Minimax Rates for Homology Inference

Don Sheehy

Joint work with
Sivaraman Balakrishan,
Alessandro Rinaldo,
Aarti Singh, and
Larry Wasserman
Something like a joke.
Something like a joke.

What is topological inference?
Something like a joke.

What is topological inference?

It’s when you infer the topology of a space given only a finite subset.
Something *like* a joke.

What is topological inference?

It’s when you infer the topology of a space given only a finite subset.
We add geometric and statistical hypotheses to make the problem well-posed.

Geometric Assumption:
The underlying space is a smooth manifold $M$.

Statistical Assumption:
The points are drawn i.i.d. from a distribution derived from $M$. 
We add geometric and statistical hypotheses to make the problem well-posed.

Geometric Assumption:
The underlying space is a smooth manifold $M$.

Statistical Assumption:
The points are drawn i.i.d. from a distribution derived from $M$. 
We add geometric and statistical hypotheses to make the problem well-posed.

Geometric Assumption:
The underlying space is a smooth manifold $M$.

Statistical Assumption:
The points are drawn i.i.d. from a distribution derived from $M$. 
We add geometric and statistical hypotheses to make the problem well-posed.

Geometric Assumption:
The underlying space is a smooth manifold $M$.

Statistical Assumption:
The points are drawn i.i.d. from a distribution derived from $M$. 
Input: $n$ points from a $d$-manifold $M$ in $D$-dimensions.
Input: $n$ points from a $d$-manifold $M$ in $D$-dimensions.

Output: The homology of $M$. 
Input: n points from a d-manifold $M$ in D-dimensions.

Output: The homology of $M$.

Upper bound: What is the worst case complexity?
Input: n points from a d-manifold M in D-dimensions.

Output: The homology of M.

Upper bound: What is the worst case complexity?

Lower Bound: What is the worst case complexity of the best possible algorithm?
Input: n points from a d-manifold M in D-dimensions.

Output: The homology of M.

Upper bound: What is the worst case complexity?

Lower Bound: What is the worst case complexity of the best possible algorithm?
Input: $n$ points from a $d$-manifold $M$ in $D$-dimensions.

Output: The homology of $M$.

Upper bound: What is the worst case complexity?

Lower Bound: What is the worst case complexity of the best possible algorithm?
Input: \( n \) points from a \( d \)-manifold \( M \) in \( D \)-dimensions.

Output: The homology of \( M \).

Upper bound: What is the worst case complexity?

Lower Bound: What is the worst case complexity of the best possible algorithm?
Input: $n$ points from a $d$-manifold $M$ in $D$-dimensions.

Output: The homology of $M$.

Upper bound: What is the worst case complexity?

Lower Bound: What is the worst case complexity of the best possible algorithm?
Input: \( n \) points from a \( d \)-manifold \( M \) in \( D \)-dimensions.

Output: The homology of \( M \).

Upper bound: What is the worst case complexity?

Lower Bound: What is the worst case complexity of the best possible algorithm?
Input: $n$ points from a $d$-manifold $M$ in $D$-dimensions.

Output: The homology of $M$.

Upper bound: What is the worst case complexity?

Lower Bound: What is the worst case complexity of the best possible algorithm?

The Goal: Matching Bounds (asymptotically)
Minimax risk is the error probability of the best estimator on the hardest examples.
Minimax risk is the error probability of the best estimator on the hardest examples.

Minimax Risk: \[ R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \]
Minimax risk is the error probability of the best estimator on the hardest examples.

Minimax Risk: \[ R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \]
Minimax risk is the error probability of the best estimator on the hardest examples.

Minimax Risk: \[ R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \]
Minimax risk is the error probability of the best estimator on the hardest examples.

Minimax Risk: \( R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \)
Minimax risk is the error probability of the best estimator on the hardest examples.

Minimax Risk: \[ R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \]

- the best estimator
- the hardest distribution
- product distribution
- the true homology
Minimax risk is the error probability of the best estimator on the hardest examples.

Minimax Risk: \[ R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \]

Sample Complexity: \[ n(\epsilon) = \min\{n : R_n \leq \epsilon\} \]
We assume manifolds without boundary of bounded volume and reach.
We assume manifolds without boundary of bounded volume and reach.

Let $\mathcal{M}$ be the set of compact $d$-dimensional Riemannian manifolds without boundary such that
We assume manifolds without boundary of bounded volume and reach.

Let \( \mathcal{M} \) be the set of compact \( d \)-dimensional Riemannian manifolds without boundary such that

\[
M \subset \text{ball}_D(0, 1)
\]
We assume manifolds without boundary of bounded volume and reach.

Let $\mathcal{M}$ be the set of compact $d$-dimensional Riemannian manifolds without boundary such that

1. $M \subset \text{ball}_D(0, 1)$
2. $\text{vol}(M) \leq c_d$
We assume manifolds without boundary of bounded volume and reach.

Let $\mathcal{M}$ be the set of compact $d$-dimensional Riemannian manifolds without boundary such that

1. $M \subset \text{ball}_D(0, 1)$
2. $\text{vol}(M) \leq c_d$
3. The reach of $M$ is at most $\tau$. 
We assume manifolds without boundary of bounded volume and reach.

Let $\mathcal{M}$ be the set of compact $d$-dimensional Riemannian manifolds without boundary such that

1. $M \subset \text{ball}_D(0, 1)$
2. $\text{vol}(M) \leq c_d$
3. The reach of $M$ is at most $\tau$.

Let $\mathcal{P}$ be the set of probability distributions supported over $M \in \mathcal{M}$ with densities bounded from below by a constant $a$. 
We consider 4 different noise models.

- **Noiseless**
- **Clutter**
- **Tubular**
- **Additive**
We consider 4 different noise models.

Noisy.\[ Q = P \]
We consider 4 different noise models.

- **Noiseless**
  \[ Q = \mathcal{P} \]

- **Clutter**
  \[ Q = (1 - \gamma)U + \gamma P \]
  \[ P \in \mathcal{P} \]
  \[ U \text{ is uniform on ball}(0, 1) \]

- **Tubular**

- **Additive**
We consider 4 different noise models.

<table>
<thead>
<tr>
<th>Noiseless</th>
<th>Clutter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q = \mathcal{P}$</td>
<td>$Q = (1 - \gamma)U + \gamma P$</td>
</tr>
<tr>
<td>$P \in \mathcal{P}$</td>
<td>$P \in \mathcal{P}$</td>
</tr>
<tr>
<td>$U$ is uniform on ball$(0, 1)$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tubular</th>
<th>Additive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Let $Q_{M,\sigma}$ be uniform on $M^\sigma$.</td>
<td>$Q = {Q_{M,\sigma} : M \in \mathcal{M}}$</td>
</tr>
</tbody>
</table>
We consider 4 different noise models.

**Noiseless**

\[ Q = \mathcal{P} \]

**Clutter**

\[ Q = (1 - \gamma)U + \gamma P \]

\[ P \in \mathcal{P} \]

\[ U \text{ is uniform on ball}(0,1) \]

**Tubular**

Let \( Q_{M,\sigma} \) be uniform on \( M^\sigma \).

\[ Q = \{ Q_{M,\sigma} : M \in \mathcal{M} \} \]

**Additive**

\[ Q = \{ P \ast \Phi : P \in \mathcal{P} \} \]

\[ \Phi \text{ is Gaussian with } \sigma \ll \tau \]

or \( \Phi \) has Fourier transform bounded away from 0 and \( \tau \) is fixed.
Le Cam’s Lemma is a powerful tool for proving minimax lower bounds.
Le Cam’s Lemma is a powerful tool for proving minimax lower bounds.

**Lemma.** Let $Q$ be a set of distributions. Let $\theta(Q)$ take values in a metric space $(X, \rho)$ for $Q \in Q$. For any $Q_1, Q_2 \in Q$,

$$\inf_{\hat{\theta}} \sup_{Q \in Q} \mathbb{E}_{Q^n} \left[ \rho(\hat{\theta}, \theta(Q)) \right] \geq \frac{1}{8} \rho(\theta(Q_1), \theta(Q_2))(1 - \text{TV}(Q_1, Q_2))^{2n}$$
Le Cam’s Lemma is a powerful tool for proving minimax lower bounds.

**Lemma.** Let $\mathcal{Q}$ be a set of distributions. Let $\theta(Q)$ take values in a metric space $(X, \rho)$ for $Q \in \mathcal{Q}$. For any $Q_1, Q_2 \in \mathcal{Q}$,

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

For homology, use the trivial metric. $\rho(x, y) = \begin{cases} 
0 & \text{if } x = y \\
1 & \text{if } x \neq y 
\end{cases}$
Le Cam’s Lemma is a powerful tool for proving minimax lower bounds.

**Lemma.** Let $\mathcal{Q}$ be a set of distributions. Let $\theta(Q)$ take values in a metric space $(X, \rho)$ for $Q \in \mathcal{Q}$. For any $Q_1, Q_2 \in \mathcal{Q}$,

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

For homology, use the trivial metric. $\rho(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$

$$\inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \geq \frac{1}{8}(1 - \text{TV}(Q_1, Q_2))^{2n}$$
Le Cam’s Lemma is a powerful tool for proving minimax lower bounds.

**Lemma.** Let $Q$ be a set of distributions. Let $\theta(Q)$ take values in a metric space $(X, \rho)$ for $Q \in Q$. For any $Q_1, Q_2 \in Q$,

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

For homology, use the trivial metric. $\rho(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$

$$R_n = \inf_{\hat{H}} \sup_{Q \in Q} Q^n(\hat{H} \neq H(M)) \geq \frac{1}{8}(1 - \text{TV}(Q_1, Q_2))^2 \frac{1}{n}$$
The lower bound requires two manifolds that are geometrically close but topologically distinct.
The lower bound requires two manifolds that are geometrically close but topologically distinct.

\[ B = \text{ball}_d(0, 1 - \tau) \]

\[ A = B \setminus \text{ball}_d(0, 2\tau) \]
The lower bound requires two manifolds that are geometrically close but topologically distinct.

\[
\begin{align*}
B &= \text{ball}_d(0, 1 - \tau) \\
A &= B \setminus \text{ball}_d(0, 2\tau) \\
M_1 &= \partial(B^\tau) \\
M_2 &= \partial(A^\tau)
\end{align*}
\]
The lower bound requires two manifolds that are geometrically close but topologically distinct.

\[ B = \text{ball}_d(0, 1 - \tau) \]

\[ A = B \setminus \text{ball}_d(0, 2\tau) \]

\[ M_1 = \partial(B^\tau) \]

\[ M_2 = \partial(A^\tau) \]

The overlap
It suffices to bound the total variation distance.
It suffices to bound the total variation distance.

Total Variation Distance:

\[ \text{TV}(Q_1, Q_2) = \sup_A |Q_1(A) - Q_2(A)| \]

\[ \leq a \max\{\text{vol}(M_1 \setminus M_2), \text{vol}(M_2 \setminus M_1)\} \]

\[ \leq C_d a \tau^d \]
It suffices to bound the total variation distance.

Total Variation Distance:

$$TV(Q_1, Q_2) = \sup_A |Q_1(A) - Q_2(A)|$$

$$\leq a \max \{\operatorname{vol}(M_1 \setminus M_2), \operatorname{vol}(M_2 \setminus M_1)\}$$

$$\leq C_d a \tau^d$$

Minimax Risk:

$$R_n \geq \frac{1}{8} (1 - TV(Q_1, Q_2))^{2n} \geq \frac{1}{8} (1 - C_d a \tau^d)^{2n} \geq \frac{1}{8} e^{-2C_d a \tau^d n}$$
It suffices to bound the total variation distance.

Total Variation Distance:

\[
TV(Q_1, Q_2) = \sup_A |Q_1(A) - Q_2(A)|
\leq a \max\{\text{vol}(M_1 \setminus M_2), \text{vol}(M_2 \setminus M_1)\}
\leq C_d a \tau^d
\]

Minimax Risk:

\[
R_n \geq \frac{1}{8} (1 - TV(Q_1, Q_2))^{2n} \geq \frac{1}{8} (1 - C_d a \tau^d)^{2n} \geq \frac{1}{8} e^{-2 C_d a \tau^d n}
\]

Sampling Rate:

\[
n(\epsilon) \geq \left(\frac{1}{\tau}\right)^d \log \frac{1}{\epsilon}
\]
The upper bound uses a union of balls to estimate the homology of M.
The upper bound uses a union of balls to estimate the homology of $M$. 
The upper bound uses a union of balls to estimate the homology of $M$. 
The upper bound uses a union of balls to estimate the homology of $M$.

1. Take a union of balls.
The upper bound uses a union of balls to estimate the homology of $M$.

1. Take a union of balls.

2. Compute the homology of the resulting Čech complex.
The upper bound uses a union of balls to estimate the homology of M.

1. Take a union of balls.
2. Compute the homology of the resulting Čech complex.
The upper bound uses a union of balls to estimate the homology of $M$.

0. Denoise the data.

1. Take a union of balls.

2. Compute the homology of the resulting Čech complex.
The upper bound uses a union of balls to estimate the homology of M.

0. Denoise the data.
1. Take a union of balls.
2. Compute the homology of the resulting Čech complex.
The upper bound uses a union of balls to estimate the homology of $M$.

0. Denoise the data.
1. Take a union of balls.
2. Compute the homology of the resulting Čech complex.
The upper bound uses a union of balls to estimate the homology of $M$.

0 Denoise the data.

1 Take a union of balls.

2 Compute the homology of the resulting Čech complex.
The upper bound uses a union of balls to estimate the homology of $M$.

0. Denoise the data.

1. Take a union of balls.

2. Compute the homology of the resulting Čech complex.
The upper bound uses a union of balls to estimate the homology of $M$.

1. Take a union of balls.
2. Compute the homology of the resulting Čech complex.

To prove: The density is bounded from below near $M$ and from above far from $M$. 

0. Denoise the data.
Many fundamental problems are still open.
Many fundamental problems are still open.

1. Is the reach the *right* parameter?
Many fundamental problems are still open.

1. Is the reach the *right* parameter?
2. What about manifolds with boundary?
Many fundamental problems are still open.

1. Is the reach the *right* parameter?
2. What about manifolds with boundary?
3. Homotopy equivalence?
Many fundamental problems are still open.

1. Is the reach the right parameter?
2. What about manifolds with boundary?
3. Homotopy equivalence?
4. How to choose parameters?
Many fundamental problems are still open.

1. Is the reach the *right* parameter?
2. What about manifolds with boundary?
3. Homotopy equivalence?
4. How to choose parameters?
5. Are there efficient algorithms?
Thank you.