The Annals of Statistics
2012, Vol. 40, No. 2, 941–963
DOI: 10.1214/12-AOS994

MANIFOLD ESTIMATION AND SINGULAR DECONVOLUTION UNDER HAUSDORFF LOSS

BY CHRISTOPHER R. GENOVESE[1], MARCO PERONE-PACIFICO[2], ISABELLA VERDINELLI[2] AND LARRY WASSERMAN[3]

Carnegie Mellon University, Sapienza University of Rome, Carnegie Mellon University and Sapienza University of Rome, and Carnegie Mellon University

We find lower and upper bounds for the risk of estimating a manifold in Hausdorff distance under several models. We also show that there are close connections between manifold estimation and the problem of deconvolving a singular measure.

1. Introduction. Manifold learning is an area of intense research activity in machine learning and statistics. Yet a very basic question about manifold learning is still open, namely, how well can we estimate a manifold from n noisy samples? In this paper we investigate this question under various assumptions.

Suppose we observe a random sample $Y_1, \dots, Y_n \in \mathbb{R}^D$ that lies on or near a d-manifold M where $d < D$. The question we address is: what is the minimax risk under Hausdorff distance for estimating M? Our main assumption is that M is a d-dimensional, smooth Riemannian submanifold in $\mathbb{R}^D$; the precise conditions on M are given in Section 2.

Let Q denote the distribution of Y_i. We shall see that Q depends on several things, including the manifold M, a distribution G supported on M and a model for the noise. We consider three noise models. The first is the *noiseless* model in which $Y_1, \dots, Y_n$ is a random sample from G. The second is the *clutter noise* model, in which

$$Y_1, \dots, Y_n \sim (1-\pi)U + \pi G, \tag{1}$$

where U is a uniform distribution on a compact set $\mathcal{K} \subset \mathbb{R}^D$ with nonempty interior, and G is supported on M. (When $\pi = 1$ we recover the noiseless case.) The third is the *additive* model,

$$Y_i = X_i + Z_i, \tag{2}$$

Received September 2011; revised January 2012.
[1]Supported by NSF Grant DMS-08-06009.
[2]Supported by Italian National Research Grant PRIN 2008.
[3]Supported by NSF Grant DMS-08-06009, Air Force Grant FA95500910373 and Sapienza University of Rome grant for visiting professors 2009.
MSC2010 subject classifications. Primary 62G05, 62G20; secondary 62H12.
Key words and phrases. Deconvolution, manifold learning, minimax.

where $X_1, \dots, X_n \sim G$, G is supported on M, and the noise variables $Z_1, \dots, Z_n$ are a sample from a distribution Φ on $\mathbb{R}^D$ which we take to be Gaussian. In this case, the distribution Q of Y is a convolution of G and Φ written $Q = G \star \Phi$.

In a previous paper [Genovese et al. (2010)], we considered a noise model in which the noise is perpendicular to the manifold. This model is also considered in Niyogi, Smale and Weinberger (2011). Since we have already studied that model, we shall not consider it further here.

In the additive model, estimating M is related to estimating the distribution G, a problem that is usually called *deconvolution* [Fan (1991)]. The problem of deconvolution is well studied in the statistical literature, but in the manifold case there is an interesting complication: the measure G is singular because it puts all its mass on a subset of $\mathbb{R}^D$ that has zero Lebesgue measure (since the manifold has dimension $d < D$). Deconvolution of singular measures has not received as much attention as standard deconvolution problems and raises interesting challenges.

Each noise model gives rise to a class of distributions $\mathcal{Q}$ for Y defined more precisely in Section 2. We are interested in the minimax risk

$$R_n \equiv R_n(\mathcal{Q}) = \inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[H(\widehat{M}, M)], \tag{3}$$

where the infimum is over all estimators $\widehat{M}$, and H is the Hausdorff distance [defined in equation (4)]. Note that finding the minimax risk is equivalent to finding the *sample complexity* $n(\varepsilon) = \inf\{n : R_n \le \varepsilon\}$. We emphasize that the goal of this paper is to find the minimax rates, not to find practical estimators. We use the Hausdorff distance because it is one of the most commonly used metrics for assessing the accuracy of set-valued estimators. One could of course create other loss functions and study their properties, but this is beyond the scope of this paper. Finally, we remark that our upper bounds sometimes differ from our lower bounds by a logarithmic factor. This is a common phenomenon when dealing with Hausdorff distance (and sup norm in function estimation problems). Currently, we do not know how to eliminate the log factor.

1.1. *Related work.* In the additive noise case, estimating a manifold is related to deconvolution problems such as those in Fan (1991), Fan and Truong (1993) and Stefanski (1990). More closely related is the problem of estimating the support of a distribution in the presence of noise as discussed, for example, in Meister (2006).

There is a vast literature on manifold estimation. Much of the literature deals with using manifolds for the purpose of dimension reduction. See, for example, Baraniuk and Wakin (2009) and references therein. We are interested instead in actually estimating the manifold itself. There is a literature on this problem in the field of computational geometry; see Dey (2007). However, very few papers allow for noise in the statistical sense, by which we mean observations drawn randomly from a distribution. In the literature on computational geometry, observations are called noisy if they depart from the underlying manifold in a very specific way: the observations have to be close to the manifold but not too close to each other. This notion of noise is quite different from random sampling from a distribution. An exception is Niyogi, Smale and Weinberger (2008), who constructed the following estimator: Let $I = \{i : \widehat{p}(Y_i) > \lambda\}$ where $\widehat{p}$ is a density estimator. They define $\widehat{M} = \bigcup_{i \in I} B_D(Y_i, \varepsilon)$ where $B_D(Y_i, \varepsilon)$ is a ball in $\mathbb{R}^D$ of radius ε centered at Y_i. Niyogi, Smale and Weinberger (2008) show that if λ and ε are chosen properly, then $\widehat{M}$ is homologous to M. This means that M and $\widehat{M}$ share certain topological properties. However, the result does not guarantee closeness in Hausdorff distance. A very relevant paper is Caillerie et al. (2011). These authors consider observations generated from a manifold and then contaminated by additive noise as we do in Section 5. Also, they use deconvolution methods as we do. However, their interest is in upper bounding the Wasserstein distance between an estimator $\widehat{G}$ and the distribution G, as a prelude to estimating the homology of M. They do not establish Hausdorff bounds. Koltchinskii (2000) considers estimating the number of connected components of a set, contaiminated by additive noise. This corresponds to estimating the zeroth order homology.

There is a also a literature on estimating principal surfaces. A recent paper on this approach with an excellent review is Ozertem and Erdogmus (2011). This is similar to estimating manifolds but, to the best of our knowledge, this literature does not establish minimax bounds for estimation in Hausdorff distance. Finally we would like to mention the related problem of testing for a set of points on a surface in a field of uniform noise [Arias-Castro et al. (2005)], but, despite some similarity, this problem is quite different.

1.2. *Notation.* We let $B_D(x, r)$ denote a D-dimensional open ball centered at x with radius r. If A is a set, and x is a point, then we write $d(x, A) = \inf_{y \in A} \|x - y\|$ where $\|\cdot\|$ is the Euclidean norm. Given two sets A and B, the *Hausdorff distance* between A and B is

$$H(A, B) = \inf\{\varepsilon : A \subset B \oplus \varepsilon \text{ and } B \subset A \oplus \varepsilon\}, \tag{4}$$

where

$$A \oplus \varepsilon = \bigcup_{x \in A} B_D(x, \varepsilon). \tag{5}$$

The L_1 distance between two distributions P and Q with densities p and q is $\ell_1(p, q) = \int |p - q|$ and the *total variation distance* between P and Q is

$$\mathsf{TV}(P, Q) = \sup_A |P(A) - Q(A)|, \tag{6}$$

where the supremum is over all measurable sets A. Recall that $\mathsf{TV}(P, Q) = (1/2)\ell_1(p, q)$.

Let $p(x) \wedge q(x) = \min\{p(x), q(x)\}$. The *affinity* between P and Q is

$$\|P \wedge Q\| = \int p \wedge q = 1 - \frac{1}{2} \int |p - q|. \tag{7}$$

Let P^n denote the n-fold product measure based on n independent observations from P. It can be shown that

$$\|P^n \wedge Q^n\| \geq \frac{1}{8}\left(1 - \frac{1}{2}\int |p-q|\right)^{2n}. \tag{8}$$

The convolution between two measures P and Φ—denoted by $P \star \Phi$—is the measure defined by

$$(P \star \Phi)(A) = \int \Phi(A-x)\, dP(x). \tag{9}$$

If Φ has density ϕ, then $P \star \Phi$ has density $\int \phi(y-u)\, dP(u)$. The Fourier transform of P is denoted by

$$p^*(t) = \int e^{it^T u}\, dP(u) = \int e^{it\cdot u}\, dP(u), \tag{10}$$

where we use both $t^T u$ and $t \cdot u$ to denote the dot product.

We write $X_n = O_P(a_n)$ to mean that for every $\varepsilon > 0$, there exists $C > 0$ such that $\mathbb{P}(\|X_n\|/a_n > C) \leq \varepsilon$ for all large n. Throughout, we use symbols like $C, C_0, C_1, c, c_0, c_1, \ldots$ to denote generic positive constants whose value may be different in different expressions. We write $\mathsf{poly}(\varepsilon)$ to denote any expression of the form $a\varepsilon^b$ for some positive real numbers a and b. We write $a_n \preceq b_n$ if there exists $c > 0$ such that $a_n \leq cb_n$ for all large n. Similarly, write $a_n \succeq b_n$ if $b_n \preceq a_n$. Finally, write $a_n \asymp b_n$ if $a_n \preceq b_n$ and $b_n \preceq a_n$.

We will use Le Cam's lemma to derive lower bounds, which we now state. This version is from Yu (1997).

LEMMA 1 (Le Cam 1973). *Let $\mathcal{Q}$ be a set of distributions. Let $\theta(Q)$ take values in a metric space with metric ρ. Let $Q_0, Q_1 \in \mathcal{Q}$ be any pair of distributions in $\mathcal{Q}$. Let $Y_1, \ldots, Y_n$ be drawn i.i.d. from some $Q \in \mathcal{Q}$ and denote the corresponding product measure by Q^n. Let $\widehat{\theta} = \widehat{\theta}(Y_1, \ldots, Y_n)$ be any estimator. Then*

$$\sup_{Q\in\mathcal{Q}} \mathbb{E}_{Q^n}[\rho(\widehat{\theta}, \theta(Q))] \geq \rho(\theta(Q_0), \theta(Q_1))\|Q_0^n \wedge Q_1^n\|$$
$$\geq \rho(\theta(Q_0), \theta(Q_1))\frac{1}{8}(1 - \mathsf{TV}(Q_0, Q_1))^{2n}.$$

2. Assumptions. We shall be concerned with d-dimensional Riemannian submanifolds of $\mathbb{R}^D$ where $d < D$. Usually, we assume that M is contained in some compact set $\mathcal{K} \subset \mathbb{R}^D$. An exception is Section 5 where we allow noncompact manifolds. Let $\Delta(M)$ be the largest r such that each point in $M \oplus r$ has a unique projection onto M. The quantity $\Delta(M)$ will be small if either M is not smooth or if M is close to being self-intersecting. The quantity $\Delta(M)$ has been rediscovered many times. It is called the *condition number* in Niyogi, Smale and Weinberger

(2008) and the *reach* in Federer (1959). Let $\mathcal{M}(\kappa)$ denote all d-dimensional manifolds embedded in $\mathbb{R}^D$ such that $\Delta(M) \geq \kappa$. Throughout this paper, κ is a fixed positive constant.

We consider three different distributional models:

(1) *Noiseless*. We observe $Y_1, \ldots, Y_n \sim G$ where G is supported on a manifold M where $M \in \mathcal{M} = \{M \in \mathcal{M}(\kappa), M \subset \mathcal{K}\}$. In this case, $Q = G$ and the observed data fall exactly on the manifold. We assume that G has density g with respect to the uniform distribution on M and that

$$0 < b(\mathcal{M}) \leq \inf_{y \in M} g(y) \leq \sup_{y \in M} g(y) \leq B(\mathcal{M}) < \infty, \tag{11}$$

where $b(\mathcal{M})$ and $B(\mathcal{M})$ are allowed to depend on the class $\mathcal{M}$, but not on the particular manifold M. Let $\mathcal{G}(M)$ denote all such distributions. In this case we define

$$\mathcal{Q} = \mathcal{G} = \bigcup_{M \in \mathcal{M}} \mathcal{G}(M). \tag{12}$$

(2) *Clutter noise*. Define $\mathcal{M}$ and $\mathcal{G}(M)$ as in the noiseless case. We observe

$$Y_1, \ldots, Y_n \sim Q \equiv (1 - \pi)U + \pi G, \tag{13}$$

where $0 < \pi \leq 1$, U is uniform on the compact set $\mathcal{K} \subset \mathbb{R}^D$ and $G \in \mathcal{G}(M)$. Define

$$\mathcal{Q} = \{Q = (1 - \pi)U + \pi G : G \in \mathcal{G}(M), M \in \mathcal{M}\}. \tag{14}$$

(3) *Additive noise*. In this case we allow the manifolds to be noncompact. However, we do require that each G put nontrivial probability in some fixed compact set. Specifically, we again fix a compact set $\mathcal{K}$. Let $\mathcal{M} = \mathcal{M}(\kappa)$. Fix positive constants $0 < b(\mathcal{M}) < B(\mathcal{M}) < \infty$. For any $M \in \mathcal{M}$, let $\mathcal{G}(M)$ be the set of distributions G supported on M, such that G has density g with respect to Hausdorff measure on M, and such that

$$0 < b(\mathcal{M}) \leq \inf_{y \in M \cap \mathcal{K}} g(y) \leq \sup_{y \in M \cap \mathcal{K}} g(y) \leq B(\mathcal{M}) < \infty. \tag{15}$$

Let $X_1, X_2, \ldots, X_n \sim G \in \mathcal{G}(M)$, and define

$$Y_i = X_i + Z_i, \qquad i = 1, \ldots, n, \tag{16}$$

where Z_i are i.i.d. draws from a distribution Φ on $\mathbb{R}^D$, and where Φ is a standard D-dimensional Gaussian. Let $Q = G \star \Phi$ be the distribution of each Y_i and Q^n be the corresponding product measure. Let $\mathcal{Q} = \{G \star \Phi : G \in \mathcal{G}(M), M \in \mathcal{M}\}$.

These three models are an attempt to capture the idea that we have data falling on or near a manifold. These appear to be the most commonly used models. No doubt, one could create other models as well which is a topic for future research. As we mentioned earlier, a different noise model is considered in Niyogi, Smale and Weinberger (2011) and in Genovese et al. (2010). Those authors consider the case where the noise is perpendicular to the manifold. The former paper considers estimating the homology groups of M while the latter paper shows that the minimax Hausdorff rate is $n^{-2/(2+d)}$ in that case.

3. Noiseless case. We now derive the minimax bounds in the noiseless case.

THEOREM 2. *Under the noiseless model, we have*

$$\inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}[H(\widehat{M}, M)] \geq C n^{-2/d}. \tag{17}$$

PROOF. Fix $\gamma > 0$. By Theorem 6 of Genovese et al. (2010) there exist manifolds M_0, M_1 that satisfy the following conditions:

(1) $M_0, M_1 \in \mathcal{M}$.
(2) $H(M_0, M_1) = \gamma$.
(3) There is a set $B \subset M_1$ such that:
  (a) $\inf_{y \in M_0} \|x - y\| > \gamma/2$ for all $x \in B$.
  (b) $\mu_1(B) \geq \gamma^{d/2}$ where μ_1 is the uniform measure on M_1.
  (c) There is a point $x \in B$ such that $\|x - y\| = \gamma$ where $y \in M_0$ is the closest point on M_0 to x. Moreover, $T_x M_1$ and $T_y M_0$ are parallel where $T_x M$ is the tangent plane to M at x.
(4) If $A = \{y : y \in M_1, y \notin M_0\}$, then $\mu_1(A) \leq C\gamma^{d/2}$ for some $C > 0$.

Let $Q_i = G_i$ be the uniform measure on M_i, for $i = 0, 1$, and let A be the set defined in the last item. Then $\mathsf{TV}(G_0, G_1) = G_1(A) - G_0(A) = G_1(A) \leq C\gamma^{d/2}$. From Le Cam's lemma,

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} H(\widehat{M}, M) \geq \gamma(1 - \gamma^{d/2})^{2n}. \tag{18}$$

Setting $\gamma = (1/n)^{2/d}$ yields the stated lower bound. $\square$

See Figure 1 for a heuristic explanation of the construction of the two manifolds, M_0 and M_1, used in the above proof. Now we derive an upper bound.

THEOREM 3. *Under the noiseless model, we have*

$$\inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}[H(\widehat{M}, M)] \leq C \left(\frac{\log n}{n} \right)^{2/d}. \tag{19}$$

Hence, the rate is tight, up to logarithmic factors. The proof is a special case of the proof of the upper bound in the next section and so is omitted.

REMARK. The Associate Editor pointed out that the rate $(1/n)^{2/d}$ might seem counterintuitive. For example, when $d = 1$, this yields $(1/n)^2$ which would seem to contradict the usual $1/n$ rate for estimating the support of a uniform distribution. However, the slower $1/n$ rate is actually a boundary effect much like the boundary effects that occur in density estimation and regression. If we embed the uniform into $\mathbb{R}^2$ and wrap it into a circle to eliminate the boundary, we do indeed get a rate of $1/n^2$. Our assumption of smooth manifolds without boundary removes the boundary effect.

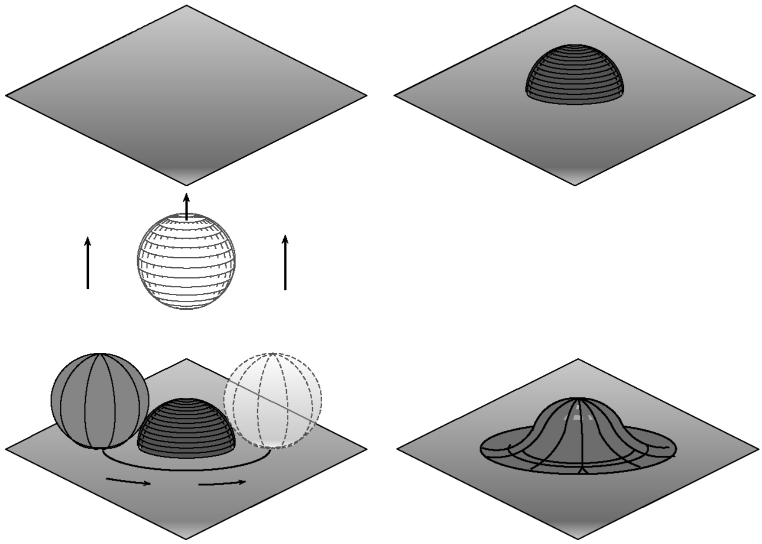

FIG. 1. *The proof of Theorem 2 uses two manifolds, M_0 and M_1. A sphere of radius κ is pushed upward into the plane M_0 (top left). The resulting manifold M_0' is not smooth (top right). A sphere is then rolled around the manifold (bottom left) to produce a smooth manifold M_1 (bottom right). The construction is made rigorous in Theorem 6 of Genovese et al. (2010).*

4. Clutter noise.

Recall that

$$Y_1, \dots, Y_n \sim Q = (1 - \pi)U + \pi G,$$

where U is uniform on $\mathcal{K}$, $0 < \pi \le 1$ and $G \in \mathcal{G}$.

THEOREM 4. *Under the clutter model, we have*

$$\inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}[H(\widehat{M}, M)] \ge C\left(\frac{1}{n\pi}\right)^{2/d}. \tag{20}$$

PROOF. We define M_0, M_1 and A as in the proof of Theorem 2. Let $Q_0 = (1-\pi)U + \pi G_0$ and $Q_1 = (1-\pi)U + \pi G_1$. Then $\mathsf{TV}(Q_0, Q_1) = \pi \mathsf{TV}(G_0, G_1)$. Hence $\mathsf{TV}(Q_0, Q_1) \le \pi(G_1(A) - G_0(A)) = \pi G_1(A) \le C\pi\gamma^{d/2}$. From Le Cam's lemma,

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}[H(\widehat{M}, M)] \ge \gamma(1 - \pi\gamma^{d/2})^{2n}. \tag{21}$$

Setting $\gamma = (1/n\pi)^{2/d}$ yields the stated lower bound. $\square$

Now we consider the upper bound. Let $\widehat{Q}_n$ be the empirical measure. Let $\varepsilon_n = (K \log n/n)^{2/d}$ where $K > 0$ is a large positive constant. Given a manifold M and a point $y \in M$ let $S_M(y)$ denote the slab, centered at y, with size $b_1\sqrt{\varepsilon_n}$

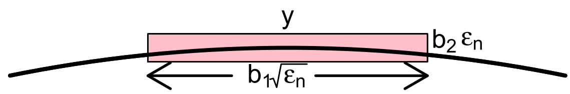

FIG. 2. *Given a manifold M and a point $y \in M$, $S_M(y)$ is a slab, centered at y, with size $O(\sqrt{\varepsilon_n})$ in the d directions corresponding to the tangent space $T_y M$ and size $O(\varepsilon_n)$ in the $D-d$ normal directions.*

in the d directions corresponding to the tangent space $T_y M$ and size $b_2\varepsilon_n$ in the $D-d$ normal directions to the tangent space. Here, b_1 and b_2 are small, positive constants. See Figure 2.

Define

$$s(M) = \inf_{y \in M} \widehat{Q}_n[S_M(y)] \quad \text{and} \quad \widehat{M}_n = \arg\max_M s(M).$$

In case of ties we take any maximizer.

THEOREM 5. *Let $\xi > 1$ and let $\varepsilon_n = (K \log n/n)^{2/d}$ where K is a large, positive constant. Then*

$$\sup_{Q \in \mathcal{Q}} Q^n\big(H(M_0, \widehat{M}_n) > \varepsilon_n\big) < n^{-\xi}$$

and hence

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}(H(M_0, \widehat{M}_n)) \le C\varepsilon_n.$$

We will use the following result, which follows from Theorem 7 of Bousquet, Boucheron and Lugosi (2004). This version of the result is from Chaudhuri and Dasgupta (2010).

LEMMA 6. *Let $\mathcal{A}$ be a class of sets with VC dimension V. Let $0 < u < 1$ and*

$$\beta_n = \sqrt{\left(\frac{4}{n}\right)\left[V \log(2n) + \log\left(\frac{8}{u}\right)\right]}.$$

Then for all $A \in \mathcal{A}$,

$$\begin{aligned} &-\min\{\beta_n\sqrt{\widehat{Q}_n(A)}, \beta_n^2 + \beta_n\sqrt{Q(A)}\} \\ &\qquad \le Q(A) - \widehat{Q}_n(A) \le \min\{\beta_n^2 + \beta_n\sqrt{\widehat{Q}_n(A)}, \beta_n\sqrt{Q(A)}\} \end{aligned}$$

with probability at least $1-u$.

The set of hyper-rectangles in $\mathbb{R}^D$ (which contains all the slabs) has finite VC dimension V, say. Hence, we have the following lemma obtained by setting $u = (1/n)^\xi$.

LEMMA 7. *Let $\mathcal{A}$ denote all hyper-rectangles in $\mathbb{R}^D$. Let $C = 4[V + \max\{3, \xi\}]$. Then for all $A \in \mathcal{A}$,*

$$\widehat{Q}_n(A) \le Q(A) + \frac{C \log n}{n} + \sqrt{\frac{C \log n}{n}}\sqrt{Q(A)} \quad \text{and} \tag{22}$$

$$\widehat{Q}_n(A) \ge Q(A) - \sqrt{\frac{C \log n}{n}}\sqrt{Q(A)} \tag{23}$$

with probability at least $1 - (1/n)^\xi$.

Now we can prove Theorem 5.

PROOF OF THEOREM 5. Let M_0 denote the true manifold. Assume that (22) and (23) hold. Let $y \in M_0$ and let $A = S_{M_0}(y)$. Note that $Q(A) = (1-\pi)U(A) + \pi G(A)$. Since $y \in M_0$ and G is singular, the term $U(A)$ is of lower order and so there exist $0 < c_1 \le c_2 < \infty$ such that, for all large n,

$$\frac{c_1 K \log n}{n} = c_1 \varepsilon_n^{d/2} \le Q(A) \le c_2 \varepsilon_n^{d/2} = \frac{c_2 K \log n}{n}.$$

Hence

$$\widehat{Q}_n(A) \ge Q(A) - \sqrt{\frac{C \log n}{n}}\sqrt{Q(A)} \ge \frac{c_1 K \log n}{n} - \sqrt{c_2' K}\frac{\log n}{n} > \frac{c_3 K \log n}{n}.$$

Thus $s(M_0) > \frac{c_3 K \log n}{n}$ with high probablity.

Now consider any M for which $H(M_0, M) > \varepsilon_n$. There exists a point $y \in M$ such that $d(y, M_0) > \varepsilon_n$. It can be seen, since $M \in \mathcal{M}$, that $S_M(y) \cap M_0 = \varnothing$. [To see this, note that $\Delta(M) \ge \kappa > 0$ implies that the interior of any ball of radius κ tangent to M at y has empty intersection with M and the slab $S_M(y)$ is strictly contained in such a ball for b_1 and b_2 small enough relative to κ.] Hence

$$\begin{aligned} Q(S_M(y)) &= (1-\pi)U(S_M(y)) = c_4 \varepsilon_n^{d/2} \varepsilon_n^{D-d} \\ &= \left(\frac{K \log n}{n}\right) c_4 \left(\frac{K \log n}{n}\right)^{2(D-d)/d} = C\left(\frac{\log n}{n}\right)^{(2D-d)/d}. \end{aligned}$$

So, from the previous lemma,

$$\begin{aligned} s(M) &= \inf_{x \in M} \widehat{Q}_n(S_M(x)) \le \widehat{Q}_n(S_M(y)) \\ &\le Q(S_M(y)) + \frac{C \log n}{n} + \sqrt{\frac{C \log n}{n}}\sqrt{Q(S_M(y))} \\ &= \left(\frac{K \log n}{n}\right)^{(2D-d)/d} + \frac{C \log n}{n} + \left(\frac{K \log n}{n}\right)^{D/d} < \frac{C_3 K \log n}{n} = s(M_0) \end{aligned}$$

since $D > d$ and K is large. Let $\mathcal{M}_n = \{M \in \mathcal{M} : H(M_0, M) > \varepsilon_n\}$. We conclude that

$$Q^n\bigl(s(M) > s(M_0) \text{ for some } M \in \mathcal{M}_n\bigr) < \left(\frac{1}{n}\right)^{\xi}. \qquad \square$$

5. Additive noise. Let us recall the model. Let $\mathcal{M} = \mathcal{M}(\kappa)$. We allow the manifolds to be noncompact. Fix positive constants $0 < b(\mathcal{M}) < B(\mathcal{M}) < \infty$. For any $M \in \mathcal{M}$ let $\mathcal{G}(M)$ be the set of distributions G supported on M such that G has density g with respect to Hausdorff measure on M and such that

$$0 < b(\mathcal{M}) \le \inf_{y \in M \cap \mathcal{K}} g(y) \le \sup_{y \in M \cap \mathcal{K}} g(y) \le B(\mathcal{M}) < \infty, \tag{24}$$

where $\mathcal{K}$ is a compact set. Let $X_1, X_2, \ldots, X_n \sim G \in \mathcal{G}(M)$, and define

$$Y_i = X_i + Z_i, \qquad i = 1, \ldots, n, \tag{25}$$

where Z_i are i.i.d. draws from a distribution Φ on $\mathbb{R}^D$, and where Φ is a standard D-dimensional Gaussian. Let $Q = G \star \Phi$ be the distribution of each Y_i and Q^n be the corresponding product measure. Let $\mathcal{Q} = \{G \star \Phi : G \in \mathcal{G}(M), M \in \mathcal{M}\}$.

Since we allow the manifolds to be noncompact, the Hausdorff distance could be unbounded. Hence we define a truncated loss function,

$$L(M, \widehat{M}) = H(M \cap \mathcal{K}, \widehat{M} \cap \mathcal{K}). \tag{26}$$

THEOREM 8. *For all large enough n,*

$$\inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[L(M, \widehat{M})] \ge \frac{C}{\log n}. \tag{27}$$

PROOF. Define $\widetilde{c} : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R}^d \to \mathbb{R}^{D-d}$ as follows: $\widetilde{c}(x) = \cos(x/(a\sqrt{\gamma}))$ and $c(u) = (\prod_{\ell=1}^d \widetilde{c}(u_\ell), 0, \ldots, 0)^T$. Let $M_0 = \{(u, \gamma c(u)) : u \in \mathbb{R}^d\}$ and $M_1 = \{(u, -\gamma c(u)) : u \in \mathbb{R}^d\}$. See Figure 3 for a picture of M_0 and M_1 when $D = 2$, $d = 1$. Later, we will show that $M_0, M_1 \in \mathcal{M}$.

Let U be a d-dimensional random variable with density ζ where ζ is d-dimensional standard Gaussian density. Let $\widetilde{\zeta}$ be a one-dimensional $N(0,1)$ density. And define G_0 and G_1 by $G_0(A) = \mathbb{P}((U, \gamma c(U)) \in A)$ and $G_1(A) = \mathbb{P}((U, -\gamma c(U)) \in A)$.

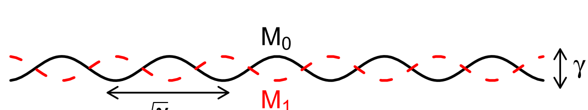

FIG. 3. *The two least favorable manifolds M_0 and M_1 in the proof of Theorem 8 in the special case where $D = 2$ and $d = 1$.*

We begin by bounding $\int |q_1 - q_0|^2$. Define the D-cube $\mathcal{Z} = [-1/(2a\sqrt{\gamma}), 1/(2a\sqrt{\gamma})]^D$. Then, by Parseval's identity, and that fact that $q_j^* = \phi^* g_j^*$,

$$\begin{aligned}(2\pi)^D \int |q_1 - q_0|^2 &= \int |q_1^* - q_0^*|^2 = \int |\phi^*|^2 |g_1^* - g_0^*|^2 \\ &= \int_{\mathcal{Z}} |\phi^*|^2 |g_1^* - g_0^*|^2 + \int_{\mathcal{Z}^c} |\phi^*|^2 |g_1^* - g_0^*|^2 \\ &\equiv I + II.\end{aligned}$$

Then

$$\begin{aligned}II &= \int_{\mathcal{Z}^c} |g_1^*(t) - g_0^*(t)|^2 |\phi^*(t)|^2 \\ &\le \int_{\mathcal{Z}^c} |\phi^*(t)|^2 \le C\left(\int_{1/(2a\sqrt{\gamma})}^{\infty} e^{-t^2}\,dt\right)^D \\ &\le \mathsf{poly}(\gamma) e^{-D/4a^2\gamma}.\end{aligned}$$

Now we bound I. Write $t \in \mathbb{R}^D$ as (t_1, t_2) where $t_1 = (t_{11}, \dots, t_{1d}) \in \mathbb{R}^d$ and $t_2 = (t_{21}, \dots, t_{2(D-d)}) \in \mathbb{R}^{D-d}$. Let $c_1(u) = \prod_{\ell=1}^d \widetilde{c}(u_\ell)$ denote the first component of the vector-valued function c. We have

$$\begin{aligned}g_1^*(t) - g_0^*(t) &= \int_{\mathbb{R}^d} \big(e^{it_1\cdot u + it_{21}\gamma c_1(u)} - e^{it_1\cdot u - it_{21}\gamma c_1(u)}\big)\zeta(u)\,du \\ &= 2i \int e^{it_1\cdot u} \sin(t_{21}\gamma c_1(u))\zeta(u)\,du \\ &= 2i \int e^{it_1\cdot u} \sum_{k=0}^{\infty} \frac{(-1)^k t_{21}^{2k+1}\gamma^{2k+1}}{(2k+1)!} c_1^{2k+1}(u)\zeta(u)\,du \\ &= 2i \sum_{k=0}^{\infty} \frac{(-1)^k t_{21}^{2k+1}\gamma^{2k+1}}{(2k+1)!} \int e^{it_1\cdot u} c_1^{2k+1}(u)\zeta(u)\,du \\ &= 2i \sum_{k=0}^{\infty} \frac{(-1)^k t_{21}^{2k+1}\gamma^{2k+1}}{(2k+1)!} \prod_{\ell=1}^d \int e^{it_{1\ell}u_\ell} \widetilde{c}^{2k+1}(u_\ell)\widetilde{\zeta}(u_\ell)\,du_\ell \\ &= 2i \sum_{k=0}^{\infty} \frac{(-1)^k t_{21}^{2k+1}\gamma^{2k+1}}{(2k+1)!} \prod_{\ell=1}^d (\widetilde{c}^{2k+1}\widetilde{\zeta})^*(t_{1\ell}) \\ &= 2i \sum_{k=0}^{\infty} \frac{(-1)^k t_{21}^{2k+1}\gamma^{2k+1}}{(2k+1)!} \prod_{\ell=1}^d m_k(t_{1\ell}),\end{aligned}$$

where

$$m_k(t_{1\ell}) = (\widetilde{c}^{2k+1}\widetilde{\zeta})^*(t_{1\ell}) = (\underbrace{\widetilde{c}^* \star \widetilde{c}^* \star \cdots \star \widetilde{c}^*}_{2k+1 \text{ times}} \star \widetilde{\zeta}^*)(t_{1\ell}). \tag{28}$$

Note that

$$\widetilde{c}^* = \tfrac{1}{2}\delta_{-1/(a\sqrt{\gamma})} + \tfrac{1}{2}\delta_{1/(a\sqrt{\gamma})},$$

where δ_y a Dirac delta function at y, that is, a generalized function corresponding to point evaluation at y. For any integer r, if we convolve $\widetilde{c}^*$ with itself r times, we have that

$$\underbrace{\widetilde{c}^* \star \widetilde{c}^* \star \cdots \star \widetilde{c}^*}_{r \text{ times}} = \left(\frac{1}{2}\right)^r \sum_{j=0}^{r} \binom{r}{j} \delta_{a_j}, \tag{29}$$

where $a_j = (2j - r)/(a\sqrt{\gamma})$. Thus

$$m_k(t_{1\ell}) = \left(\frac{1}{2}\right)^{2k+1} \sum_{j=0}^{2k+1} \binom{2k+1}{j} \widetilde{\zeta}^*(t_{1\ell} - a_j). \tag{30}$$

Now $\widetilde{\zeta}^*(t_{1\ell}) = \exp(-\frac{t_{1\ell}^2}{2})$ and $\widetilde{\zeta}^*(s) \le 1$ for all $s \in \mathbb{R}$. For $t \in \mathcal{Z}$, $\widetilde{\zeta}^*(t_{1\ell} - a_j) \le e^{-1/(2a^2\gamma)}$, and thus $|m_k(t_{1\ell})| \le e^{-1/(2a^2\gamma)}$. Hence, $\prod_{\ell=1}^{d} |m_k(t_{1\ell})| \le e^{-d/(2a^2\gamma)}$. It follows that for $t \in \mathcal{Z}$,

$$\begin{aligned}
|g_1^*(t) - g_0^*(t)| &\le 2\sum_{k=0}^{\infty} \frac{|t_{21}|^{2k+1}\gamma^{2k+1}}{(2k+1)!} \prod_{\ell=1}^{d} |m_k(t_{1\ell})| \\
&\le e^{-d/(2a^2\gamma)} \sum_{k=0}^{\infty} \frac{|t_{21}|^{2k+1}\gamma^{2k+1}}{(2k+1)!} \\
&\le e^{-d/(2a^2\gamma)} \sinh(|t_{21}|\gamma) \le e^{-d/(2a^2\gamma)}.
\end{aligned}$$

So,

$$\begin{aligned}
I &= \int_{\mathcal{Z}} |g_1^*(t) - g_0^*(t)|^2 |\phi^*(t)|^2 \, dt \\
&\le \int_{\mathcal{Z}} |g_1^*(t) - g_0^*(t)|^2 \, dt \\
&\le \mathsf{Volume}(\mathcal{Z}) e^{-d/(a^2\gamma)} = \mathsf{poly}(\gamma) e^{-d/(a^2\gamma)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\int |q_1 - q_0|^2 &\le I + II \le \mathsf{poly}(\gamma) e^{-d/a^2\gamma} + \mathsf{poly}(\gamma) e^{-D/4a^2\gamma} \\
&= \mathsf{poly}(\gamma) e^{-2w/\gamma},
\end{aligned} \tag{31}$$

where $2w = \min\{d/a^2, D/(4a^2)\}$.

Next we bound $\int |q_1 - q_0|$ so that we can apply Le Cam's lemma. Let T_γ be a ball centered at the origin with radius $1/\gamma$. Then, by Cauchy–Schwarz,

$$\begin{aligned}\int |q_1 - q_0| &= \int_{T_\gamma} |q_1 - q_0| + \int_{T_\gamma^c} |q_1 - q_0| \\ &\le \sqrt{\mathsf{Volume}(T_\gamma)}\sqrt{\int |q_1 - q_0|^2} + \int_{T_\gamma^c} |q_1 - q_0| \\ &\le \mathsf{poly}(\gamma) e^{-w/\gamma} + \int_{T_\gamma^c} |q_1 - q_0|.\end{aligned}$$

For all small γ we have that $\mathcal{K} \subset T_\gamma$. Hence,

$$\begin{aligned}\int_{T_\gamma^c} |q_1 - q_0| &\le \int_{M_1}\int_{T_\gamma^c} \phi(\|y-u\|) + \int_{M_0}\int_{T_\gamma^c} \phi(\|y-u\|) \le \mathsf{poly}(\gamma) e^{-D/\gamma^2} \\ &\le \mathsf{poly}(\gamma) e^{-w/\gamma}.\end{aligned}$$

Putting this all together, we have that $\int |q_1 - q_0| \le \mathsf{poly}(\gamma) e^{-w/\gamma}$.

Now we apply Lemma 1 and conclude that, for every $\gamma > 0$,

$$\sup_Q \mathbb{E}(L(M, \widehat{M})) \ge \frac{\gamma}{8}\bigl(1 - \mathsf{poly}(\gamma) e^{-w/\gamma}\bigr)^{2n}.$$

Set $\gamma \asymp w/\log n$ and conclude that, for all large n,

$$\sup_Q \mathbb{E}(L(M, \widehat{M})) \ge \frac{w}{8e^2}\frac{1}{\log n}.$$

This concludes the proof of the lower bound except that it remains to show that $M_0, M_1 \in \mathcal{M}(\kappa)$. Note that $|\widetilde{c}''(u)| = a^{-2}|\cos(u/(a\sqrt{\gamma})|$. Hence, as long as $a > \sqrt{\kappa}$, $\sup_u |\widetilde{c}''(u)| < 1/\kappa$. It now follows that $M_0, M_1 \in \mathcal{M}(\kappa)$. This completes the proof. $\square$

REMARK. Consider the special case where $D = 2$, $d = 1$ and the manifold has the special form $\{(u, m(u)) : u \in \mathbb{R}\}$ for some function $m : \mathbb{R} \to \mathbb{R}$. In this case, estimating the manifold is like estimating a regression function with errors in variables. (More on this in Section 6.) The rate obtained for estimating a regression function with errors in variables under these conditions [Fan and Truong (1993)] is $1/\log n$ in agreement with our rate. However, the proof technique is not quite the same as we explain in Section 6.

REMARK. The proof of the lower bound is similar to other lower bounds in deconvolution problems. There is an interesting technical difference, however. In standard deconvolution, we can choose G_0 and G_1 so that $g_1^*(t) - g_0^*(t)$ is zero in a large neighborhood around the origin. This simplifies the proof considerably. It appears we cannot do this in the manifold case since G_0 and G_1 have different supports.

Next we construct an upper bound. We use a standard deconvolution density estimator $\widehat{g}$ (even thought G has no density), and then we threshold this estimator.

THEOREM 9. *Fix any $0 < \delta < 1/2$. Let $h = 1/\sqrt{\log n}$. Let λ_n be such that*

$$C'\left(\frac{1}{h}\right)^{D-d} < \lambda_n < C''\left(\frac{1}{L}\right)^{2k}\left(\frac{1}{h}\right)^{D-d},$$

where $k \ge d/(2\delta)$, C' is defined in Lemma 11 and C'' and L are defined in Lemma 12. Define $\widehat{M} = \{y : \widehat{g}(y) > \lambda_n\}$ where $\widehat{g}$ is defined in (34). Then for all large n,

$$\inf_{\widehat{M}} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[L(M, \widehat{M})] \le C\left(\frac{1}{\log n}\right)^{(1-\delta)/2}. \tag{32}$$

Let us now define the estimator in more detail. Define $\psi_k(y) = \operatorname{sinc}^{2k}(y/(2k))$. By elementary calculations, it follows that

$$\psi_k^*(t) = 2k B_{2k}\left(\frac{t}{2k}\right),$$

where $B_r = \underbrace{J \star \cdots \star J}_{r \text{ times}}$ where $J = \frac{1}{2} I_{[-1,1]}$. The following properties of ψ_k and ψ_k^* follow easily:

(1) The support of ψ_k^* is $[-1, 1]$.
(2) $\psi_k \ge 0$ and $\psi_k^* \ge 0$.
(3) $\int \psi_k^*(t)\, dt = \psi_k(0) = 1$.
(4) ψ_k^* and ψ_k are spherically symmetric.
(5) $|\psi_k(y)| \le 1/((2k)^{2k}|y|^{2k})$ for all $|y| > \pi/(2k)$.

Abusing notation somewhat, when u is a vector, we take $\psi_k(u) \equiv \psi_k(\|u\|)$.

Define

$$\widehat{g}^*(t) = \frac{\widehat{q}^*(t)}{\phi^*(t)} \psi_k^*(ht), \tag{33}$$

where $\widehat{q}^*(t) = \frac{1}{n}\sum_{i=1}^n e^{-it^T Y_i}$ is the empirical characteristic function. Now define

$$\widehat{g}(y) = \left(\frac{1}{2\pi}\right)^D \int e^{-it^T y} \frac{\psi_k^*(ht)\widehat{q}^*(t)}{\phi^*(t)}\, dt. \tag{34}$$

Let $\overline{g}(y) = \mathbb{E}(\widehat{g}(y))$.

LEMMA 10. *For all $y \in \mathbb{R}^D$,*

$$\overline{g}(y) = \left(\frac{1}{2\pi h}\right)^D \int \psi_k\left(\frac{\|y-u\|}{h}\right) dG(u).$$

PROOF. Let $\psi_{k,h}(x) = h^{-D}\psi_k(x/h)$. Hence, $\psi^*_{k,h}(t) = \psi^*_k(th)$. Now,

$$\begin{aligned}
\overline{g}(y) &= \left(\frac{1}{2\pi}\right)^D \int e^{-it^T y} \frac{\psi^*_k(th)q^*(t)}{\phi^*(t)}\,dt \\
&= \left(\frac{1}{2\pi}\right)^D \int e^{-it^T y} \frac{\psi^*_k(th)g^*(t)\phi^*(t)}{\phi^*(t)}\,dt \\
&= \left(\frac{1}{2\pi}\right)^D \int e^{-it^T y} \psi^*_k(th)g^*(t)\,dt \\
&= \left(\frac{1}{2\pi}\right)^D \int e^{-it^T y} \psi^*_{k,h}(t)g^*(t)\,dt = \left(\frac{1}{2\pi}\right)^D \int e^{-it^T y} (g \star \psi_{k,h})^*(t)\,dt \\
&= \left(\frac{1}{2\pi}\right)^D (g \star \psi_{k,h})(y) = \left(\frac{1}{2\pi}\right)^D \int \psi_{k,h}(y-u)\,dG(u) \\
&= \frac{1}{h^D}\left(\frac{1}{2\pi}\right)^D \int \psi_k\left(\frac{y-u}{h}\right) dG(u). \qquad \square
\end{aligned}$$

LEMMA 11. *We have that* $\inf_{y \in M \cap \mathcal{K}} \overline{g}(y) \ge C' h^{d-D}$.

PROOF. Choose any $x \in M \cap \mathcal{K}$ and let $B = B(x, Ch)$. Note that $G(B) \ge b(\mathcal{M})ch^d$. Hence,

$$\begin{aligned}
\overline{g}(x) &= (2\pi)^{-D} h^{-D} \int \psi_k\left(\frac{x-u}{h}\right) dG(u) \\
&\ge (2\pi)^{-D} h^{-D} \int_B \psi_k\left(\frac{x-u}{h}\right) dG(u) \\
&\ge (2\pi)^{-D} h^{-D} G(B) = C' h^{d-D}. \qquad \square
\end{aligned}$$

LEMMA 12. *Fix* $0 < \delta < 1/2$. *Suppose that* $k \ge d/(2\delta)$. *Then,*

$$\sup\{\overline{g}(y) : y \in \mathcal{K}, d(y, M) > Lh^{1-\delta}\} \le C'' L^{-2k} \left(\frac{1}{h}\right)^{D-d}. \tag{35}$$

PROOF. Let y be such that $d(y, M) > Lh^{1-\delta}$. For integer $j \ge 1$, define

$$A_j = \bigl[B\bigl(y, (j+1)Lh^{1-\delta}\bigr) - B\bigl(y, jLh^{1-\delta}\bigr)\bigr] \cap M \cap \mathcal{K}.$$

Then

$$\begin{aligned}
\overline{g}(y) &= \left(\frac{1}{2\pi h}\right)^D \int \psi_k\left(\frac{\|u-y\|}{h}\right) dG(u) \\
&\le \left(\frac{1}{2\pi h}\right)^D \sum_{j=1}^{\infty} \int_{A_j} \psi_k\left(\frac{\|u-y\|}{h}\right) dG(u)
\end{aligned}$$

$$\leq \left(\frac{1}{2\pi h}\right)^D \sum_j \int_{A_j} \left(\frac{2kh}{\|u-y\|}\right)^{2k} dG(u)$$

$$\leq C\left(\frac{1}{h}\right)^D \sum_j \int_{A_j} \left(\frac{h}{jLh^{1-\delta}}\right)^{2k} dG(u)$$

$$\leq C\left(\frac{1}{h}\right)^D L^{-2k} h^{2k\delta} \sum_j \left(\frac{1}{j}\right)^{2k} G(A_j)$$

$$\leq C\left(\frac{1}{h}\right)^D L^{-2k} h^{2k\delta} \tag{*}$$

$$\leq C\left(\frac{1}{h}\right)^D L^{-2k} h^{d} \tag{**}$$

$$\leq C'' L^{-2k} \left(\frac{1}{h}\right)^{D-d},$$

where equation (*) follows because G is a probability measure and $\sum_j j^{-2k} < \infty$, and equation (**) follows because $2k\delta \geq d$. $\square$

Now define $\Gamma_n = \sup_y |\widehat{g}(y) - \overline{g}(y)|$.

LEMMA 13. *Let $h = 1/\sqrt{\log n}$, and let $\xi > 1$. Then, for large n,*

$$\Gamma_n = \left(\frac{1}{\sqrt{\log n}}\right)^{4k+4-D} \tag{36}$$

on an event $\mathcal{A}_n$ of probability at least $1 - n^{-\xi}$.

PROOF. We proceed as in Theorem 2.3 of Stefanski (1990). Note that

$$\widehat{g}(y) - \overline{g}(y) = \left(\frac{1}{2\pi}\right)^D \int e^{-it^T y} \frac{\psi_k^*(th)}{\phi^*(t)} \big(\widehat{q}^*(t) - q^*(t)\big)\, dt, \tag{37}$$

and also note that the integrand is 0 for $\|t\| > 1/h$. So

$$\sup_y |\widehat{g}(y) - \overline{g}(y)| \leq \frac{\Delta_n}{(2\pi)^D} \left| \int_{\|t\|\leq 1/h} \frac{\psi_k^*(th)}{\phi^*(t)}\, dt \right|, \tag{38}$$

where $\Delta_n = \sup_{\|t\|<1/h} |\widehat{q}^*(t) - q^*(t)|$.

For $D = 1$, it follows from Theorem 4.3 of Yukich (1985) that

$$Q^n(\Delta_n > 4\varepsilon) \leq 4N(\varepsilon) \exp\left(-\frac{n\varepsilon^2}{8 + 4\varepsilon/3}\right) + 8N(\varepsilon) \exp\left(-\frac{n\varepsilon}{96}\right), \tag{39}$$

where $N(\varepsilon)$ is the bracketing number of the set of complex exponentials, which is given by $N(\varepsilon) = 1 + \frac{24 M_\varepsilon T_n}{\varepsilon}$, and M_ε is defined by $Q(\|Y\| > M_\varepsilon) \le \varepsilon/4$. By a similar argument, we have that in D dimensions,

$$\sup_{Q\in\mathcal{Q}_n} Q^n(\Delta_n > 4\varepsilon) \le 4N(\varepsilon)\exp\left(-\frac{n\varepsilon^2}{8+4\varepsilon/3}\right) + 8N(\varepsilon)\exp\left(-\frac{n\varepsilon}{96}\right), \tag{40}$$

where now

$$N(\varepsilon) = C\left[1 + \frac{24 M_\varepsilon T_n}{\varepsilon}\right]\varepsilon^{-(D-1)}, \tag{41}$$

and M_ε is defined by $\sup_{Q\in\mathcal{Q}_n} Q(\|Y\| > M_\varepsilon) \le \varepsilon/4$. Note that $M_\varepsilon = O(1)$. It follows that $\Delta_n \le \sqrt{\frac{C\log n}{n}}$ except on a set of probability $n^{-\xi}$ where ξ can be made arbitrarily large by taking C large.

Now, note that $\psi_k^*(ht)/\phi^*(t)$ is a spherically symmetric function $R(\|t\|)$. Hence,

$$\int_{\|t\|\le 1/h} \frac{\psi_k^*(ht)}{\phi^*(t)}\,dt = C\int_{s=0}^{1/h} R(s)s^{D-1}\,ds \le Ch^{4k+4-D}e^{1/(2h^2)},$$

where the last result follows from Lemma 3.1 in Stefanski (1990) using parameters $\delta = 2$, $\gamma = 1/2$, $r = 2k+2$, $\beta = D-1$, with $\lambda = h$. The value of r follows from the definition of ψ_k^*. The result now follows by combining this bound with (38). $\square$

Now we can complete the proof of the upper bound.

PROOF OF THEOREM 9. On the event $\mathcal{A}_n$ where $\Gamma_n \le (1/\sqrt{\log n})^{4k+4-D}$ (defined in the previous lemma), we have

$$\begin{aligned}\inf_{y\in M\cap\mathcal{K}} \widehat{g}(y) &\ge \inf_{y\in M\cap\mathcal{K}} \overline{g}(y) - \Gamma_n \ge C\left(\frac{1}{h}\right)^{D-d} - \left(\frac{1}{\sqrt{\log n}}\right)^{4k+4-D}\\ &\ge (C/2)\left(\frac{1}{h}\right)^{D-d} > \lambda_n.\end{aligned}$$

This implies that $M\cap\mathcal{K} \subset \widehat{M}\cap\mathcal{K}$

Next, we have

$$\begin{aligned}\sup_{\substack{y\in\mathcal{K}\\ d(y,M)\ge Lh^{1-\delta}}} \widehat{g}(y) &\le \sup_{\substack{y\in\mathcal{K}\\ d(y,M)\ge Lh^{1-\delta}}} \overline{g}(y) + \Gamma_n\\ &\le CL^{-2k}\left(\frac{1}{h}\right)^{D-d} + \left(\frac{1}{\sqrt{\log n}}\right)^{4k+4-D}\\ &\le 2CL^{-2k}\left(\frac{1}{h}\right)^{D-d} < \lambda_n\end{aligned}$$

for large enough L. This implies that

$$\{y : y \in \mathcal{K} \text{ and } d(y, M) \geq Lh^{1-\delta}\} \cap \widehat{M} = \varnothing.$$

Therefore, on $\mathcal{A}_n$, $L(M, \widehat{M}) \leq C(\frac{1}{\log n})^{(1-\delta)/2}$ and hence,

$$\begin{aligned}\mathbb{E}(L(M, \widehat{M})) &= \mathbb{E}(L(M, \widehat{M})1_{\mathcal{A}_n}) + \mathbb{E}(L(M, \widehat{M})1_{\mathcal{A}_n^c}) \\ &\leq C\left(\frac{1}{\log n}\right)^{(1-\delta)/2} + Q^n(\mathcal{A}_n^c) \\ &\leq C\left(\frac{1}{\log n}\right)^{(1-\delta)/2} + n^{-\xi} \leq C\left(\frac{1}{\log n}\right)^{(1-\delta)/2},\end{aligned}$$

and the theorem is proved. □

REMARK. Again, the proof of the upper bound is similar to proofs used in other deconvolution problems. But once more, there are interesting differences. In particular, the density estimator $\widehat{g}$ is not estimating any underlying density since the measure G is singular and thus does not have a density. Hence, the usual bias calculation is meaningless.

REMARK. Note that $\widehat{M}$ is a set not a manifold; if desired, we can replace $\widehat{M}$ with any manifold in $\{M \in \mathcal{M} : M \subset \widehat{M}\}$, and then the estimator is a manifold and the rate is the same.

REMARK. The upper bound is slightly slower than the lower bound. The rate is consistent with the results in Caillerie et al. (2011) who show that $\mathbb{E}(W_2(\widehat{g}, G)) \leq C/\sqrt{\log n}$ where W_2 is the Wasserstein distance. In the special case where the manifold has the form $\{(u, m(u)) : u \in \mathbb{R}\}$ for some function m, the problem can be viewed as nonparametric regression with measurement error; see Section 6. In this special case, we can use the deconvolution kernel regression estimator in Fan and Truong (1993) which achieves the rate $1/\log n$. We do not know of any estimator in the general case that achieves the rate $1/\log n$, although we conjecture that the following estimator might have a better rate: let $(\widehat{M}, \widehat{G})$ minimize $\sup_{\|t\| \leq T_n} |\widehat{q}^*(y) - q^*_{M,G}(t)|$ where $T_n = O(\sqrt{\log n})$. In any case, as with all Gaussian deconvolution problems, the rate is very slow, and the difference between $1/\log n$ and $1/\sqrt{\log n}$ is not of practical consequence.

6. Singular deconvolution. Estimating a manifold under additive noise is related to deconvolution. It is also related to regression with errors in variables. The purpose of this section is to explain the connections between the problems.

6.1. *Relationship to density deconvolution.* Recall that the model is $Y = X + Z$ where $X \sim G$, G is supported on a manifold M and $Z \sim \Phi$. G is a singular measure supported on the d-dimensional manifold M.

Now consider a somewhat simpler model: suppose again that $Y_i = X_i + Z_i$, but suppose that X has a density g on $\mathbb{R}^D$ (instead of being supported on a manifold). All three distributions Q, G and Φ have D-dimensional support and $Q = G \star \Phi$. The problem of recovering the density g of X from $Y_1, \dots, Y_n$ is the usual density deconvolution problem. A key reference is Fan (1991).

Most of the existing literature on deconvolution assumes that X and Y have the same support, or at least that the supports have the same dimension; an exception is Koltchinskii (2000). Manifold learning may be regarded as the problem of deconvolution for singular measures.

It is instructive to compare the least favorable pair used for proving the lower bounds in the ordinary case versus the singular case. Figure 4 shows a typical least favorable pair for proving a lower bound in ordinary deconvolution. The top left plot is a density g_0, and the top right plot is a density g_1 which is a perturbed version of g_0. The L_1 distance between the densities is ε. The bottom plots are $q_0 = \int \phi(y-x)g_0(x)\,dx$ and $q_1 = \int \phi(y-x)g_1(x)\,dx$. These densities are nearly indistinguishable, and, in fact, their total variation distance is of order $e^{-1/\varepsilon}$. Of course, these distributions have the same support and hence such a least favorable pair will not suffice for proving lower bounds in the manifold case where we will need two densities with different support.

Figure 5 shows the type of least favorable pair we used for manifold learning. The top two plots do not show the densities; rather they show the support of the densities. The distribution g_0 is uniform on the circle in the top left plot. The distribution g_1 is uniform on the perturbed circle in the top right plot. The

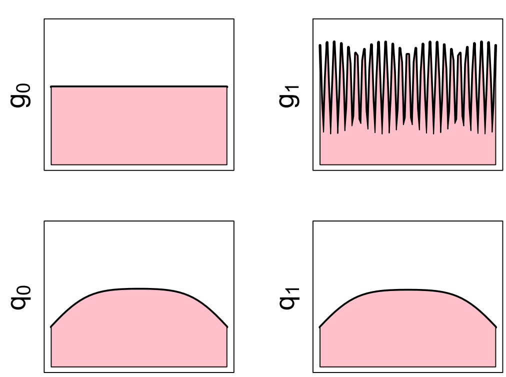

FIG. 4. *A typical least favorable pair for proving a lower bounds in ordinary deconvolution. The top left plot is a density g_0 and the top right plot is a density g_1 which is a perturbed version of g_0. The L_1 distance between the densities is ε. The bottom plots are $q_0 = \int \phi(y-x)g_0(x)\,dx$ and $q_1 = \int \phi(y-x)g_1(x)\,dx$. These densities are nearly indistinguishable and, in fact, their total variation distance is $e^{-1/\varepsilon}$.*

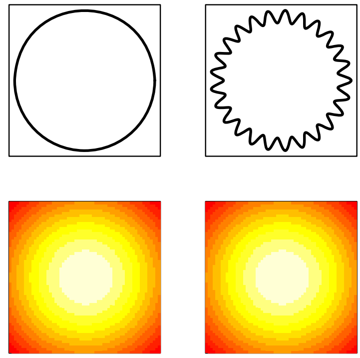

FIG. 5. *The type of least favorable pair needed for proving lower bounds in manifold learning. The distribution g_0 is uniform on the circle in the top left plot. The distribution g_1 is uniform on the perturbed circle in the top right plot. The Hausdorff distance between the supports of the densities is ε. The bottom plots are heat maps of $q_0 = \int \phi(y-x)g_0(x)\,dx$ and $q_1 = \int \phi(y-x)g_1(x)\,dx$. These densities are nearly indistinguishable and, in fact, their total variation distance is $e^{-1/\varepsilon}$.*

Hausdorff distance between the supports of densities is ε. The bottom plots are $q_0 = \int \phi(y-x)g_0(x)\,dx$ and $q_1 = \int \phi(y-x)g_1(x)\,dx$. Again, these densities are nearly indistinguishable, and, in fact, their total variation distance is $e^{-1/\varepsilon}$. In this case, however, g_0 and g_1 have different supports.

6.2. *Relationship to regression with measurement error.* We can also relate the manifold estimation problem with nonparametric regression with measurement error. Suppose that

$$\begin{aligned} U_i &= X_i + Z_{2i}, \\ Y_i &= m(X_i) + Z_{1i}, \end{aligned} \tag{42}$$

and we want to estimate the regression function m. If we observe $(X_1, Y_1), \dots, (X_n, Y_n)$, then this is a standard nonparametric regression problem. But if we only observe $(U_1, Y_1), \dots, (U_n, Y_n)$, then this is the usual nonparametric regression with measurement error problem. The rates of convergence are similar to deconvolution. Indeed, Fan and Truong (1993) have an argument that converts nonparametric regression with measurement error into a density deconvolution problem. Let us see how this related to manifold learning.

Suppose that $D = 2$ and $d = 1$. Futher, suppose that the manifold is *function-like*, meaning that the manifold is a curve of the form $M = \{(u, m(u)) : u \in \mathbb{R}\}$ for some function m. Then each Y_i can be written in the form

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} U_i \\ m(U_i) \end{pmatrix} + \begin{pmatrix} Z_{1i} \\ Z_{2i} \end{pmatrix}$$

which is exactly of the form (42). Let $\mathcal{Q}$ be all such distributions obtained this way with $|m''(u)| \le 1/\kappa$. However, this only holds when the manifold has the function-like form. Moreover, the lower bound argument in Fan and Truong (1993) cannot directly be transferred to the manifold setting as we now explain.

In our lower bound proof, we defined a least favorable pair q_0 and q_1 for the distribution of Y as follows. Take $M_0 = \{(u, 0) : u \in \mathbb{R}\}$ and $M_1 = \{(u, m(u)) : u \in \mathbb{R}\}$. [In fact, we used $(u, m(u))$ and $(u, -m(u))$, but the present discussion is clearer if we use $(u, 0)$ and $(u, m(u))$.] Let $Y = (Y_1, Y_2)$. For M_0, the distribution q_0 for Y is based on

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} U \\ 0 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}.$$

The density of (U, Y_2) is $f_0(u, y_2) = \zeta(u)\phi(y_2)$ where ζ is some density for U. Then

$$q_0(y_1, y_2) = f_0 \star \Phi = \int f_0(y_1 - Z_1, y_2)\, d\Phi(z_1),$$

where the convolution symbol here and in what follows, refers to convolution only over $U + Z_1$.

Now let $q_1(y_1, y_2)$ denote the distribution of Y in the model

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} U \\ m(U) \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}.$$

This generates the least favorable pair q_0, q_1 used in our proof (restricted to this special case).

The least favorable pair used by Fan and Truong is different in a subtle way. The first distribution q_0 is the same. The second, which we will denote w_1, is constructed as follows. Let

$$w_1(y_1, y_2) = f_1 \star \Phi,$$

where the convolution is only over U,

$$f_1(\xi, y_2) = f_0(\xi, y_2) + \gamma H(\xi/\sqrt{\gamma}) h_0(y_2),$$

where $f_1(\xi) = g(\xi)$, $\gamma H(\xi/\sqrt{\gamma})/g(\xi) = b(\xi)$, H is a perturbation function such as a cosine, and h_0 is chosen so that $\int h_0(y_2)\, dy_2 = 0$ and $\int y_2 h_0(y_2)\, dy_2 = 1$. Now we show that $w_1(y_1, y_2) \neq q_1(y_1, y_2)$. In fact, w_1 is not in $\mathcal{Q}$. Note that

$$w_1(y_1, y_2) = f_1 \star \Phi = q_0(y_1, y_2) + \gamma h_0(y_2) \int H\left(\frac{y_1 - z_1}{\sqrt{\gamma}}\right) d\Phi(z_1).$$

Now,

$$q_1(y_2|u) = \phi\big(y_2 - m(u)\big),$$

but

$$f_1(y_2|u) = \frac{f_1(y_2, u)}{f_1(u)} = \phi(y_2) + m(u) h_0(y_2).$$

These both have mean $m(u)$ but the distributions are different. Indeed, the marginals $w_1(y_2)$ and $q_1(y_2)$ are different. In fact,

$$w_1(y_2) = q_0(y_2) + ch_0(y_2)$$

for some c. This is not in our class because it is not of the form $\phi(y_2 - m(u))$. Hence, w_1 is not in our class $\mathcal{Q}$: it does not correspond to drawing a point on a manifold and adding noise.

The point is that manifold learning reduces to nonparametric regression with errors only in the special case that the manifold is function-like. And even in this case, the proofs of the bounds are somewhat different than the usual proofs.

7. Discussion. The purpose of this paper is to establish minimax bounds on estimating manifolds. The estimators used to prove the upper bounds are theoretical constructions for the purposes of the proofs. They are not practical estimators.

There is a large literature on methodology for estimating manifolds. However, these estimators are not likely to be optimal except under stringent conditions. In current work we are trying to bridge the gap between the theory and the methodology.

Probably the most realistic noise condition is the additive model. In this case, we are dealing with a singular deconvolution problem. The upper bound used deconvolution techniques. Such methods require that the noise distribution is known (or is at least restricted to some narrow class of distributions). This seems unrealistic in real problems. A more realistic goal is to estimate some proxy manifold M^* that, in some sense, approximates M. We are currently working on such techniques.

REFERENCES

ARIAS-CASTRO, E., DONOHO, D. L., HUO, X. and TOVEY, C. A. (2005). Connect the dots: How many random points can a regular curve pass through? *Adv. in Appl. Probab.* **37** 571–603. MR2156550

BARANIUK, R. G. and WAKIN, M. B. (2009). Random projections of smooth manifolds. *Found. Comput. Math.* **9** 51–77. MR2472287

BOUSQUET, O., BOUCHERON, S. and LUGOSI, G. (2004). Introduction to statistical learning theory. *Machine Learning* **3176** 169–207.

CAILLERIE, C., CHAZAL, F., DEDECKER, J. and MICHEL, B. (2011). Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Stat.* **5** 1394–1423. MR2851684

CHAUDHURI, K. and DASGUPTA, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 23* (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta) 343–351.

DEY, T. K. (2007). *Curve and Surface Reconstruction: Algorithms with Mathematical Analysis. Cambridge Monographs on Applied and Computational Mathematics* **23**. Cambridge Univ. Press, Cambridge. MR2267420

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272. MR1126324

FAN, J. and TRUONG, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21** 1900–1925. MR1245773

FEDERER, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2010). Minimax manifold estimation. Available at arXiv:1007.0549.

KOLTCHINSKII, V. I. (2000). Empirical geometry of multivariate data: A deconvolution approach. *Ann. Statist.* **28** 591–629. MR1790011

MEISTER, A. (2006). Estimating the support of multivariate densities under measurement error. *J. Multivariate Anal.* **97** 1702–1717. MR2298884

NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. MR2383768

NIYOGI, P., SMALE, S. and WEINBERGER, S. (2011). A topological view of unsupervised learning from noisy data. *SIAM J. Comput.* **40** 646–663.

OZERTEM, U. and ERDOGMUS, D. (2011). Locally defined principal curves and surfaces. *J. Mach. Learn. Res.* **12** 1249–1286. MR2804600

STEFANSKI, L. A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statist. Probab. Lett.* **9** 229–235. MR1045189

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963

YUKICH, J. E. (1985). Laws of large numbers for classes of functions. *J. Multivariate Anal.* **17** 245–260. MR0813235

C. R. GENOVESE
L. WASSERMAN
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: genovese@stat.cmu.edu
larry@stat.cmu.edu

M. PERONE-PACIFICO
DEPARTMENT OF STATISTICAL SCIENCES
SAPIENZA UNIVERSITY OF ROME
ROME
ITALY
E-MAIL: marco.peronepacifico@uniroma1.it

I. VERDINELLI
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
AND
DEPARTMENT OF STATISTICAL SCIENCES
SAPIENZA UNIVERSITY OF ROME
ROME
ITALY
E-MAIL: isabella@stat.cmu.edu