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.

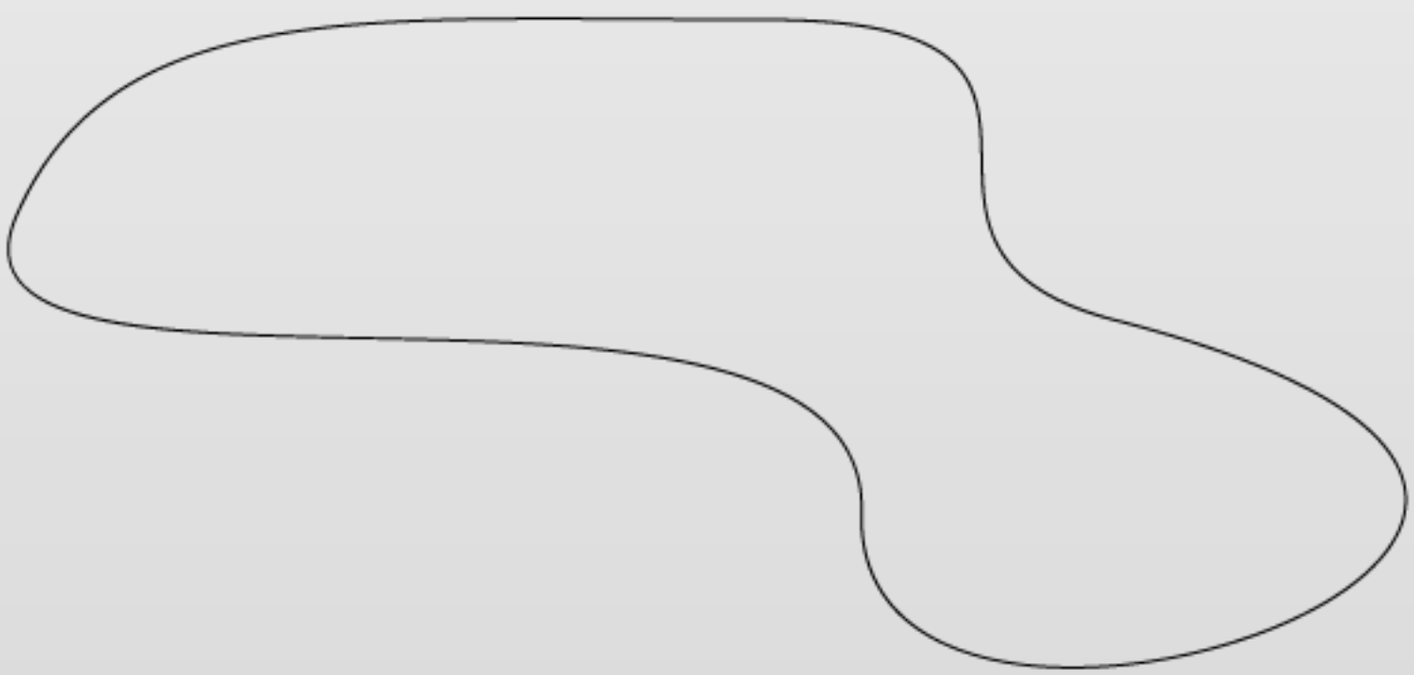
Geometric Assumption:
The underlying space is a smooth manifold M.

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?

sampled i.i.d.

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?

sampled i.i.d.

distribution supported on

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?

sampled i.i.d.

distribution supported on

Input: n points from a d-manifold M in D-dimensions.

with noise

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?

sampled i.i.d.

distribution supported on

Input: n points from a d-manifold M in D-dimensions.

with noise

an estimate of

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?

sampled i.i.d.

distribution supported on

Input: n points from a d-manifold M in D-dimensions.

with noise

an estimate of

Output: The homology of M.

Upper bound: What is the worst case ~~complexity~~?

probability of giving a wrong answer

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

sampled i.i.d.

distribution supported on

Input: n points from a d-manifold M in D-dimensions.

with noise

an estimate of

Output: The homology of M.

Upper bound: What is the worst case ~~complexity~~?

probability of giving a wrong answer

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: $$R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M))$$

the best estimator

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

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

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: $$R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M))$$

the best estimator

the hardest distribution

product distribution

the true homology

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

1. $M \subset \mathrm{ball}_D(0,1)$

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. $M \subset \mathrm{ball}_D(0,1)$
2. $\mathrm{vol}(M) \leq c_d$

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

We assume manifolds without boundary of bounded volume and reach.

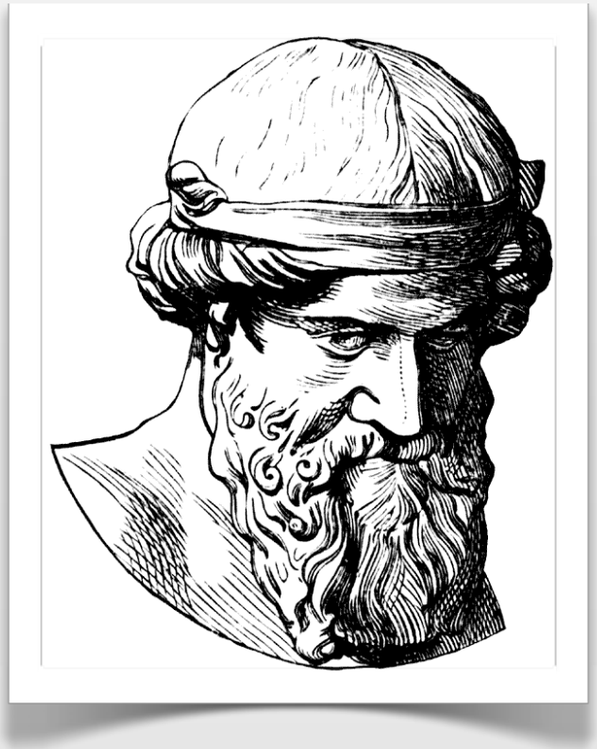
Let $\mathcal{M}$ be the set of compact d-dimensional Riemannian manifolds without boundary such that

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

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.

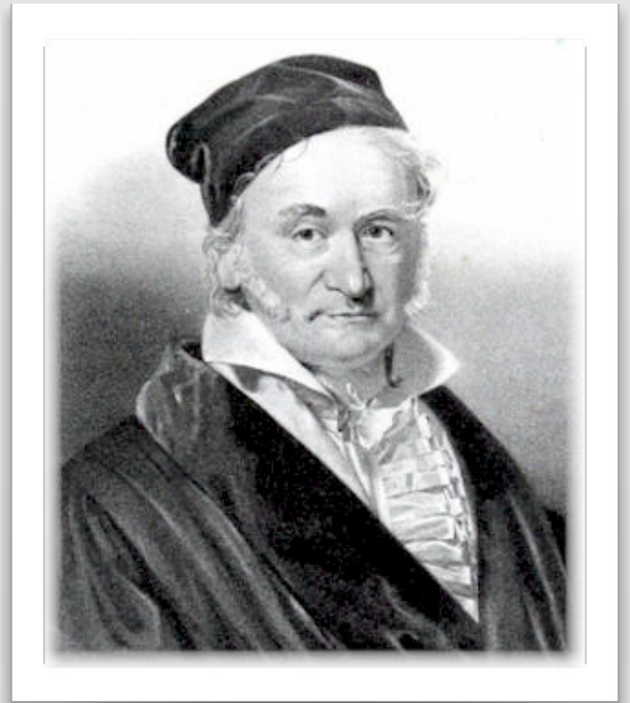
Noiseless



Clutter

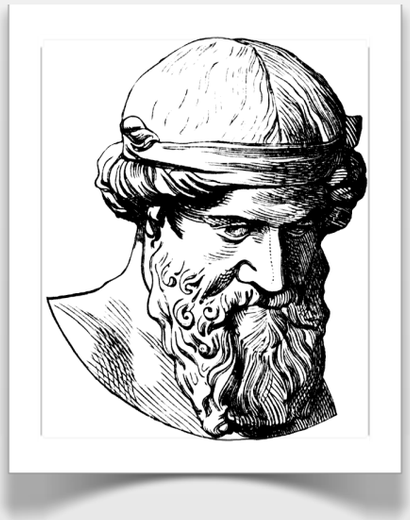
Tubular

Additive



We consider 4 different noise models.

Noiseless

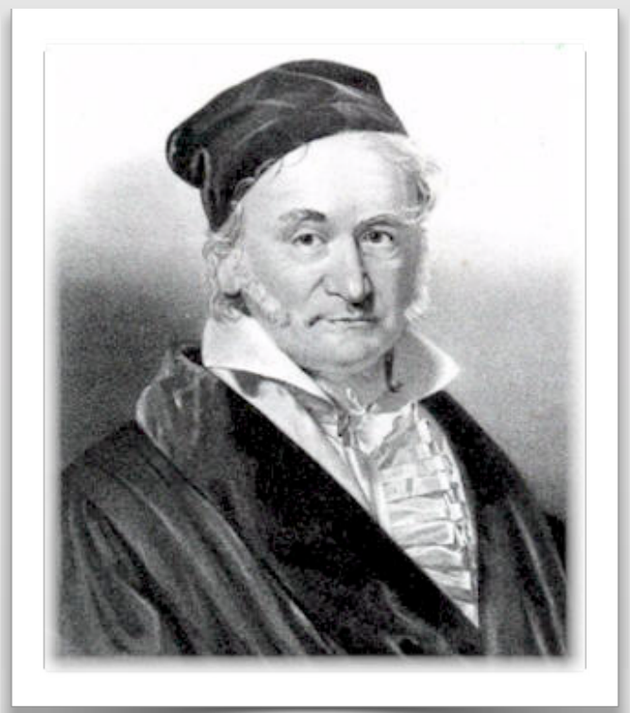


$$Q = \mathcal{P}$$

Clutter

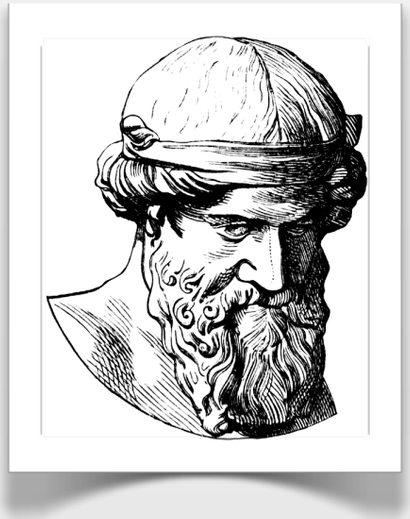
Tubular

Additive



We consider 4 different noise models.

Noiseless



$$\mathcal{Q} = \mathcal{P}$$

Clutter

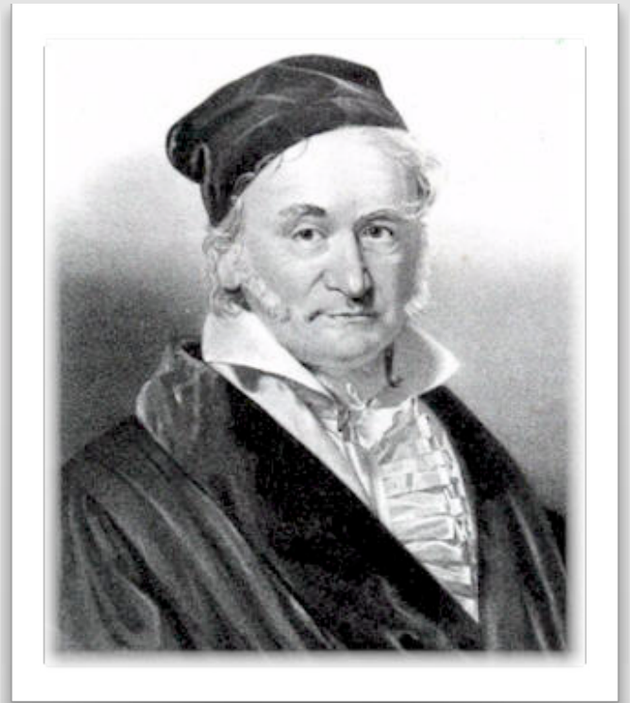
$$\mathcal{Q} = (1-\gamma)U + \gamma P$$

$$P \in \mathcal{P}$$

U is uniform on ball$(0, 1)$

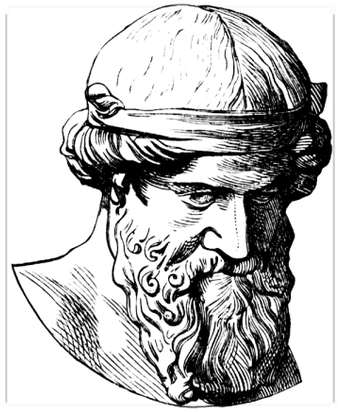
Tubular

Additive



We consider 4 different noise models.

Noiseless



$$\mathcal{Q} = \mathcal{P}$$

Clutter

$$\mathcal{Q} = (1-\gamma)U + \gamma P$$

$$P \in \mathcal{P}$$

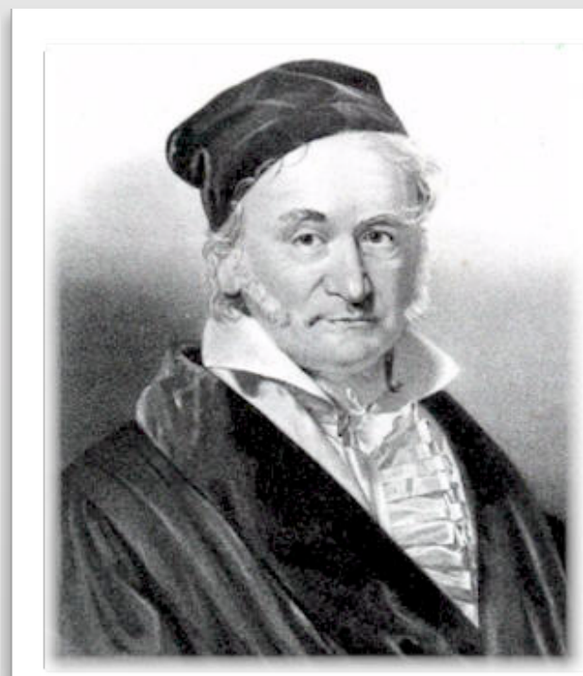
U is uniform on ball$(0, 1)$

Tubular

Let $Q_{M,\sigma}$ be uniform on M^σ.

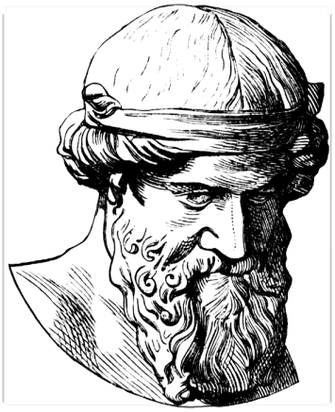
$$\mathcal{Q} = \{Q_{M,\sigma} : M \in \mathcal{M}\}$$

Additive



We consider 4 different noise models.

Noiseless



$$\mathcal{Q} = \mathcal{P}$$

Clutter

$$\mathcal{Q} = (1-\gamma)U + \gamma P$$

$$P \in \mathcal{P}$$

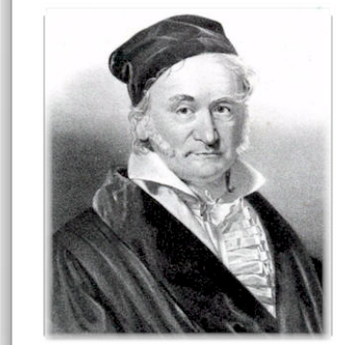
U is uniform on ball$(0,1)$

Tubular

Let $Q_{M,\sigma}$ be uniform on M^σ.

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

Additive



$$\mathcal{Q} = \{P \star \Phi : P \in \mathcal{P}\}$$

Φ is Gaussian with $\sigma \ll \tau$

or Φ has Fourier transform bounded away from 0 and τ 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 $\mathcal{Q}$ be a set of distributions. Let $\theta(Q)$ take values in a metric space (X, ρ) for $Q \in \mathcal{Q}$. For any $Q_1, Q_2 \in \mathcal{Q}$,*

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

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, ρ) for $Q \in \mathcal{Q}$. For any $Q_1, Q_2 \in \mathcal{Q}$,*

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

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, ρ) for $Q \in \mathcal{Q}$. For any $Q_1, Q_2 \in \mathcal{Q}$,*

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

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

$$\inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \geq \frac{1}{8} (1 - \mathrm{TV}(Q_1, Q_2))^{2n}$$

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, ρ) for $Q \in \mathcal{Q}$. For any $Q_1, Q_2 \in \mathcal{Q}$,*

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

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

$$R_n = \inf_{\hat{H}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{H} \neq H(M)) \geq \frac{1}{8}(1 - \mathrm{TV}(Q_1, Q_2))^{2n}$$

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.

$B = \mathrm{ball}_d(0, 1-\tau)$

$A = B \setminus \mathrm{ball}_d(0, 2\tau)$

The lower bound requires two manifolds that are geometrically close but topologically distinct.

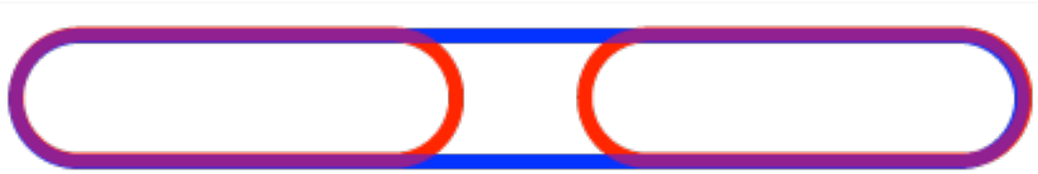
$$B = \mathrm{ball}_d(0, 1-\tau) \qquad A = B \setminus \mathrm{ball}_d(0, 2\tau)$$

$$M_1 = \partial(B^\tau) \qquad M_2 = \partial(A^\tau)$$

The lower bound requires two manifolds that are geometrically close but topologically distinct.

$$B = \mathrm{ball}_d(0, 1-\tau) \qquad A = B \setminus \mathrm{ball}_d(0, 2\tau)$$

$$M_1 = \partial(B^\tau) \qquad M_2 = \partial(A^\tau)$$



The overlap

It suffices to bound the total variation distance.

It suffices to bound the total variation distance.

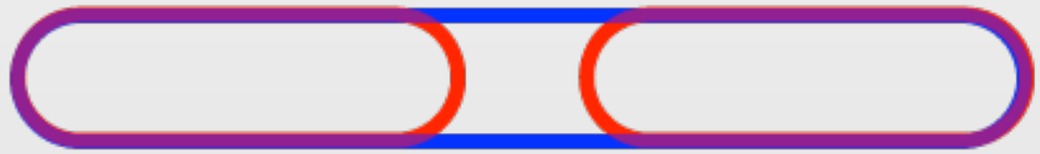
Total Variation Distance:

$$\begin{aligned}\mathrm{TV}(Q_1, Q_2) &= \sup_A |Q_1(A) - Q_2(A)| \\ &\leq a \max\{\mathrm{vol}(M_1 \setminus M_2), \mathrm{vol}(M_2 \setminus M_1)\} \\ &\leq C_d a \tau^d\end{aligned}$$

It suffices to bound the total variation distance.

Total Variation Distance:

$$\begin{aligned}\mathrm{TV}(Q_1, Q_2) &= \sup_A |Q_1(A) - Q_2(A)| \\ &\leq a \max\{\mathrm{vol}(M_1 \setminus M_2), \mathrm{vol}(M_2 \setminus M_1)\} \\ &\leq C_d a \tau^d\end{aligned}$$



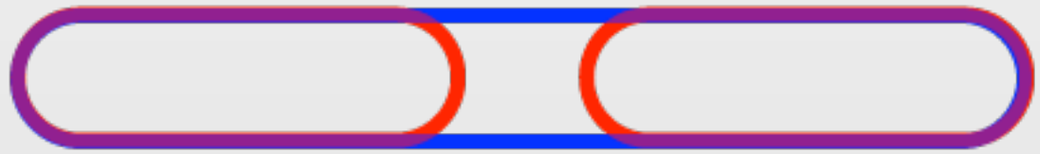
Minimax Risk:

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

It suffices to bound the total variation distance.

Total Variation Distance:

$$\begin{aligned}\mathrm{TV}(Q_1, Q_2) &= \sup_A |Q_1(A) - Q_2(A)| \\ &\leq a \max\{\mathrm{vol}(M_1 \setminus M_2), \mathrm{vol}(M_2 \setminus M_1)\} \\ &\leq C_d a \tau^d\end{aligned}$$



Minimax Risk:

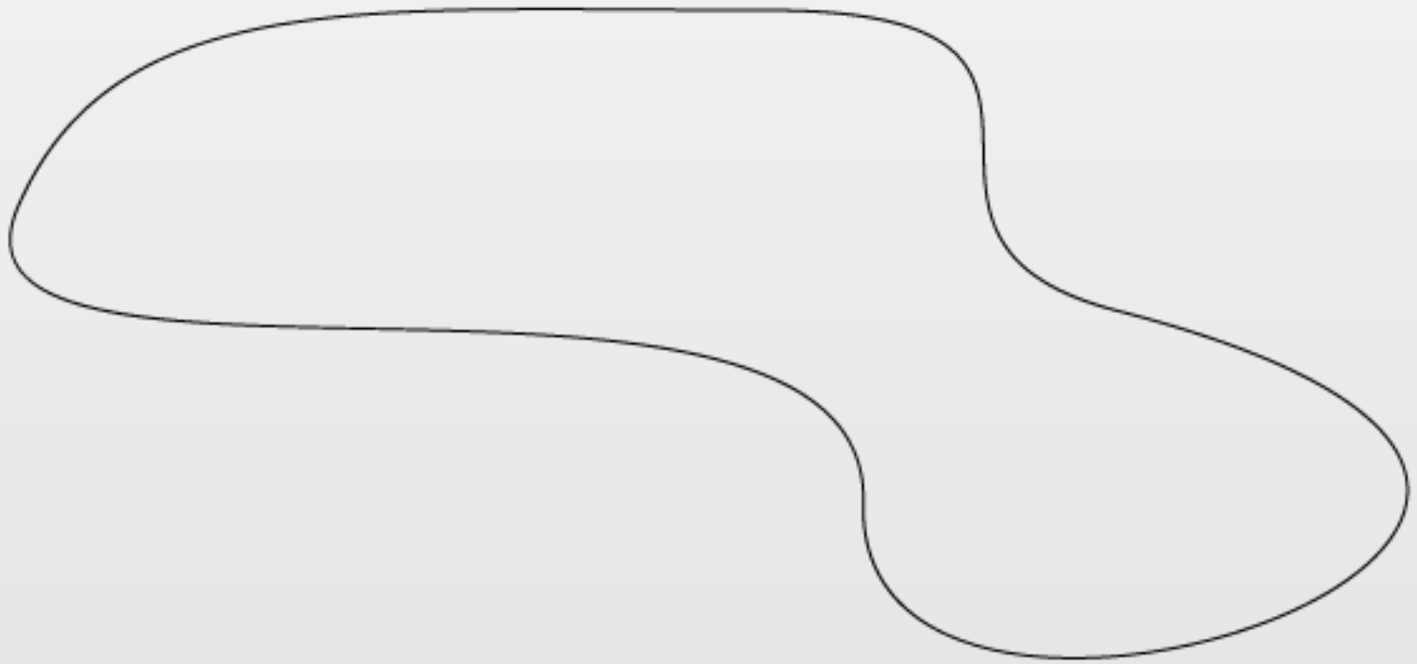
$$R_n \geq \frac{1}{8}(1 - \mathrm{TV}(Q_1, Q_2)^{2n} \geq \frac{1}{8}(1 - C_d a \tau^d)^{2n} \geq \frac{1}{8} e^{-2 C_d a \tau^d n}$$

Sampling Rate:

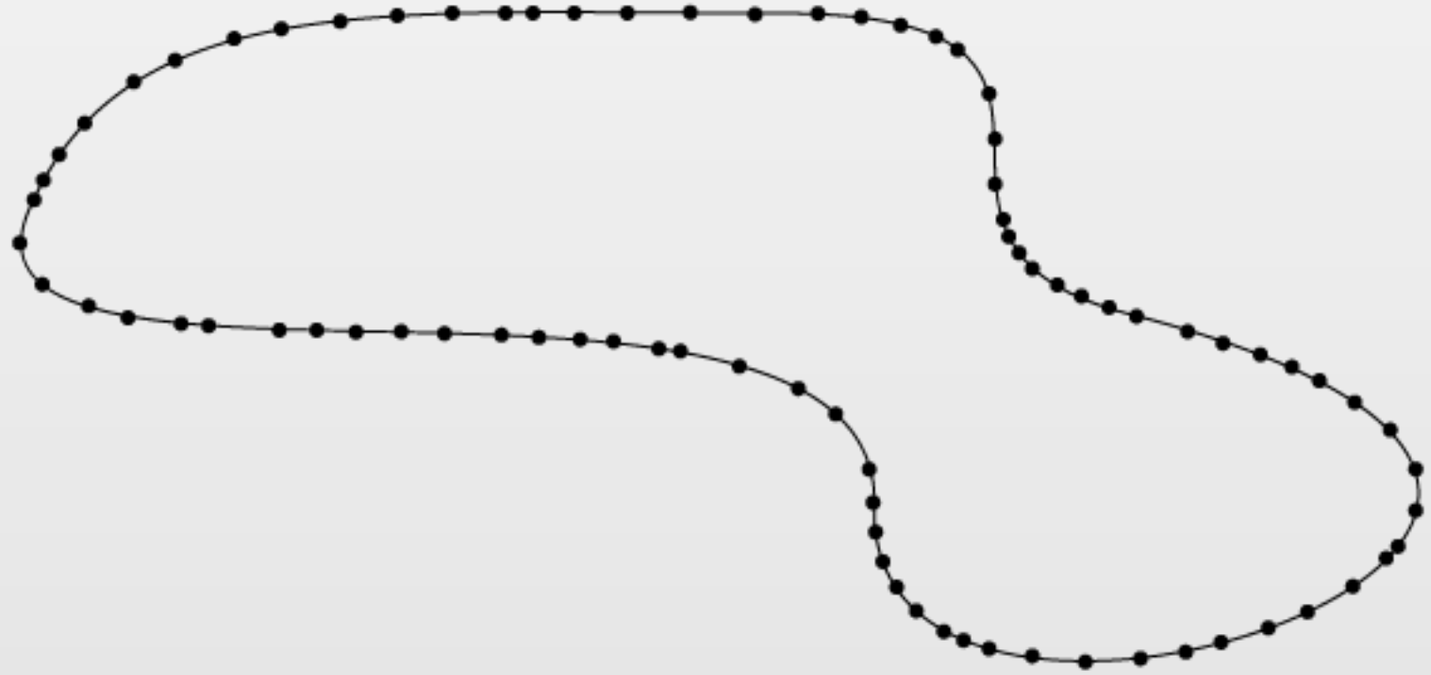
$$n(\epsilon) \geq \left(\frac{1}{\tau}\right)^d \log \frac{1}{\varepsilon}$$

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

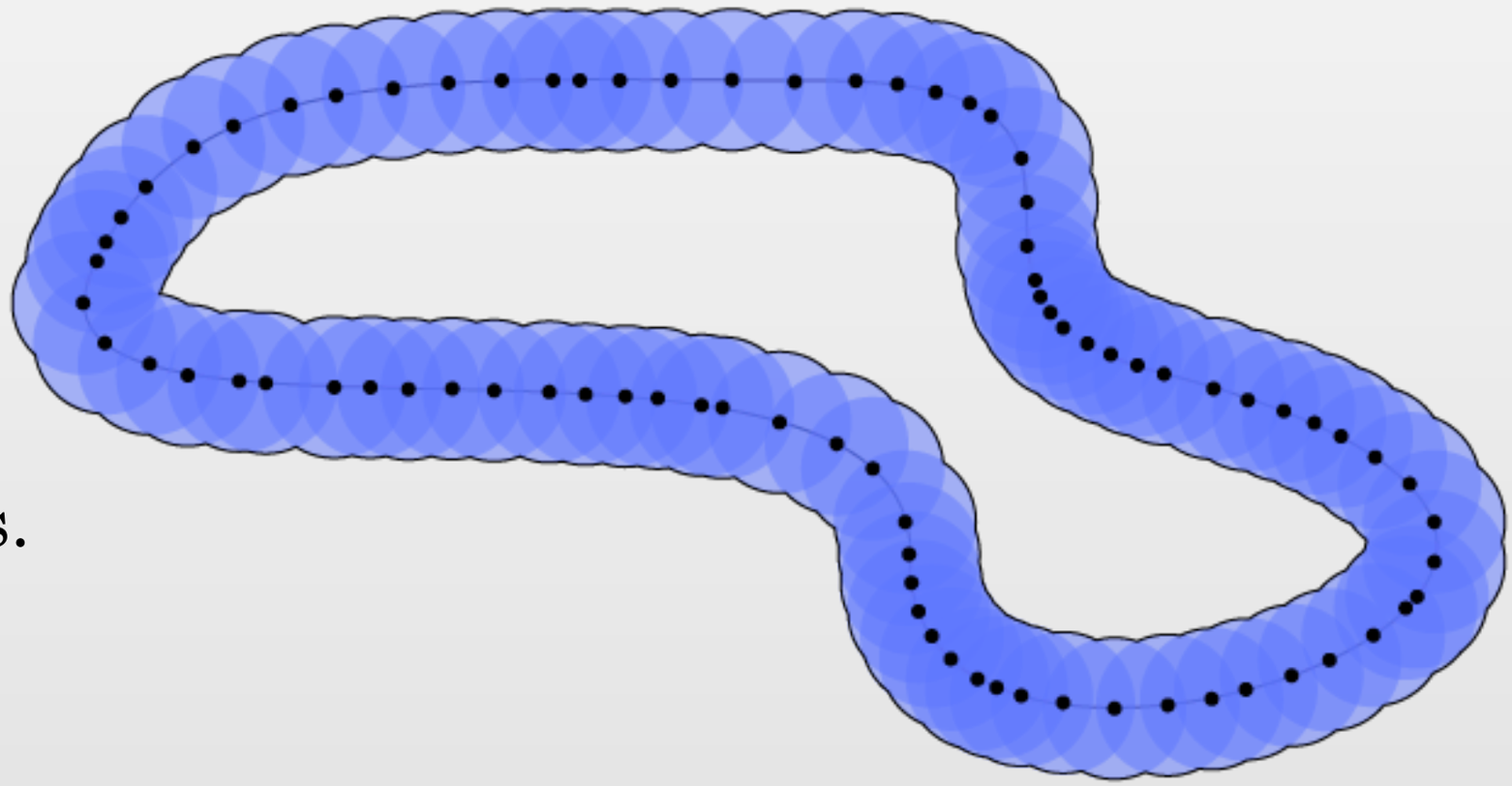
The upper bound uses a union of balls to estimate the homology of M.



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

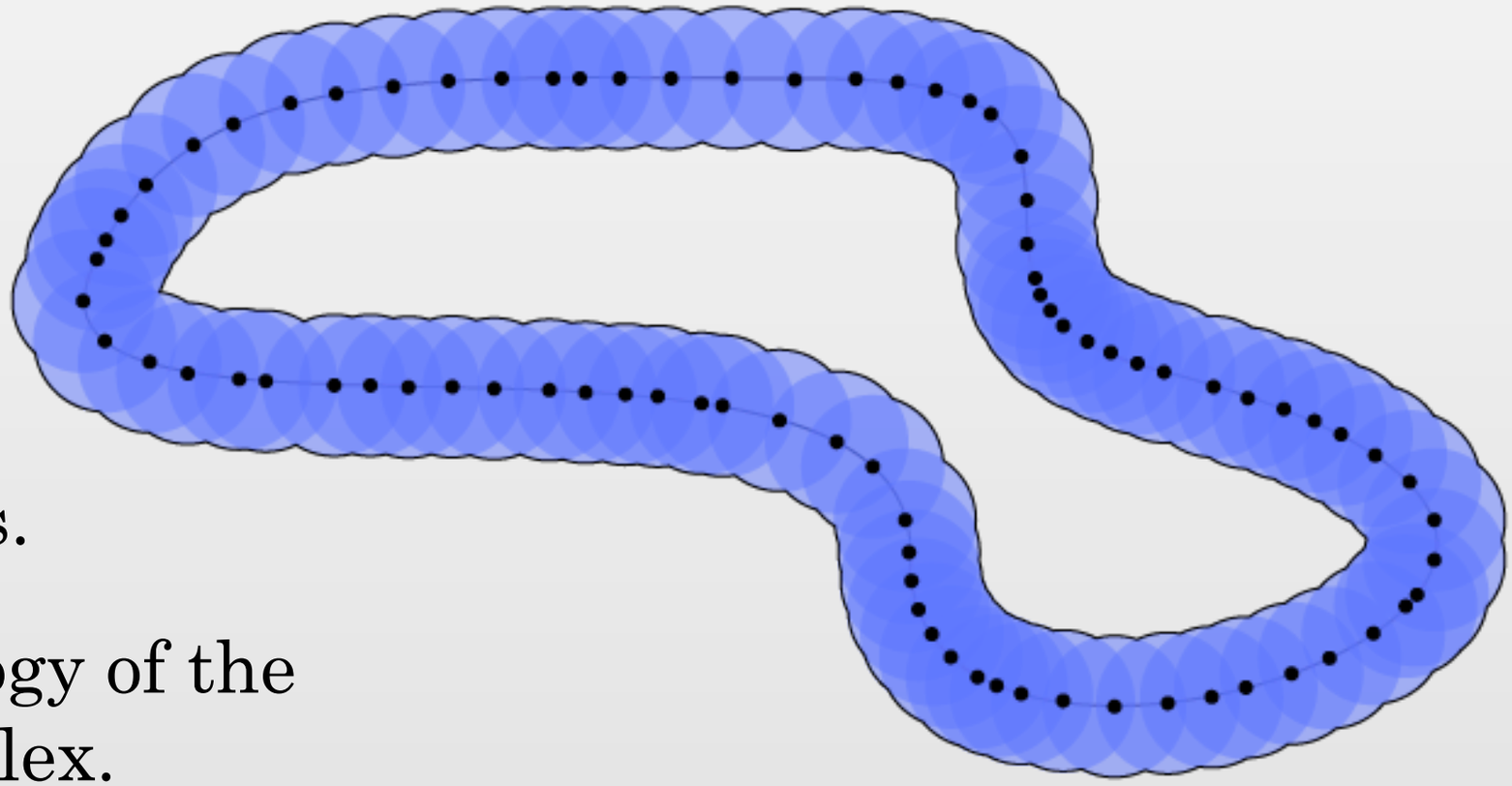


The upper bound uses a union of balls to estimate the homology of M.



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 Cech 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 Cech complex.

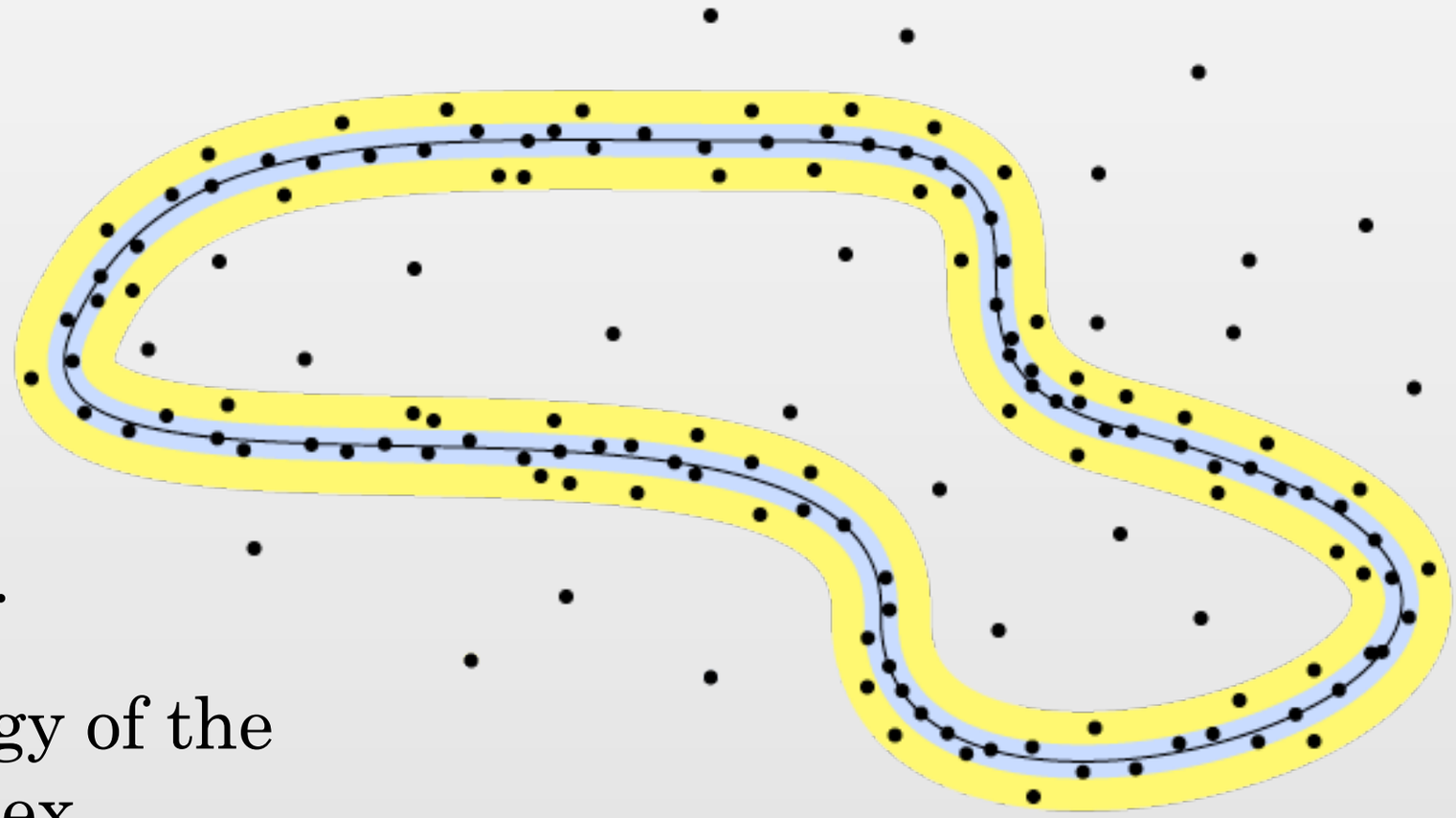
The upper bound uses a union of balls to estimate the homology of M.

0 Denoise the data.

1 Take a union of balls.

2 Compute the homology of the resulting Cech complex.

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

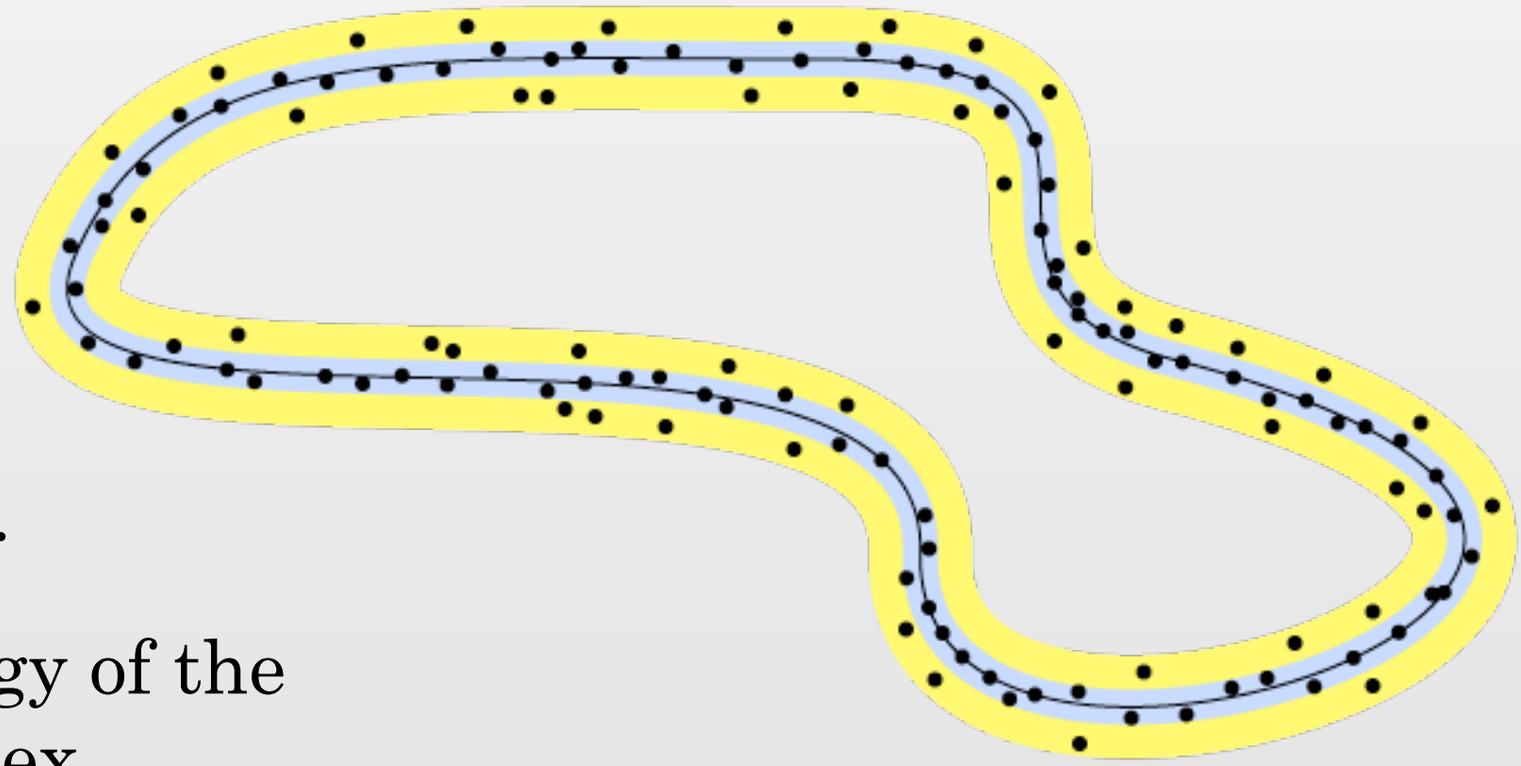


0 Denoise the data.

1 Take a union of balls.

2 Compute the homology of the resulting Cech complex.

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

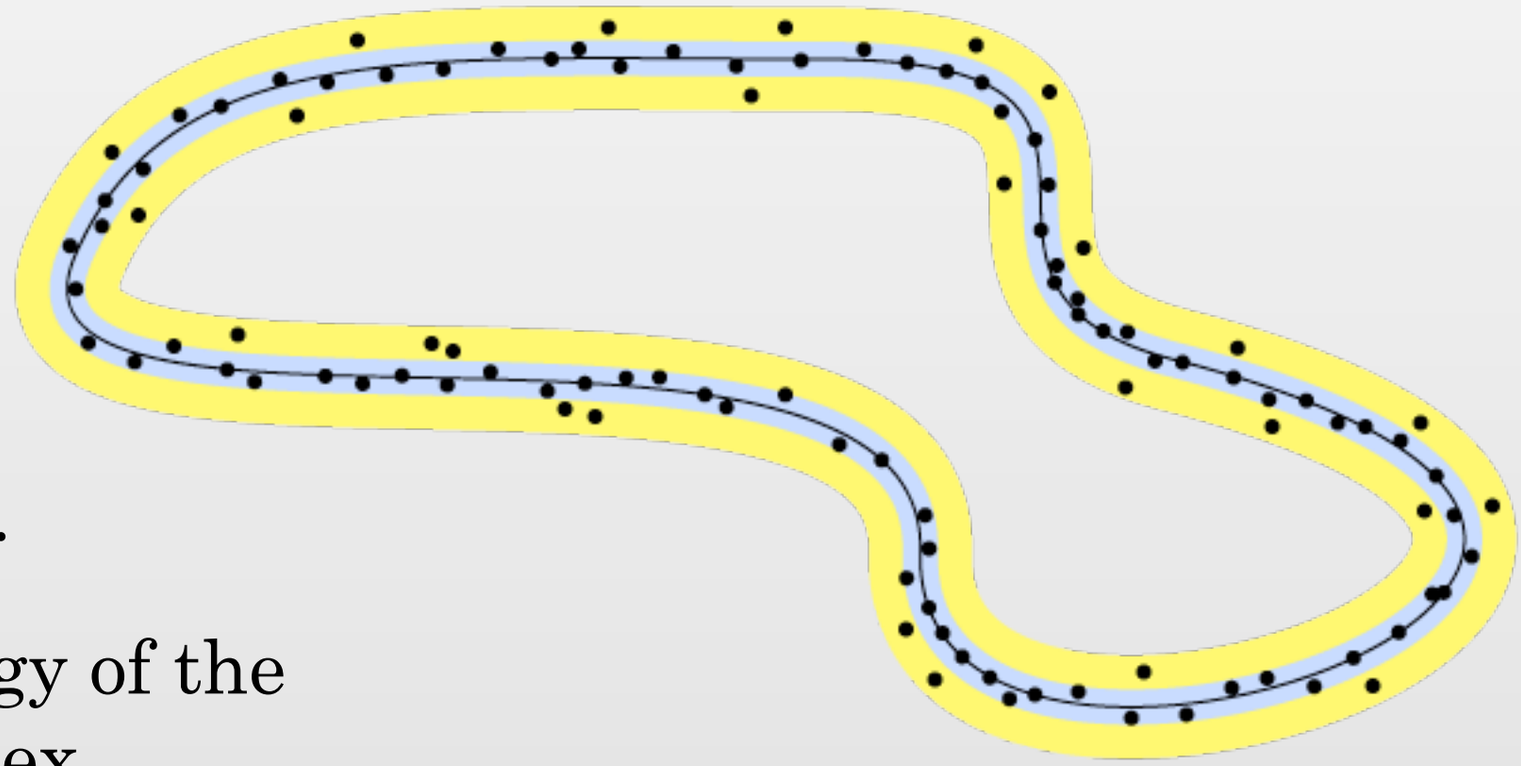


0 Denoise the data.

1 Take a union of balls.

2 Compute the homology of the resulting Cech 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.

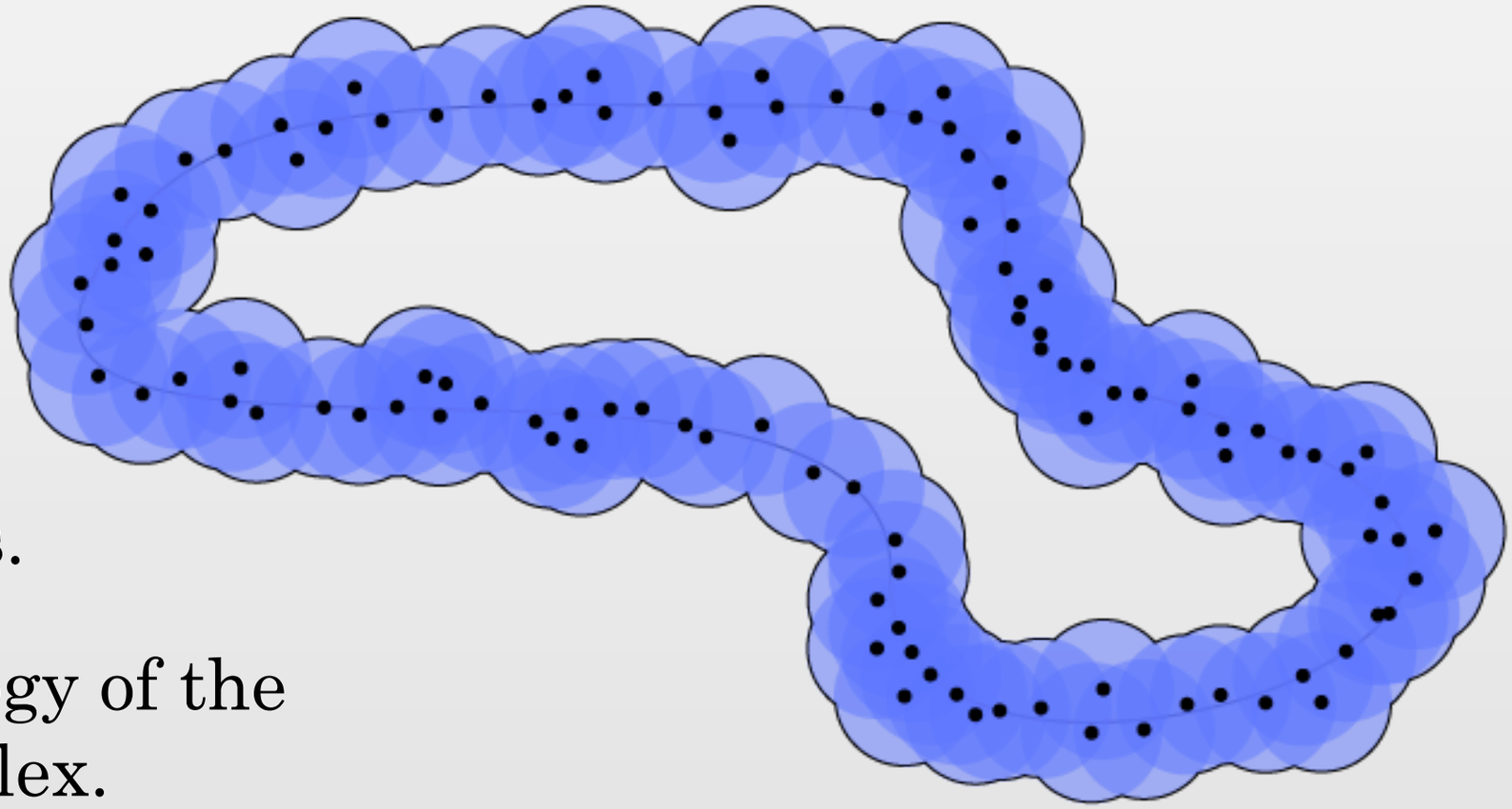
2 Compute the homology of the resulting Cech complex.

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

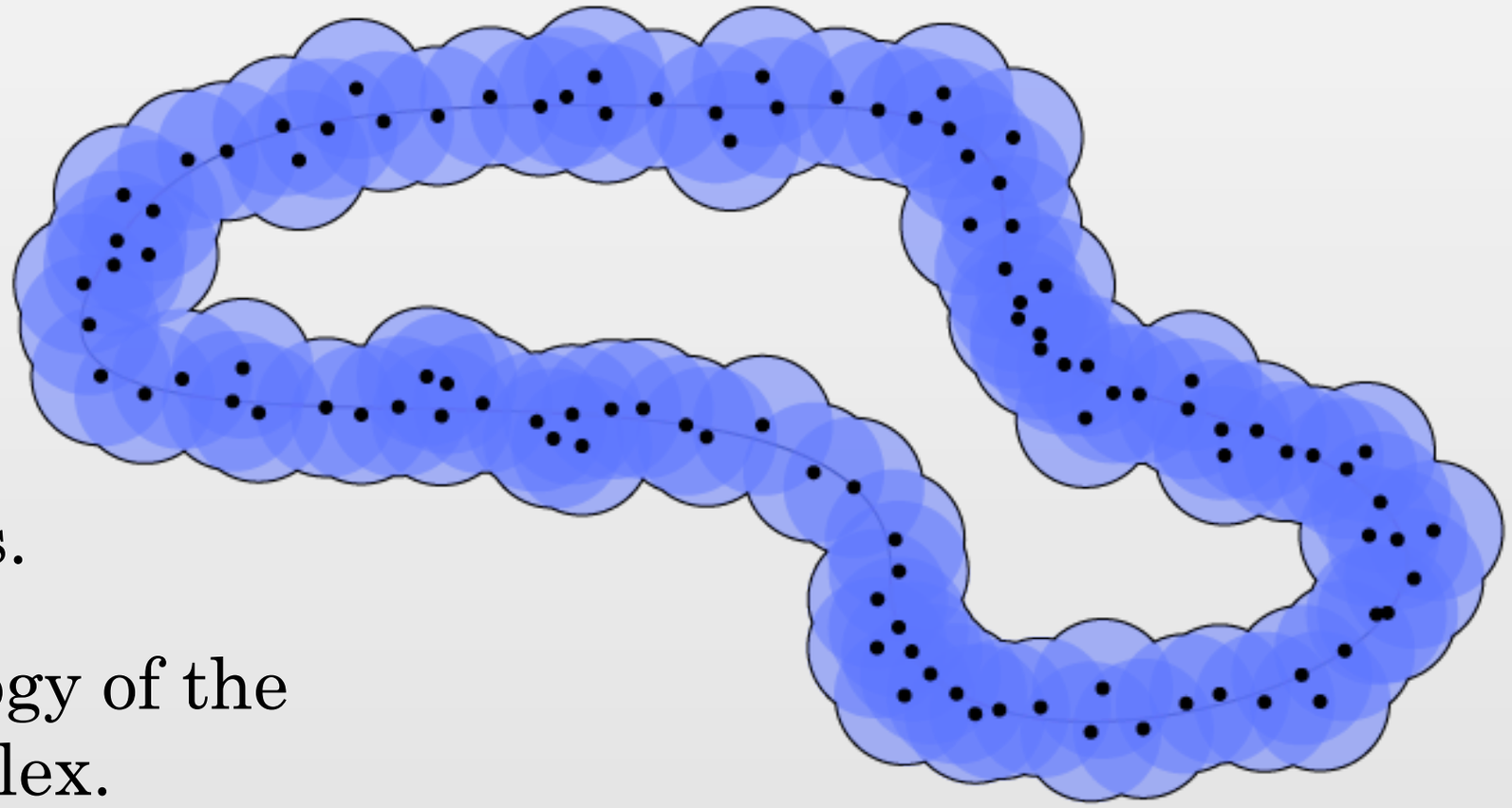
0 Denoise the data.

1 Take a union of balls.

2 Compute the homology of the resulting Cech 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 Cech complex.

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

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.