A Appendix – Supplementary Material

A.1 Key technical lemmas from [17]

We will need two technical lemmas, which follow from [17].

Lemma 6 (**Ball volume lemma**, Lemma 5.3 in [17])**.** *Let $p \in M$. Now consider $A = M \cap B_\epsilon(p)$. Then $vol(A) \geq (\cos(\theta))^d vol(B^d_\epsilon(p))$ where $B^d_\epsilon(p)$ is the a d-dimensional ball in the tangent space at p, $\theta = \sin^{-1}\frac{\epsilon}{2\tau}$.*

Next, consider a collection of balls $\{B_r(p_i)\}_{i=1,\ldots,n}$ centered around points p_i on the manifold and such that $M \subset \cup_{i=1}^l B_r(p_i)$.

Lemma 7 (**Sampling lemma**, Lemma 5.1 in [17])**.** *Let $A_i = B_r(p_i)$ be a collection of sets such that $\cup_{i=1}^l A_i$ forms a minimal cover of M. If $Q(A_i) \geq \alpha$, and*

$$n > \frac{1}{\alpha}\left(\log l + \log\left(\frac{2}{\delta}\right)\right)$$

then w.p. at least $1 - \delta/2$, each A_i contains at least one sample point, and $M \subset \cup_{i=1}^n B_{2r}(x_i)$. Further we have that $l \leq \frac{\mathrm{vol}(M)}{\cos^d(\theta) v_d r^d}$.

A.1.1 Proofs for the noiseless case

Lower bound Here we describe the densities on the two manifolds M_1 and M_2. There are two sets of interest to us: $W_1 = M_1 \setminus M_2$ which corresponds to the two "holes" of radius 4τ in the annulus, and $W_2 = M_2 \setminus M_1$ which corresponds to the d-dimensional piece added to smoothly join the inner pieces of the two annuli in M_2.

By construction, $\mathrm{vol}(W_1) = 2v_d(4\tau)^d$ where v_d is the volume of the unit d-ball. $\mathrm{vol}(W_2)$ is somewhat tricky to calculate exactly due to the curvature of W_2 but it is easy to see that $\mathrm{vol}(W_2)$ is also $O(\tau^d)$ with the constant depending on d.

One of the densities is constructed in the following way, on the set of larger volume (between W_1 and W_2) we set $p(x) = a$, and evenly distribute the rest of the mass over the remaining portion of the manifold (we are guaranteed that the mass on the rest of the manifold is at least a since otherwise the constraint $p(x) \geq a$ can never be satisfied).

The other density is constructed to be equal (to the first density) outside the set on which the two manifolds differ. The remaining mass is spread evenly on the set where they do differ. We are again guaranteed that $p(x) \geq a$ by construction.

Let us now calculate the TV between these two densities. This is just the integral of the difference of the densities over the set where one of the densities is larger. Since the two densities are equal outside $W_1 \cup W_2$ and disjoint over $W_1 \cup W_2$ it is clear that

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

with the constant depending on d. The lower bound follows from the calculations in the main paper.

Upper bound The NSW lemma tells us that for $n > \zeta_1\left(\log(\zeta_2) + \log\left(\frac{1}{\delta}\right)\right)$, with $\zeta_1 = \frac{\mathsf{vol}(M)}{a\cos^d\theta_1 \mathsf{vol}(B^d_{\epsilon/4})}$, $\zeta_2 = \frac{\mathsf{vol}(M)}{\cos^d\theta_2 \mathsf{vol}(B^d_{\epsilon/8})}$, $\theta_1 = \sin^{-1}\frac{\epsilon}{8\tau}$ and $\theta_2 = \sin^{-1}\frac{\epsilon}{16\tau}$, we have $\mathbb{P}(\widehat{\mathcal{H}} \neq \mathcal{H}(M)) < \delta$.

By assumption, we have $\mathrm{vol}(M) \leq C$. We further take $\epsilon = \tau/2$. It is clear that in ζ_1 and ζ_2 all terms except the ball volumes are constant. This gives us that $\zeta_1 = C_1/(a\tau^d)$ and $\zeta_2 = C_2/(a\tau^d)$.

Now, the NSW lemma can be restated as if $n = C_1/\tau^d(\log(C_2/\tau^d) + \log(1/\delta))$ we recover the homology with probability at least $1 - \delta$. Notice that this means that the minimax risk $\leq \delta$.

A straightforward rearrangement of this gives us

$$R_n \leq C_2/(a\tau^d)\exp(-na\tau^d/C_1)$$

for appropriate C_1, C_2. To bound the resolution we rewrite this as

$$R_n \leq \exp\left(-\frac{na\tau^d}{C_1} + \log\left(\frac{C_2}{a\tau^d}\right)\right)$$

One can verify that if

$$\tau^d \leq C\frac{\log n \log(1/\epsilon)}{n}$$

for an appropriately large C, we have $R_n \leq \epsilon$ as desired.

A.1.2 Proofs for the clutter noise case

Lower bound This is a straightforward extension of the noiseless case. The densities are constructed in an identical manner. The contribution to the densities from the clutter noise is identical in each case. As in the analysis for the noiseless case we bound the total variation distance between the two densities. We have an additional factor of π which is the mixture weight of the component corresponding to the density on the manifold.

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

Given this bound the calculations are identical to those in the noiseless case.

Upper bound As a preliminary step we will need to clean the data to eliminate points that are far away

from the manifold. Our analysis will show that Algorithm 1 will achieve this, with high probability. We will 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.

Let $a = \inf_{x \in M} p(x)$, which is strictly positive by assumption. Define, $A = \mathsf{tube}_r(M)$ and $B = \mathbb{R}^D - \mathsf{tube}_{2r}(M)$ where $r < \frac{(\sqrt{9}-\sqrt{8})\tau}{2}$. Following [18], we define $\alpha_s = \inf_{t \in A} Q(B_s(t))$ and $\beta_s = \sup_{t \in B} Q(B_s(t))$ where $s = 2r$. Then $\alpha_s \geq \frac{v_D s^D (1-\pi)}{\mathrm{vol(Box)}} + \pi a v_d r^d \cos^d \theta = \alpha$ and $\beta_s \leq \frac{v_D s^D (1-\pi)}{\mathrm{vol(Box)}} = \beta$ where $\theta = \sin^{-1}(\frac{r}{2\tau})$. The second term of the bound on α_s follows in two steps: first observe that for any point x in A, $B_s(x) \supseteq B_r(t)$ where t is the closest point on M to x. Now, we use Lemma 6 to bound $Q(B_r(t))$.

We will now 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.

Define the events $\mathcal{E}_1 = \left\{ \{X_i : \ i \in I\} \supseteq \{X_i \in A\} \quad \text{and} \quad \{X_i : \ i \in I^c\} \supseteq \{X_i \in B\} \right\}$ and $\mathcal{E}_2 = \left\{ M \subset \bigcup_{i \in I} B_{2r}(X_i) \right\}$. We will show that $\mathcal{E}_1$ and $\mathcal{E}_2$ both hold with high probability.

For $\mathcal{E}_1$ to hold, we need β to be not too close to α, in particular $\beta < \alpha/2$ will suffice. This happens with probability 1, for τ small if $d < D$. By Lemma 13 in the Appendix, $\mathcal{E}_1$ happens with probability at least $1 - \delta/2$, provided that $n > 4\kappa \log \kappa$, where

$$\kappa = \max\left(1 + \frac{200}{3\pi a v_d r^d \cos^d(\theta)} \log\left(\frac{2}{\delta}\right), 4\right).$$

Now we turn to $\mathcal{E}_2$. Let $p_1, \ldots, p_N \in M$ be such that $B_r(p_1), \ldots, B_r(p_N)$ forms a minimal covering of M. From Lemma 7, we have that $N \leq \frac{\mathrm{vol}(M)}{\cos^d(\theta) v_d r^d}$. Let $A_j = B_r(p_j)$. Then

$$\begin{aligned} Q(A_j) &\geq \frac{v_D s^D (1-\pi)}{\mathrm{vol(Box)}} + \pi a v_d r^d \cos^d(\theta) \\ &\geq \pi a v_d r^d \cos^d(\theta) \equiv \gamma. \end{aligned}$$

Using again Lemma 7, if $n > \frac{1}{\gamma}\left(\log N + \log\left(\frac{2}{\delta}\right)\right)$, then with probability at least $1 - \delta/2$, each A_i contains at least one sample point, and hence $M \subset \bigcup_{i \in I} B_{2r}(X_i)$, which implies that $\mathcal{E}_2$ holds.

Combining these we are now ready to again apply the main result from NSW. We restate this lemma in a slightly different form here.

Lemma 8. *[NSW] Let S be a set of points in the tubular neighborhood of radius R around M. Let $U = \bigcup_{x \in S} B_\epsilon(x)$. If S is R-dense in M then $\widehat{\mathcal{H}}(U) = \mathcal{H}(M)$ for all $R < (\sqrt{9} - \sqrt{8})\tau$, if $\epsilon = \frac{R+\tau}{2}$.*

Combining the previously established facts with the lemma above we obtain Lemma 3 from the main paper. Taking $r = (\sqrt{9} - \sqrt{8})\tau/4$ in that lemma, we can see that if $n \geq \frac{C_1}{\pi \tau^d}(\log \frac{C_2}{\tau^d} + \log(C_3/\epsilon))$ then we recover the correct homology with probability at least $1 - \epsilon$.

This is a sample complexity upper bound. Corresponding upper bounds on the minimax risk and resolution follow the arguments of the noiseless case.

A.1.3 Proofs for the tubular noise case

Lower bound In this setting we get samples uniformly in a full dimensional tube around the manifold. We are interested in the case when $\sigma \leq C_0 \tau$ for a small constant C_0.

Let us denote the density q_1 at a point in the tube around M_1 by θ_1 and the density q_2 around M_2 by θ_2. Since, it is not straightforward to decide whether $\theta_1 \leq \theta_2$ or not we will need to consider both possibilities. We will show the calculations assuming $\theta_1 \leq \theta_2$ (the other calculation follows similarly).

Now, remember from the definition of total variation $TV = q_1(G) - q_2(G)$ where G is the set where $q_1 > q_2$. We need an upper bound on total variation and so it suffices to use $TV \leq q_1(G^+) - q_2(G^-)$ where G^+ and G^- are sets containing and contained in G respectively.

Since, $\theta_1 < \theta_2$ we have G is contained in the holes (of radius 4τ) of the two annuli, and G contains a strip of width at least $2\tau - 2\sigma$ in these holes. These are G^+ and G^-.

We need to upper bound the mass under q_1 in G^+ and lower bound the mass under q_2 in G^-. We can now follow the a similar argument to the one made below (in the tubular noise upper bound) to obtain bounds on the various volumes. In each case, the volume of the tubular region is $\Omega(\mathrm{vol}(M)\sigma^{D-d})$, and both M_1 and M_2 have constant volume, in particular $c_1 \leq \mathrm{vol}(M) \leq C_1$. Giving us that the tubular region has volume $\Omega(\sigma^{D-d})$.

It is also clear that both G^+ and G^- have volumes that are $\Omega(\sigma^{D-d}\tau^d)$ (these can be calculated *exactly* since they are cylindrical with no additional curvature but we will not need this here). Here we use that σ is not too close to τ (and in particular is at most a constant fraction of τ).

Since q_1 and q_2 are both uniform in their respective tubes, it follows that

$$TV(q_1, q_2) \leq \Omega\left(\frac{\sigma^{D-d}\tau^d}{\sigma^{D-d}}\right) = \Omega(\tau^d)$$

Notice, that we assumed $\theta_1 \leq \theta_2$ above. The other calculation is nearly identical and we will not reproduce it here.

Upper bound Denote by M_σ the tube of radius σ around M. Recall that we are interested in the case when $\sigma \ll \tau$, and $\epsilon = \tau/2$.

Lemma 9. *If $\epsilon \gg \sigma$ (in particular $\epsilon \geq 2\sigma$ will suffice)*

$$k_\epsilon = \Omega(\epsilon^d).$$

Proof. For any $p \in M$,

$$Q(B_\epsilon(p)) = \frac{vol(B_\epsilon(p) \cap M_\sigma)}{vol(M_\sigma)}.$$

We will prove the claim by deriving derive an upper bound on the denominator and a lower bound on the numerator using packing/covering arguments, both bounds holding uniformly in p.

Upper bound on $vol(M_\sigma)$
We consider a covering of M by γ-balls of d dimensions, and denote the number of balls required N_γ, and the centers $\mathcal{C}_\gamma$. It is clear N_γ is bounded by the number of balls of radius $\gamma/2$ one can pack in M. A simple volume argument then gives

$$N_\gamma \leq C \frac{vol(M)}{(\gamma/2)^d},$$

for some constant C. Given this covering of M, it is easy to see that $\gamma+\sigma$ D-dimensional balls around each of the centers in $\mathcal{C}_\gamma$ covers the tubular region. Thus, we have

$$vol(M_\sigma) \leq v_D N_\gamma (\gamma+\sigma)^D \leq v_D C \frac{vol(M)}{(\gamma/2)^d}(\gamma+\sigma)^D,$$

for any γ. Selecting $\gamma = \sigma$, we have

$$vol(M_\sigma) \leq C_{D,d} vol(M) \sigma^{D-d}$$

for some constant $C_{D,d}$ depending on the manifold and ambient dimensions, independent of σ.

Lower bound on $vol(B_\epsilon(p) \cap M_\sigma)$
Define

$$A(p) = M \cap B_{\epsilon-\sigma}(p),$$
$$B(p) = M \cap B_\epsilon(p),$$
$$B_\sigma(p) = M_\sigma \cap B_\epsilon(p).$$

Denote with N_σ the number of points we can "pack" in $A(p)$ such that the distance between any two points is at least 2σ. Denote the points themselves by the set $\mathcal{C}$. Then,

$$vol(B_\sigma) \geq N_\sigma v_D \sigma^D$$

where v_D is the volume of the unit ball in D-dimensions. To see this just note that every point that is at most σ away from any point in $\mathcal{C}$ is contained in B_σ, and these sets are disjoint so the union of σ balls around $\mathcal{C}$ is contained in B_σ.

Now, to prove a lower bound on N_σ we invoke some ideas from [17]. Consider, the map f described in Lemma 5.3 in [17], which projects the manifold onto its tangent space, and observe its action on $A(p)$. It is clear by their discussion that this map projects the manifold onto a superset of a ball of radius $(\epsilon-\sigma)\cos\theta$, for $\theta = \sin^{-1}(\frac{\epsilon-\sigma}{2\tau})$. In addition to being invertible, this map is a projection, and only shrinks distances between points. So if we can derive a lower bound on the number of points we can "pack" in this projection then it is also a lower bound on N_σ. Now, the set is just a ball in d-dimensions of radius $(\epsilon-\sigma)\cos\theta$. Using, the fact that 2σ balls around each of the points in $\mathcal{C}$ must cover this set a simple volume argument shows

$$N_\sigma (2\sigma)^d \geq v_d((\epsilon-\sigma)\cos\theta)^d,$$

i.e.

$$N_\sigma \geq C_{D,d}\left(\frac{(\epsilon-\sigma)\cos\theta}{\sigma}\right)^d,$$

which gives a lower bound.

Putting the upper and lower bound together, we get

$$\begin{aligned} k_\epsilon &= \inf_{p\in M} Q(B_\epsilon(p)) \\ &\geq C'_{D,d}\frac{1}{vol(M)\sigma^{D-d}}\left(\frac{(\epsilon-\sigma)\cos\theta}{\sigma}\right)^d \sigma^D \\ &= C'_{D,d}\frac{[(\epsilon-\sigma)\cos\theta]^d}{vol(M)}, \end{aligned}$$

for some quantity $C'_{D,d}$, independent of σ. □

We will prove the following main lemma.

Lemma 10. *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 $\epsilon < \frac{(\sqrt{9}-\sqrt{8})\tau}{2}$.*

Proof. This is a straightforward consequence of Lemma 8 and Lemma 7. □

A.1.4 Proof of Theorem 4 (additive case)

Lower Bound

From Lemma 14 we see that convolution only decreases the total variation distance, and so the lower bound for the noiseless case is still valid here.

Upper Bound

We will again proceed by a similar argument to the clutter noise case. Let $\sqrt{D}\sigma < r$, $R = 8r$ and $s = 4r$ and set $\alpha_s = \inf_{p\in A} Q(B_s(p))$ and $\beta_s = \sup_{p\in B} Q(B_s(p))$, where $A = \text{tube}_r(M)$, $B = \mathbb{R}^D - \text{tube}_R(M)$.

As in the clutter noise case, we will need the two events $\mathcal{E}_1$ and $\mathcal{E}_2$ to hold with high probability.

We will use the following version of a common χ^2 inequality, established by [18].

Lemma 11. *For a D-dimensional Gaussian random vector*

$$\mathbb{P}(||\epsilon|| > \sqrt{T}) \leq (ze^{1-z})^{D/2}$$

where $z = \frac{T}{D\sigma^2}$

Using this inequality,

$$\mathbb{P}(||\epsilon|| \geq 4r) \leq (16\exp\{-15\})^{D/2} \equiv t$$

and

$$\mathbb{P}(||\epsilon|| \geq 2r) \leq (4\exp\{-3\})^{D/2} \equiv \gamma.$$

Observe that these are both constants. Next, it is easy to see that

$$\alpha_s \geq Q(B_{s-r}(p)) \geq av_d r^d(\cos\theta)^d(1-\gamma) \equiv \alpha,$$

where $\theta = \sin^{-1}(r/(2\tau))$, and

$$\beta_s \leq v_D(8r)^D t \equiv \beta.$$

As in the clutter noise, we need β to be sufficiently smaller than α if we are to successfully clean the data. As we are interested in the case when r is small, if $D > d$ then we can take $\beta \leq \alpha/2$, while, if $D = d$ then we will need that the dimension is quite large (observe that both γ and t tend to zero rapidly rapidly as D grows).

We are now in a position to invoke the Lemma 13 to ensure $\mathcal{E}_1$ holds with high probability for n large enough. Further, one can see that the mass of an $r/2$-ball close to manifold is at least

$$Q(A_i) \geq av_d(1-\gamma)(\cos\theta)^d(r/2)^d$$

for $\theta = \sin^{-1}(r/(4\tau))$. This quantity is also $O(r^d)$ as desired, and for n large enough we can ensure $\mathcal{E}_2$ holds with high probability. Under the condition on σ, and r we have $r \leq \frac{(\sqrt{9}-\sqrt{8})\tau}{8}$. At this point we can invoke Theorem 5.1 from [18] to see that for $n \asymp^* \frac{1}{\tau^d}$ we recover the correct homology with high probability.

A.1.5 Deconvolution

Upper bound Recall, that the kernel Ψ satisfies

$$\Psi\{x : |x| \geq \epsilon\} \leq \gamma \tag{2}$$

with ϵ and γ being small constants that we will specify in our proof.

The starting point of our proof will be a uniform concentration result from Koltchinskii [16].

Lemma 12. *Consider the event*

$$A = \{\max_x |\widehat{P}_n(B_{2\epsilon}(x)) - \widehat{P}_\Psi(B_{2\epsilon}(x))| < \gamma\}$$

For any small constants ϵ *and* γ*, there exists* $q \in (0,1)$ *such that*

$$P(A^c) \leq 4q^n$$

This lemma tells us that the deconvolved measure is uniformly close to a smoothed (by the kernel Ψ) version of the true density.

Our first step will be to draw

$$m > \frac{1}{\omega}\left(\log l + \log\left(\frac{2}{\delta}\right)\right)$$

samples from $\widehat{P}_n$, where $\omega = \inf_{x\in M} \widehat{P}_n(B_{2\epsilon}(x))$, and l is the 2ϵ covering number of the manifold, and $\delta = 8q^n$. Denote, this sample Z. We know that $l \leq \frac{\text{vol}(M)}{\cos^d(\theta)v_d(2\epsilon)^d}$.

Let us first show that we can choose ϵ and γ so that ω is at least a small positive constant.

$$\begin{aligned}\omega &= \inf_{x\in M} \widehat{P}_n(B_{2\epsilon}(x)) \\ &\geq \inf_{x\in M} P_\Psi(B_{2\epsilon}(x)) - \gamma\end{aligned}$$

Notice that,

$$P_\Psi(B_{2\epsilon}) \geq P(B_\epsilon)\Psi(x : |x| \leq \epsilon)$$

So, we have,

$$\omega \geq \inf_{x\in M} P(B_\epsilon(x))(1-\gamma) - \gamma$$

Using the ball volume lemma we have,

$$\omega \geq av_d\epsilon^d\cos^d\theta(1-\gamma) - \gamma$$

where $\theta = \sin^{-1}(\epsilon/2\tau)$. Notice, that τ is a fixed constant, and ϵ and γ are constants to be chosen appropriately. It is clear that for $\gamma \leq C_{d,\tau}\epsilon$, with $C_{d,\tau}$ small we have

$$\omega \geq c$$

for a small constant c which depends on τ,d and our choices of ϵ and γ.

We now use the sampling lemma 7 to conclude that w.p. at least $1 - 4q^n$,

1. The m samples are 4ϵ dense around M.
2. $M \subset \cup_{i=1}^{m} B_{4\epsilon}(x_i)$

Our next step will be a cleaning step. This cleaning procedure differs from the Algorithm CLEAN in that we use the deconvolved measure to clean the data. In particular, we will remove all points from Z for which $\widehat{P}_n(B_{4\epsilon}(Z_i)) \leq 2\gamma$. Denote the remaining points by W. Our estimator will then be constructed from

$$H = \bigcup B_{\frac{5\epsilon+\tau}{2}}(W_i)$$

To analyze this cleaning procedure, we use the uniform concentration lemma 12 above, and consider the case when event A happens.

1. **All points far away from M are eliminated**: In particular, for any point x if we have

   $$\text{dist}(B_{4\epsilon}(x), M) \geq \epsilon$$

   then the corresponding point is eliminated.

   To see this is simple. We eliminated all points with deconvolved empirical mass $\widehat{P}_n(B_{4\epsilon}) < 2\gamma$. Since, we are assuming event A happened, we have for any remaining point $P_\Psi(B_{4\epsilon}) > \gamma$. Now, we have that

   $$\Psi\{x : |x| \geq \epsilon\} \leq \gamma$$

   From this we see that some part of $B_{4\epsilon}$ must be within ϵ of M, and we have arrived at a contradiction.

2. **All points close to M are kept**: In particular, for any point x if

   $$\text{dist}(x, M) \leq 2\epsilon$$

   then the corresponding point is kept.

   We need to show $\widehat{P}_n(B_{4\epsilon}(x)) \geq 2\gamma$. Notice, that $\widehat{P}_n(B_{4\epsilon}(x)) \geq \widehat{P}_n(B_{2\epsilon}(\pi(x)))$ where $\pi(x)$ is the projection of x onto M. This quantity is just ω.

   To finish, we need to show that we can choose ϵ and γ such that $\omega \geq 2\gamma$. Since, $\omega \geq av_d\epsilon^d \cos^d\theta(1-\gamma) - \gamma$ which as a function of γ is continuous, bounded from below by a constant depending on τ, d and ϵ and monotonically increasing as γ decreases we have for γ small enough

   $$\omega \geq 2\gamma$$

3. **The set H has the right homology**: We have shown that the cleaning eliminates all points outside a tube of radius 5ϵ, and further keeps all points in a tube of radius 2ϵ. From the sampling result we know the points that we keep are 4ϵ dense and that $M \subset \cup_{i=1}^{m} B_{4\epsilon}(x_i)$. We can now apply lemma 8 to conclude that H has the right homology provided

   $$\epsilon < \frac{(\sqrt{9} - \sqrt{8})\tau}{5}$$

   Since τ is a fixed constant we can always choose ϵ small enough to satisfy this condition. To review, we need to select γ and ϵ to satisfy three conditions

   (a) $\omega \geq av_d\epsilon^d \cos^d\theta(1-\gamma) - \gamma$ has to be atleast a small positive constant.
   (b) $\omega \geq 2\gamma$
   (c) $\epsilon < \frac{(\sqrt{9}-\sqrt{8})\tau}{5}$

   Each of these can be satisfied by choosing γ and ϵ small enough.

   Now, returning to m. We have

   $$m > \frac{1}{\omega}\left(\log l + \log\left(\frac{2}{\delta}\right)\right)$$

   where $\omega = \inf_{x \in M} \widehat{P}_n(B_{2\epsilon}(x))$, and l is the 2ϵ covering number of the manifold $l \leq \frac{\text{vol}(M)}{\cos^d(\theta)v_d(2\epsilon)^d}$, and $\delta = 8q^n$. It is clear that all terms except those in n are constant. In particular it is easy to see that

   $$m \geq Cn$$

   for C large enough is sufficient.

From this we can conclude with probability at least $1 - 8q^n$ our procedure will construct an estimator with the correct homology. Since, $q \in (0, 1)$ the success probability can be re-written as at least $1 - e^{-cn}$ for c small enough. Together this gives us the deconvolution lemma from the main paper.

A.2 Additional technical lemmas

A.2.1 The cleaning lemma

In this section we sharpen Lemma 4.1 of [18], also known as the A-B lemma, by using Bernstein's inequality instead of Hoeffding's inequality. This modification is crucial to obtain minimax rates.

Lemma 13. *Let $\beta_s \leq \beta < \alpha/2 \leq \alpha_s/2$. If $n > 4\beta \log \beta$, where*

$$\beta = \max\left(1 + \frac{200}{3\alpha}\log\left(\frac{1}{\delta}\right), 4\right),$$

then procedure CLEAN$(\frac{\alpha+\beta}{2})$ *will remove all points in region B and keep all points in region A with probability at least $1 - \delta$.*

Proof. We use the notation established in section 5.2. We first analyze the set A.

For a point X_i in A, let $q = q(i) = Q(B_s(X_i))$, and define,

$$Z_j = \mathbb{I}(X_j \in B_s(X_i)), \quad j \neq i,$$

where $\mathbb{I}$ denotes the indicator function. Notice that the random variables $\{Z_j, j \neq i\}$ are independent Bernoulli with common mean q.

We will consider two cases.

Case 1: $\alpha \leq q \leq 2\alpha$.
Notice that if

$$q - \frac{1}{n-1}\sum_{j\neq i} Z_j \leq \frac{\alpha}{4}$$

the point X_i will not be removed. By Bernstein's inequality, the probability that X_i will instead be removed is

$$\begin{aligned}\mathbb{P}\left(q - \frac{1}{n-1}\sum_{j\neq i} Z_j \geq \frac{\alpha}{4}\right) &\leq \exp\left\{-\frac{1}{2}\frac{(n-1)(\alpha/4)^2}{2\alpha + \alpha/12}\right\} \\ &\leq \exp\left\{-\frac{3}{200}(n-1)\alpha\right\}.\end{aligned}$$

Case 2: $q > 2\alpha$.
In this case if

$$q - \frac{1}{n-1}\sum_{j\neq i} Z_j \leq q - \frac{3\alpha}{4}$$

the point X_i *will* be removed. Another application of Bernstein's inequality yields

$$\begin{aligned}&\mathbb{P}\left(q - \frac{1}{n-1}\sum_{j\neq i} Z_j \geq q - \frac{3\alpha}{4}\right) \\ &\leq \exp\left\{-\frac{1}{2}\frac{(n-1)(q-3\alpha/4)^2}{q + (q-3\alpha/4)/3}\right\} \\ &\leq \exp\left\{-\frac{1}{2}(n-1)\left[\frac{q}{2} + \frac{9\alpha^2}{32p} - \frac{3\alpha}{4}\right]\right\} \\ &\leq \exp\left\{-\frac{(n-1)\alpha}{8}\right\}.\end{aligned}$$

Now, consider a point X_i in the region B, and define q and the Z_js in an identical way. This time if

$$\frac{1}{n-1}\sum_{j\neq i} Z_j - q \leq \frac{\alpha}{4},$$

the point X_i will not be removed. By Bernstein's inequality,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n-1}\sum_{j\neq i} Z_j - q \geq \frac{\alpha}{4}\right) &\leq \exp\left\{-\frac{1}{2}\frac{(n-1)(\alpha/4)^2}{\alpha/2 + \alpha/12}\right\} \\ &\leq \exp\left\{-\frac{3}{56}(n-1)\alpha\right\}\end{aligned}$$

Putting all the pieces together, we obtain that the cleaning procedure succeeds on all points with probability at least $n\exp\left\{-\frac{3}{200}(n-1)\alpha\right\}$. This requires,

$$\begin{aligned}n - 1 &> \frac{200}{3\alpha}\left(\log n + \log\left(\frac{1}{\delta}\right)\right) \quad \text{i.e.} \\ n &> 1 + \frac{200}{3\alpha}\log\left(\frac{1}{\delta}\right) + \frac{200}{3\alpha}\log n\end{aligned}$$

If $\delta < 1/2$, then $1 + \frac{200}{3\alpha}\log\left(\frac{1}{\delta}\right) > \frac{200}{3\alpha}$, so it is enough to solve

$$n > x + x\log n$$

with $x = 1 + \frac{200}{3\alpha}\log\left(\frac{1}{\delta}\right)$. The result of the lemma follows. □

A.2.2 Convolution only decreases total variation

Lemma 14. *Let P and Q two probability measures in $\mathbb{R}^D$ with common dominating measure μ. Then,*

$$\mathsf{TV}(P \star \Phi, Q \star \Phi) \leq C_\phi \mathsf{TV}(P, Q).$$

where $\star$ denotes deconvolution and Φ is a probability measure on $\mathbb{R}^D$.

Proof. This is a standard result, but we provide a proof for completeness. Let $p \star \phi$ denote the Lebesgue density of the probability distribution $P \star \Phi$, i.e.

$$p \star \phi(z) = \int \phi(z - x)p(x)d\mu(x), \quad z \in \mathbb{R}^D.$$

Similarly, $q \star \phi$ denotes the analogous quantity for $Q \star \Phi$.

Then,

$$
\begin{aligned}
2\mathsf{TV}(P \star \Phi, Q \star \Phi) &= \int_{\mathbb{R}^D} |p \star \phi(z) - q \star \phi(z)| \, dz \\
&= \int_{\mathbb{R}^D} \left| \int \phi(z-x) p(x) d\mu(x) \right. \\
&\quad \left. - \int \phi(z-x) p(x) d\mu(x) \right| dz \\
&= \int_{\mathbb{R}^D} \left| \int \phi(z-x)(p(x) \right. \\
&\quad \left. - q(x)) d\mu(x) \right| dz \\
&\leq \int_{\mathbb{R}^D} \int |\phi(z-x)(p(x) \\
&\quad - q(x))| \, d\mu(x) dz \\
&\leq \int \int_{\mathbb{R}^D} \phi(z-x) dz \, |p(x) - q(x)| \, d\mu(x) \\
&= \int |(p(x) - q(x)| \, d\mu(x) \\
&= 2\mathsf{TV}(P, Q)
\end{aligned}
$$

□