball. This is $\Omega(1)$ as desired. A similar argument works for M_{1i} .

Consider the following testing problem:

$$\begin{aligned} H_0 &: \mathbf{X} \sim P_0 \\ H_1 &: \mathbf{X} \sim P_{1i} \text{ with } i \sim \pi \end{aligned}$$

A test T , is a measurable function of \mathbf{X} , in particular $T : \mathbf{X} \rightarrow \{0, 1\}$, and its risk is defined as

$$R_n^T := \mathbb{P}_{H_0}(T(\mathbf{X}) = 1) + \mathbb{P}_{H_1}(T(\mathbf{X}) = 0)$$

The relationship between testing and estimation is standard [4]. In our case it is easy to see that the estimation minimax risk of Equation 1 satisfies,

$$R_n^T \leq 2R_n$$

and so it suffices to lower bound R_n^T to obtain a lower bound on R_n . This relation is a straightforward consequence of the fact that $\mathcal{H}(M_0) \neq \mathcal{H}(M_{1i})$ for every i (since they have different number of connected components), and so any estimator can be used in the testing problem described.

The optimal test for the hypothesis testing problem described is the likelihood ratio test,

$$T(\mathbf{X}) = 0 \text{ if and only if } L(\mathbf{X}) \leq 1$$

where

$$L(\mathbf{X}) = \frac{L_1(\mathbf{X})}{L_0(\mathbf{X})}$$

where $L_1(\mathbf{X})$ and $L_0(\mathbf{X})$ are likelihoods of the data under the alternate and null respectively.

1.2. *Coupon collector lower bound.* We begin with a theorem from [5].

LEMMA 1 (Theorem 3.8 of [5]). *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then for any constant $c \in \mathbb{R}$, and $m = n \log n - cn$,*

$$\lim_{n \rightarrow \infty} \mathbb{P}(X > m) = 1 - \exp(-\exp(c))$$

2. Main result.

THEOREM 2. *For any constant $\delta < 1$, we have*

$$R_n \geq \Omega \left(\min \left(\frac{1}{\tau^d} \exp(-n\tau^d), \delta \right) \right)$$

as $n \rightarrow \infty$.

PROOF. Notice that since

$$m = \Theta \left(\frac{1}{(4\tau)^d} \right)$$

the theorem is implied by the statement that

$$n = m \log m + m \log \left(\frac{1}{\delta} \right) \implies R_n \geq c\delta$$

for some constant c . We will focus on proving this claim.

Let us consider the case when samples are drawn according to P_0 . From Lemma 1 we have that if

$$n = m \log m + m \log \left(\frac{1}{\delta} \right)$$

then the probability with which we do not see a point in each of the m spheres is

$$1 - \exp(-\exp(-\log 1/\delta)) \geq c\delta$$

since $\delta < 1$, for some constant c . It is easy to see that if we do not see a point in each of the m spheres then

$$L(\mathbf{X}) \geq \frac{1}{m} \frac{1}{(1 - 1/m)^n} := T_{m,n}$$

When $n = m \log m + m \log \left(\frac{1}{\delta} \right)$,

$$T_{m,n} \rightarrow \frac{1}{\delta} > 1$$

so asymptotically the likelihood ratio test always rejects the null.

From this we can see the probability of a Type I error $\rightarrow c\delta$, and $R_n^T \geq c\delta$, which gives

$$R_n \geq \frac{c}{2}\delta$$

as desired. □

3. Discussion. In this note we have established tight minimax rates for the problem of homology inference in the noiseless case. The intuition behind the construction extends to the noisy cases considered in [1] in a straightforward way.

Although the bound we have shown is an asymptotic lower bound, a finite sample lower bound follows in a straightforward way by replacing the asymptotic calculation in Lemma 1 with finite sample estimates.

We also expect similar constructions to be useful in establishing tight lower bounds for the problems of manifold estimation in Hausdorff distance considered in [6, 7], and for the problem of estimation of persistence diagrams in bottleneck distance considered in [8].

REFERENCES

- [1] Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. *AISTATS*, 2012.
- [2] 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.
- [3] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.
- [4] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 2005.
- [5] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- [6] Christopher R. Genovese, Marco Perone-Pacífico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40(2):941–963, 2012.
- [7] Christopher R. Genovese, Marco Perone-Pacífico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291, 2012.
- [8] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Optimal rates of convergence for persistence diagrams in topological data analysis, 2013.

SCHOOL OF COMPUTER SCIENCE
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PA 15213
E-MAIL: sbalakri@cs.cmu.edu
E-MAIL: aarti@cs.cmu.edu

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PA 15213
E-MAIL: arinaldo@cmu.edu
E-MAIL: larry@cmu.edu