Statistical Topological Data Analysis (TDA)

Lecture 2 - part 2

Wolfgang Polonik

Department of Statistics, UC Davis

IHP, Sept. 15, 2022 - Dec. 1, 2022

Estimation of homotopy type

Some background:

Definition (medial axis)

Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^d$. The medial axis *of $\mathcal{M}$, denoted by $\mathrm{ax}(\mathcal{M})$, is defined as the closure of the set of points in $\mathbb{R}^d$ that have more than one closest point on $\mathcal{M}$, meaning that*

$$\mathrm{ax}(\mathcal{M}) = \overline{\left\{x \in \mathbb{R}^d : \arg\min_{m \in \mathcal{M}} \|x - m\| \text{ is not unique}\right\}}.$$

Estimation of homotopy type

Some background:

Definition (medial axis)

Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^d$. The medial axis *of $\mathcal{M}$, denoted by $\mathrm{ax}(\mathcal{M})$, is defined as the closure of the set of points in $\mathbb{R}^d$ that have more than one closest point on $\mathcal{M}$, meaning that*

$$\mathrm{ax}(\mathcal{M}) = \overline{\left\{x \in \mathbb{R}^d : \arg\min_{m \in \mathcal{M}} \|x - m\| \text{ is not unique}\right\}}.$$

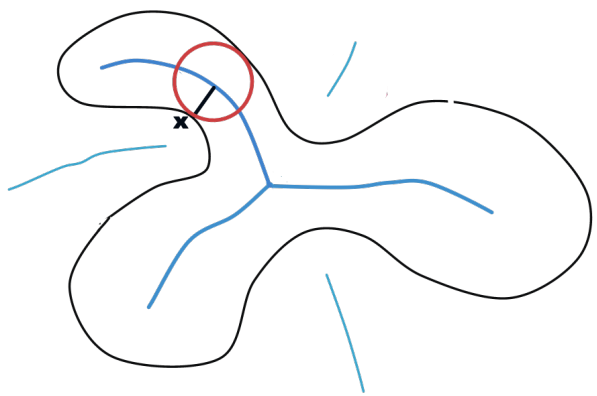

Figure: A submanifold $\mathcal{M}$ (black) with its medial axis $\mathrm{ax}(\mathcal{M})$(blue); a point x with the radius of the ball (red) being the local feature size at x; the reach of the manifold is the smallest local feature size.

The reach

Definition (reach)

For a submanifold $\mathcal{M} \subset \mathbb{R}^d$, the reach of $\mathcal{M}$ is the smallest distance from $\mathrm{ax}(\mathcal{M})$ to $\mathcal{M}$, i.e.

$$\mathrm{reach}(\mathcal{M}) = \inf_{a \in \mathrm{ax}(\mathcal{M})} d(a, \mathcal{M}).$$

In other words, consider $x \in \mathbb{R}^d$ lying in the r-enlargement of $\mathcal{M}$, i.e $x \in \mathcal{M}^r = \bigcup_{y \in \mathcal{M}} B_r(y)$. Then, if $r < \mathrm{reach}(\mathcal{M})$ there is a unique closest point (projection) to x on $\mathcal{M}$.

Theorem by Niyogi et al.

Theorem

Let $\mathcal{M}$ be a k-dimensional compact Riemannian submanifold (without boundary) of $\mathbb{R}^d$ with reach r. For $X_1, \ldots, X_n$ an i.i.d. sample drawn from the uniform distribution on $\mathcal{M}$, let $U_n(\epsilon) = \bigcup_{i=1}^n B_\epsilon(X_i)$, where $0 < \epsilon < r/2$. Then, with probability at least $1 - \delta$, for all

$$n > \beta_1 \Big(\log(\beta_2) + \log 1/\delta \Big),$$

$U_n(\epsilon)$ deformation retracts to $\mathcal{M}$ (and thus the homology of $U_n(\epsilon)$ equals the homology of $\mathcal{M}$ - see below). Here,

$$\beta_1 = \frac{\mathrm{vol}(\mathcal{M})}{\cos^k(\theta_1)\,(\frac{\epsilon}{4})^k V_k} \qquad \text{and} \qquad \beta_2 = \frac{\mathrm{vol}(\mathcal{M})}{\cos^k(\theta_2)\,(\frac{\epsilon}{8})^k V_k}$$

with $\theta_1 = \arcsin\left(\frac{r\epsilon}{8}\right)$, $\theta_2 = \arcsin\left(\frac{r\epsilon}{16}\right)$, and V_k the volume of the unit ball in $\mathbb{R}^k$.

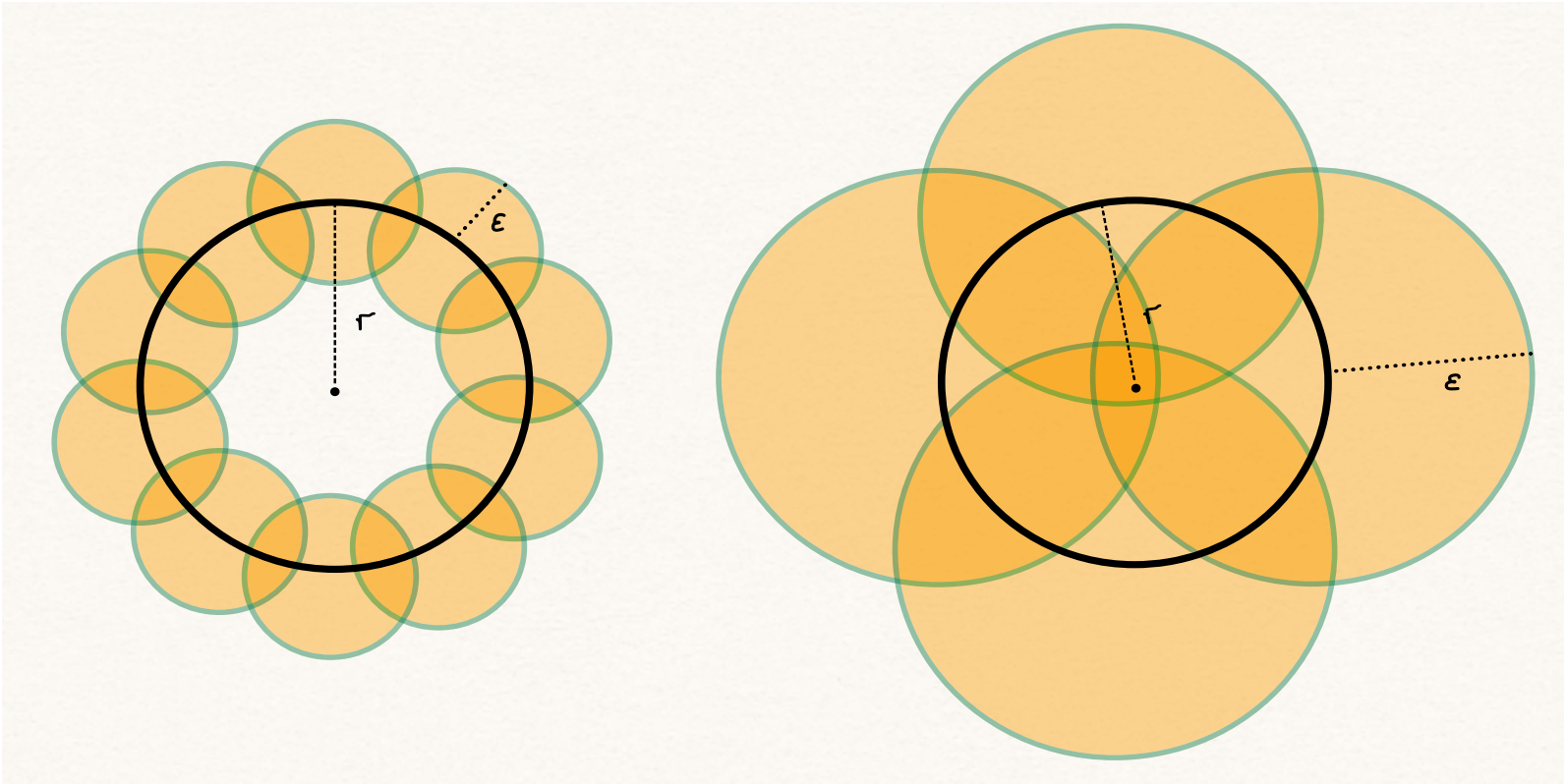

Illustration for the assumption that ϵ needs to be smaller than the reach; here the reach r equals the radius of the circle; left: here ϵ is small enough, and circle is deformation retract of union of balls; right: ϵ too large: union of balls is contractible

Two steps used in proof

Step (i): Show that, under the conditions of the theorem, with probability at least $1 - \delta$, the sample $X_1, \ldots, X_n$ is so dense in $\mathcal{M}$ that

$$\mathcal{M} \subset \bigcup_{i=1}^{n} B_{\epsilon/2}(X_i),$$

where the balls $B_{\epsilon/2}(X_i)$ are balls in the ambient space $\mathbb{R}^d$.

Two steps used in proof

Step (i): Show that, under the conditions of the theorem, with probability at least $1 - \delta$, the sample $X_1, \ldots, X_n$ is so dense in $\mathcal{M}$ that

$$\mathcal{M} \subset \bigcup_{i=1}^{n} B_{\epsilon/2}(X_i),$$

where the balls $B_{\epsilon/2}(X_i)$ are balls in the ambient space $\mathbb{R}^d$.

Step (ii) If $\mathcal{M} \subset \bigcup_{i=1}^{n} B_{\epsilon/2}(x_i)$ for some $\{x_1, \ldots, x_n\}$, and $\epsilon < \sqrt{\frac{3r}{5}}$, then $\bigcup_{i=1}^{n} B_{\epsilon}(x_i)$ deformation retracts to $\mathcal{M}$.

We will focus on Step (i).

Hausdorff distance

Definition (Hausdorff distance)

Let A, B be two compact sets in a metric space $(\mathbb{X}, d)$. Then

$$d_H(A, B) = \max\left(\sup_{a \in A} d_B(a), \sup_{b \in B} d_A(b)\right),$$

where $d_A(b)$ and $d_B(a)$ denote the distance functions to A and B, respectively.

Hausdorff distance

Definition (Hausdorff distance)

Let A, B be two compact sets in a metric space $(\mathbb{X}, d)$. Then

$$d_H(A, B) = \max\left(\sup_{a \in A} d_B(a), \sup_{b \in B} d_A(b)\right),$$

where $d_A(b)$ and $d_B(a)$ denote the distance functions to A and B, respectively.

Observation: If $A \subset B$, then

$$d_H(A, B) = \sup_{a \in B} d_A(a).$$

Hausdorff distance

Definition (Hausdorff distance)

Let A, B be two compact sets in a metric space $(\mathbb{X}, d)$. Then

$$d_H(A, B) = \max\left(\sup_{a \in A} d_B(a), \sup_{b \in B} d_A(b)\right),$$

where $d_A(b)$ and $d_B(a)$ denote the distance functions to A and B, respectively.

Observation: If $A \subset B$, then

$$d_H(A, B) = \sup_{a \in B} d_A(a).$$

This means that

$$\mathcal{M} \subset \bigcup_{i=1}^{n} B_{\epsilon/2}(X_i) \quad \Leftrightarrow \quad d_H(\mathbb{X}_n, \mathcal{M}) < \epsilon/2, \tag{1}$$

So in Step (i), we have to control the Hausdorff distance.

ϵ-dense sets

Definition (ϵ-dense sets)

Let $(\mathbb{X}, d_{\mathbb{X}})$ *be a metric space. Given a subspace* $A \subset \mathbb{X}$ *and* $\epsilon > 0$*, we call* $D = \{x_1, \ldots, x_k\}$ *an* ϵ*-dense set in* A *if* $\sup_{x \in A} d_D(x) < \epsilon$.

ϵ-dense sets

Definition (ϵ-dense sets)

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space. Given a subspace $A \subset \mathbb{X}$ and $\epsilon > 0$, we call $D = \{x_1, \ldots, x_k\}$ an ϵ-dense set in A if $\sup_{x \in A} d_D(x) < \epsilon$.

A set A being ϵ-dense in $\mathbb{X}$ is equivalent to

- $A \subset \bigcup_{i=1}^{k} B_{\epsilon/2}(x_i)$,

and to

- $d_H(A, \mathbb{X}) \leq \epsilon/2$

In other words, we have to study when our sample is $\epsilon/2$-dense.

Step (i)

This next result essentially gives Step (i). It is a slightly simpler version, not using θ_1 and θ_2.

Proposition

Let P be a probability measure with support $\mathcal{M} \subset \mathbb{R}^d$, a k-dim. compact Riemannian submanifold of $\mathbb{R}^d$ with reach $r > 0$. Suppose that P has a density f with respect to Hausdorff measure on $\mathcal{M}$. Let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ with $X_i, i = 1, \ldots, n$ iid with density f, such that $0 < m < f$ for some $m < 1$. Then, for $t \leq r/2$, and with V_k the k-dimensional volume of the unit ball in $\mathbb{R}^k$,

$$P\Big(d_H(\mathcal{M}, \mathbb{X}_n) > t\Big) \leq \frac{4^k}{m\, t^k} \exp\Big\{ - n \frac{V_k\, m\, t^k}{2^k} \Big\}.$$

In particular, ϵ-dense samples exist with probability $1 - \delta$ for

$$n \geq \frac{2^k}{V_k\, m\, \epsilon^k} \left(\log\left(\frac{4^k}{V_k\, m \epsilon^k} \right) + \log\left(\frac{1}{\delta} \right) \right).$$

Also, there exists a constant C such that

$$\limsup_{n \to \infty} \left(\frac{\log n}{n} \right)^{1/k} d_H(\mathcal{M}, \mathbb{X}_n) \leq C \qquad \textit{almost surely.}$$

(a, b)-standard assumption

Definition ((a, b)-standard assumption)

A probability measure P with compact support $A \subset \mathbb{X}$, where $\mathbb{X}$ is a metric space, satisfies the (a, b)-standard assumption, if there exist constants $a, b > 0$, such that

$$P(B_\epsilon(x)) \geq \min(1, a\epsilon^b) \qquad \text{for all } x \in A \text{ and for all } \epsilon > 0.$$

Note that here $B_\epsilon(x)$ is an open ball in $\mathbb{X}$, and since P has support A we have $P(B_\epsilon(x)) = P(B_\epsilon(x) \cap A)$.

A key result

Proposition

Let $\mathbb{X}$ be a metric space, $\mathcal{M} \subset \mathbb{X}$ compact, and P a probability measure with support $\mathcal{M}$ satisfying the so-called (a, b)-standard assumption. *If $\mathbb{X}_n$ is an iid sample from P, then*

$$P\Big(d_H(\mathcal{M}, \mathbb{X}_n) > 2\epsilon\Big) \leq \frac{2^b}{a\,\epsilon^b} \exp\Big\{ - na\epsilon^b \Big\}. \tag{2}$$

In particular, ϵ-dense random samples exist with probability $1 - \delta$, with $\delta > 0$, for

$$n \geq \frac{2^b}{a\epsilon^b}\left(\log\left(\frac{4^b}{a\epsilon^b}\right) + \log\left(\frac{1}{\delta}\right) \right). \tag{3}$$

A key result

Proposition

Let $\mathbb{X}$ be a metric space, $\mathcal{M} \subset \mathbb{X}$ compact, and P a probability measure with support $\mathcal{M}$ satisfying the so-called (a, b)-standard assumption. *If $\mathbb{X}_n$ is an iid sample from P, then*

$$P\Big(d_H(\mathcal{M}, \mathbb{X}_n) > 2\epsilon\Big) \leq \frac{2^b}{a\,\epsilon^b} \exp\Big\{ - na\epsilon^b \Big\}. \tag{2}$$

In particular, ϵ-dense random samples exist with probability $1 - \delta$, with $\delta > 0$, for

$$n \geq \frac{2^b}{a\epsilon^b}\left(\log\left(\frac{4^b}{a\epsilon^b}\right) + \log\left(\frac{1}{\delta}\right) \right). \tag{3}$$

Comment: For a probability measure P on a k-dimensional compact manifold $\mathcal{M}$ in $\mathbb{R}^d$ with density f bounded away from zero, Niyogi et al. showed that the assumption of a positive reach implies that P is standard with $b = k$ and with a depending on the reach.

Tools for the proof.

Definition (covering numbers)

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space. The minimal cardinality of ϵ-dense sets in A is called ϵ-covering number of A and is denoted by $N(\epsilon, A, d_{\mathbb{X}})$, i.e.

$$N(\epsilon, A, d_{\mathbb{X}}) = \min \left\{ k \in \mathbb{N} : \exists \{x_1, \ldots, x_k\} \subset A \text{ with } A \subset \bigcup_{i=1}^{k} B_\epsilon(x_i) \right\}.$$

Tools for the proof.

Definition (covering numbers)

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space. The minimal cardinality of ϵ-dense sets in A is called ϵ-covering number of A and is denoted by $N(\epsilon, A, d_{\mathbb{X}})$, i.e.

$$N(\epsilon, A, d_{\mathbb{X}}) = \min\left\{k \in \mathbb{N} : \exists\, \{x_1, \ldots, x_k\} \subset A \text{ with } A \subset \bigcup_{i=1}^{k} B_\epsilon(x_i)\right\}.$$

Definition (packing numbers)

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space and $A \subset \mathbb{X}$. The maximum cardinality of ϵ-packings of A, denoted by $N'(\epsilon, A, d_{\mathbb{X}})$, is called ϵ-packing number of A, i.e.

$$N'(\epsilon, A, d_{\mathbb{X}}) = \max\left\{k \in \mathbb{N} : \exists\, \{x_1, \ldots, x_k\} \subset A \text{ with } \; d(x_i, x_j) > \epsilon \;\text{ for all } i \neq j\right\}.$$

Tools for the proof.

Definition (covering numbers)

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space. The minimal cardinality of ϵ-dense sets in A is called ϵ-covering number of A and is denoted by $N(\epsilon, A, d_{\mathbb{X}})$, i.e.

$$N(\epsilon, A, d_{\mathbb{X}}) = \min\left\{k \in \mathbb{N} : \exists\, \{x_1, \ldots, x_k\} \subset A \text{ with } A \subset \bigcup_{i=1}^{k} B_\epsilon(x_i)\right\}.$$

Definition (packing numbers)

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space and $A \subset \mathbb{X}$. The maximum cardinality of ϵ-packings of A, denoted by $N'(\epsilon, A, d_{\mathbb{X}})$, is called ϵ-packing number of A, i.e.

$$N'(\epsilon, A, d_{\mathbb{X}}) = \max\left\{k \in \mathbb{N} : \exists\, \{x_1, \ldots, x_k\} \subset A \text{ with } \; d(x_i, x_j) > \epsilon \;\text{ for all } i \neq j\right\}.$$

Tools for the proof.

Lemma

For any $\epsilon > 0$

$$N'(2\epsilon, A, d_{\mathbb{X}}) \leq N(\epsilon, A, d_{\mathbb{X}}) \leq N'(\epsilon, A, d_{\mathbb{X}}).$$

PROOF: Use notation $d_{\mathbb{X}} = d$.

Consider a minimal 2ϵ-packing $P = \{x_1, \ldots, x_{N'}\}$, and a minimal ϵ-dense set $D = \{y_1, \ldots, y_N\}$. Suppose that for x_i, x_j $(i \neq j)$ there exists $y \in D$ such that y is the nearest neighbor in D to both x_i and x_j, then

$$d(x_i, x_j) \leq d(x_i, y) + d(x_j, y) < 2\epsilon,$$

which is a contradiction to the assumption of P being a 2ϵ-packing. Thus, there can be at most one y_j in each $B_\epsilon(x_i)$, implying $N'(2\epsilon, A) \leq N(\epsilon, A)$.

The second inequality follows by observing that an ϵ-packing $P = \{x_1, \ldots, x_{N'}\}$ is also ϵ-dense, because if it were not, i.e. if $\bigcup_{i=1}^{N'} B_\epsilon(x_i) \not\subset A$, then at least one more point x could be added to P with $d(x, x_i) > \epsilon$, contradicting the assumption of P being an ϵ-packing.

Comments

It is their behavior for small ϵ that is usually important. The quantity

$$\log N(\epsilon, A)$$

is often called (metric) *entropy* of A. The behavior of the entropy is closely related to the dimension of A. Indeed, for a metric space A, the limit

$$\dim(A) = \lim_{\epsilon \to 0} \frac{\log N(\epsilon, A)}{\log \frac{1}{\epsilon}}$$

is called *Minkowski dimension* or *Box-counting dimension* of A.

If A is a k-dimensional smooth manifold, then $\dim(A) = k$. The box-counting dimension can also be used to define fractal dimensions.

Packing number bound

Lemma (packing number bound)

Let $\mathbb{X}$ be a metric space and μ a measure satisfying the (a, b)-standard assumption. Then for any finite measure μ,

$$N'(\epsilon, \mathbb{X}, \mu) \leq \frac{\mu(\mathbb{X})2^b}{a\epsilon^b}.$$

PROOF. Consider a maximal ϵ-packing $\{s'_1, \ldots, s'_{N'}\}$ of $\mathbb{X}$ of size $N' = N'(\epsilon, \mathbb{X}, \mu)$. By definition of a packing, $\bigcup_{i=1}^{N'} B_{\epsilon/2}(s'_i) \subset \mathbb{X}$, and the balls $B_{\epsilon/2}(s'_i)$ are disjoint. Thus,

$$\mu(\mathbb{X}) \geq \mu\Big(\bigcup_{i=1}^{N'} B_{\epsilon/2}(x_i)\Big) = \sum_{i=1}^{N'} \mu(B_{\epsilon/2}(x_i)) \geq N' a\Big(\frac{\epsilon}{2}\Big)^b,$$

which implies the desired inequality.

Proof of key result.

Let $S_n = \{s_1, \ldots, s_N\}$ be a *minimal* ϵ-dense set in $\mathbb{X}$. Then

$$d_H(\mathbb{X}, \mathbb{X}_n) \leq d_H(\mathbb{X}, S_N) + d_H(S_N, \mathbb{X}_n) < \epsilon + d_H(S_N, \mathbb{X}_n),$$

where the first inequality is triangle inequality for the Hausdorff distance. The second follows because on the one hand, $\sup_{x \in \mathbb{X}} d_{S_N}(x) < \epsilon$ by definition of an ϵ-dense set, and on the other hand, $\sup_{x \in S_N} d_{\mathbb{X}}(x) = 0$ (because $S_N \subset \mathbb{X}$). Consequently, we have $d_H(\mathbb{X}, S_N) < \epsilon$. Using the above, we obtain

$$P\big[d_H(\mathbb{X}, \mathbb{X}_n) > 2\epsilon\big] \leq P\big[\epsilon + d_H(S_N, \mathbb{X}_n) > 2\epsilon\big] = P\big[d_H(S_N, \mathbb{X}_n) > \epsilon\big],$$

and we will bound the probability on the right.

Proof of key result.

For this, recall that S_n is ϵ-dense in $\mathbb{X}$ (and $\mathbb{X}_n \subset \mathbb{X}$). In particular, for each $X_i \in \mathbb{X}_n$, $d_{S_N}(X_i) < \epsilon$. Thus, $d_H(S_N, \mathbb{X}_n) > \epsilon$ is equivalent to $d_{\mathbb{X}_n}(s_i) > \epsilon$ for some $s_i \in S_N$, meaning that $B_\epsilon(s_i) \cap \mathbb{X}_n = \emptyset$ for some i. We obtain

$$
\begin{aligned}
P\big(d_H(S_N, \mathbb{X}_n) > \epsilon\big) &= P\big(B_\epsilon(s_i) \cap \mathbb{X}_n = \emptyset \text{ for some } i\big) \\
&\leq \sum_{i=1}^{N} P\big(B_\epsilon(s_i) \cap \mathbb{X}_n = \emptyset\big) && \text{(union bound)} \\
&= \sum_{i=1}^{N} \prod_{j=1}^{n} P\big(X_j \notin B_\epsilon(s_i)\big) && \text{(independence)} \\
&= \sum_{i=1}^{N} \big(1 - P\big(B_\epsilon(s_i)\big)\big)^n && \text{(identical distribution)} \\
&\leq \sum_{i=1}^{N} \exp\big\{ - nP\big(B_\epsilon(s_i)\big)\big\} && \text{(using } 1 - x \leq e^{-x}\text{)} \\
&\leq N \exp\{-na\epsilon^b\} && (P \text{ is } (a, b)\text{-standard}) \\
&\leq \frac{2^b}{a\epsilon^b} \exp\{-na\epsilon^b\}. \\
& && \text{(use } N \leq N' \text{ and packing number bound)}
\end{aligned}
$$

The (a, b)-standard assumptions for manifolds

Lemma

If $\mathbb{X} = \mathcal{M}$ is a compact Riemannian manifold with reach r, then, for $\epsilon < r/2$, the (a, b)-standard assumption holds for the Hausdorff measure with $a = \left(\frac{1}{2}\right)^k V_k$ and $b = k$. If P is a probability measure on $\mathcal{M}$ with density f bounded below by m, then the (a, b)-standard assumption holds with $a = \left(\frac{1}{2}\right)^k m V_k$ and $b = k$.

Indeed, Niyogi et al. (2008) show a slightly more refined result, which gives some dependence on the geometry of the manifold. It is shown that

$$\mathrm{vol}\big(\mathcal{M} \cap B_\epsilon(x)\big) \geq \big(\cos(\theta_\epsilon)\big)^k V_k \, \epsilon^k,$$

where $\theta_\epsilon = \arcsin\left(\frac{\epsilon}{2r}\right)$. Since $\epsilon < r/2$, we obtain that $\theta_\epsilon < \arcsin\left(\frac{1}{4}\right)$ and $\cos\left(\arcsin\left(\frac{1}{4}\right)\right) \approx 0.96 > \frac{1}{2}$.
Combining the result on ϵ-dense sets with the last result on the (a, b)-standard assumptions for distributions on compact Riemannian manifolds immediately gives the proof of the Proposition of step (1).