Electronic Journal of Statistics
Vol. 9 (2015) 1562–1582
ISSN: 1935-7524
DOI: 10.1214/15-EJS1039

Tight minimax rates for manifold estimation under Hausdorff loss

Arlene K. H. Kim

University of Cambridge
e-mail: a.kim@statslab.cam.ac.uk

and

Harrison H. Zhou[*]

Yale University
e-mail: huibin.zhou@yale.edu

Abstract: This paper deals with minimax rates of convergence for manifold estimation. A new lower bound is obtained by a novel construction of two sets of manifolds and an application of convex hull testing method of Le Cam (1973). The minimax lower bound matches the upper bound up to a constant factor considered by Genovese et al. (2012b).

MSC 2010 subject classifications: Primary 62C25, 62G86; secondary 65C50.
Keywords and phrases: minimax lower bound, manifold estimation, convex hull testing.

Received November 2013.

1. Introduction

We observe an i.i.d. random sample $Y_1, \ldots, Y_n \in \mathbb{R}^D$ from a distribution Q that lies on a d dimensional manifold M with $d < D$. The goal is to estimate the unknown manifold M based on the sample $\{Y_i\}$.

Manifold learning is an active area of research in machine leaning, applied mathematics as well as statistics, but not much optimality theory regarding the rates of convergence has been developed. To the best of our knowledge the optimal convergence rates for estimating manifolds are only considered by Genovese et al. (2012a,b) (henceforth GPVW) under a minimax criterion. Convergence rates of their theoretical estimators are compared to the lower bounds. However, their upper and lower bounds do not match. To fill in the gap, this paper establishes the optimal rates of convergence by a novel lower bound argument.

GPVW considered three noise models–noiseless, clutter and additive model–for Q. Noiseless model assumes that sample is obtained from a distribution G

[*]The research of Harrison Zhou was supported in part by NSF Career Award DMS-0645676, NSF FRG Grant DMS-0854975 and NSF Grant DMS-1209191.

supported on M. The clutter model assumes that sample is from G with probability π and from U with probability $1-\pi$, where U is a uniform distribution on a compact set $\mathcal{K} \subset \mathbb{R}^D$ with nonempty interior. We only consider the noiseless model.

Following GPVW we measure the loss by means of the Hausdorff distance

$$H(A,B) = \inf\{\epsilon > 0 : A \subseteq B \oplus \epsilon \quad \text{and} \quad B \subseteq A \oplus \epsilon\}$$

where $A \oplus \epsilon = \{x \in \mathbb{R}^D : ||x - A|| \leq \epsilon\}$ where $||\cdot||$ denotes Euclidean distance. For a suitably chosen set $\mathcal{M}$ of compact, d-dimensional manifolds in $\mathbb{R}^D$ and, for each $\mathsf{M} \in \mathcal{M}$, a suitable set $\mathbb{Q}(\mathsf{M})$ of probability measures concentrated on M (see Section 2), they worked with the the maximum expected loss

$$\Lambda_n(\widehat{\mathsf{M}}) := \sup_{\mathsf{M}\in\mathcal{M}} \sup_{Q\in\mathbb{Q}(\mathsf{M})} \mathbb{E}_Q H(\widehat{\mathsf{M}}_n, \mathsf{M})$$

based on the sample $Y_1, \ldots, Y_n$ from Q. They constructed a sequence of estimators for which

$$\Lambda_n(\widehat{\mathsf{M}}) = O\left(\gamma_n^{2/d}\right) \qquad \text{where } \gamma_n = n^{-1}\log n$$

and also showed that

$$\inf_{\widehat{\mathsf{M}}} \Lambda_n(\widehat{\mathsf{M}}) \geq c\gamma_n^{2/d}(\log n)^{-2/d} = cn^{-2/d} \tag{1}$$

for some constant c that depended on $\mathcal{M}$. The proof of the lower bound is based on the testing between two smooth manifolds where the first manifold looks like a squashed ball with a flat region and the second manifold coincides with the first one except a small bump on the flat area, such that these two manifolds are not statistically distinguished.

The main contribution of our paper (Theorem 1 in Section 2) is to establish a uniform lower bound for $\Lambda_n(\widehat{\mathsf{M}})$ of order $\gamma_n^{2/d}$, thereby determining the true minimax rate for one of the manifold estimation problems considered by GPVW.

We use a method, first presented by Le Cam (1973), in the form used by Yu (1997). In our setting the method becomes: if $\mathcal{M}_0$ and $\mathcal{M}_1$ are subsets of $\mathcal{M}$ for which $\inf\{H(\mathsf{M}_0, \mathsf{M}_1) : \mathsf{M}_0 \in \mathcal{M}_0, \mathsf{M}_1 \in \mathcal{M}_1\} \geq 2\gamma$, for some positive constant γ, then

$$\Lambda_n(\widehat{\mathsf{M}}) \geq \gamma \sup_{\mathbb{P}_i\in\text{co}(\mathbb{Q}_i^n)} |\mathbb{P}_0 \wedge \mathbb{P}_1| \tag{2}$$

where $\mathbb{Q}_i^n$ denotes the set of all n-fold product measures Q_i^n with $Q_i \in \mathbb{Q}(\mathcal{M}_i)$ for $i = 1, 2$ and $\text{co}(\cdot)$ denotes the convex hull. The quantity $|\mathbb{P}_0 \wedge \mathbb{P}_1|$ is called the testing affinity. It equals 1 minus half of the L_1 distance between $\mathbb{P}_0$ and $\mathbb{P}_1$.

The general analog of inequality (2) is best known for the case where both $\mathcal{M}_0$ and $\mathcal{M}_1$ are singleton subsets, a situation sometimes referred to as "two-point testing". GPVW used that method to establish their lower bound (1). The situation where only one of the $\mathcal{M}_i$'s is a singleton set has been used by many authors (see Tsybakov 2009, Section 2.7.5 for example). The full power of inequality (2) has, to our knowledge, only been effectively applied in the paper by Cai and Low (2011). In the present paper we employ (2) by means of an embedding of an occupancy problem into the manifold problem. We bound the affinity bewteen two convex hulls (mixtures) using combinatorial arguments involving the hypergeometric distribution. In Section 2, we give some intuition behind our construction and the reasons why it does not suffice to have even one of the $\mathcal{M}_i$ a singleton set.

Proofs for the main results are given in Section 3. We collect proofs for auxiliary lemmas in Section 4.

2. Main theorems

In this section we introduce lower bound results and intuitions behind them. The lower bounds, together with upper bounds in Genovese et al. (2012b), yield minimax rates of convergence.

First, we describe the setting in detail. Assume that M is a compact C^1 Riemannian submanifold without boundary in $\mathbb{R}^D$, and contained in some compact set $\mathcal{K} \subset \mathbb{R}^D$ with nonempty interior. In addition, we need a regularity condition for the curvature of the manifold M. Define $\Delta(\mathsf{M})$ to be the largest r such that each point in $\mathsf{M} \oplus r$ has a unique projection onto M. As proved by Niyogi et al. (2006, Section 6), $\Delta(\mathsf{M})$, which is usually called condition number, controls the curvature of the manifold. We assume that the d dimensional manifold M satisfies $\Delta(\mathsf{M}) \geq \kappa$, where κ is a fixed positive constant. Let $\mathcal{M}(\kappa) := \{\mathsf{M} : \Delta(\mathsf{M}) \geq \kappa, \mathsf{M} \subseteq \mathcal{K}\}$, and $\mathcal{G}(\mathsf{M})$ be a set of distributions Q whose densities q with respect to the uniform measure on M satisfy

$$0 < b(\mathcal{M}(\kappa)) \leq \inf_{y \in \mathsf{M}} q(y) \leq \sup_{y \in \mathsf{M}} q(y) \leq B(\mathcal{M}(\kappa)) < \infty \tag{3}$$

where $b(\mathcal{M}(\kappa))$ and $B(\mathcal{M}(\kappa))$ may depend on $\mathcal{M}(\kappa)$ but not on the particular manifold M.

Here is the intuition behind our main result, Theorem 1. Consider a one ($d = 1$) dimensional closed smooth curve M in a two dimensional space ($D = 2$), and observe n points uniformly distributed on M. Without loss of generality, assume that the length of the curve is 1. The maximum gap among those points is of an order of $\frac{\log n}{n}$ with high probability. That means there exists at least one connected piece of the manifold of length of order $\frac{\log n}{n}$ on which we have no observations. The locally quadratic approximation error of the interpolation of smooth manifold yields a possibly unavoidable estimation error of $(\frac{\log n}{n})^2$. This idea can be carried over to a general d dimensional manifold in $\mathbb{R}^D$ with $d < D$ by dividing the manifold into an order of $\frac{n}{\log n}$ disjoint pieces with a diameter of an order $(\frac{\log n}{n})^{1/d}$ for each.

Theorem 1. *Let $Y_1, \ldots, Y_n$ be i.i.d. sample from a distribution Q where Q is supported on a manifold $\mathsf{M} \in \mathcal{M}(\kappa)$, and $Q \in \mathcal{G}(\mathsf{M})$. Then there is a constant c,* *not depending on n but depending on d, and b and B defined in Equation (3), such that*

$$\inf_{\hat{\mathsf{M}}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}[H(\hat{\mathsf{M}}, \mathsf{M}(Q))] \geq c \left(\frac{\log n}{n}\right)^{2/d}.$$

The following two remarks explain why a naive "one versus a mixture" testing can not lead to the desired lower bound in Theorem 1 and the intuition for the testing of two convex hulls. We use $a_n \asymp b_n$ if $a_n \leq C_1 b_n$ and $b_n \leq C_2 a_n$ where C_1, C_2 are constants not depending on n. Denote by $L_1(P,Q) := \int |p-q| d\mu$ the L_1 distance between P and Q, where μ is a dominating measure, and p and q are the densities of P and Q respectively.

Remark 1 (One versus a mixture)**.** In many cases including sparse support recovery in high dimensional estimation, one (null) versus a mixture (alternative) testing gives tight bounds. For instance, consider the problem of estimating the multivariate standard normal mean vector $\theta \in \mathbb{R}^n$ with covariance matrix I_n/n. Then it can be shown that for the parameter space satisfying $\sum_{i=1}^n \mathbb{1}\{\theta_i \neq 0\} = 1$, the magnitude of the nonzero θ_i needs to be at least of an order of $\sqrt{\frac{\log n}{n}}$ for consistent support recovery by one versus a mixture testing of Le Cam. However, this same reasoning does not work for manifold estimation by simply considering a test of one manifold versus a mixture of many manifolds with one bump for each. Consider a base manifold M_0 (defined in (14)) as the null. For the alternatives, we can construct the set of manifolds $\mathcal{M}$ having one bump deviated from M_0 such that $H(\mathsf{M}_0, \mathsf{M}) \asymp (\frac{\log n}{n})^{2/d}$ for all $\mathsf{M} \in \mathcal{M}$. Define the uniform distributions $Q_0 := U(\mathsf{M}_0)$ and $Q := U(\mathsf{M}) \in \mathcal{Q}$ on these manifolds. We find that $L_1(Q_0^n, \frac{1}{|\mathcal{Q}|}\sum_{Q \in \mathcal{Q}} Q^n)$ converges to 2. These two distributions are very different from each other. This can be understood as follows. Based on one sample from any manifold of $\mathcal{M}$, with high probability there is at least one observation lying on the bump, thus we instantly know that the null hypothesis is wrong. This is different from the multivariate normal mean estimation problem, for which based on a sample generated from the alternative one can not tell whether there is a nonzero θ_i and where the location of the nonzero θ_i is. See Remark 3 for the exact calculation of the L_1 distance.

Remark 2 (A mixture versus a mixture)**.** From Remark 1, we see that the problem of estimating a set of manifolds with at most one bump for each is not hard enough for establishing the desired lower bound. A large set of manifolds with $2m = n/(t \log n)$, for some $t \in (0, 1/2)$, number of inward and outward bumps are constructed. Two subsets of manifolds $\mathcal{M}_0 = \{\mathsf{M}_{0j}\}$ and $\mathcal{M}_1 = \{\mathsf{M}_{1j'}\}$ are selected. Manifolds in $\mathcal{M}_0$ have $m = n/(2t \log n)$ outward bumps, while manifolds in $\mathcal{M}_1$ have either $m+1$ or $m-1$ outward bumps on M_0, such that $H(\mathsf{M}_{0j}, \mathsf{M}_{1j'}) \asymp (\frac{\log n}{n})^{2/d}$ for all $\mathsf{M}_{0j} \in \mathcal{M}_0$ and $\mathsf{M}_{1j'} \in \mathcal{M}_1$. Consider the uniform distributions on M_{0j}, that is, $Q_{0j} := U(\mathsf{M}_{0j})$, and similarly $Q_{1j'} := U(\mathsf{M}_{1j'})$. In Section 3 it will be shown that $L_1(\frac{1}{|\mathcal{Q}_0|}\sum_{Q_{0j} \in \mathcal{Q}_0} Q_{0j}^n, \frac{1}{|\mathcal{Q}_1|}\sum_{Q_{1j'} \in \mathcal{Q}_1} Q_{1j'}^n)$ converges to 0, which implies we can not distinguish two convex hulls from observations. The intuition behind the L_1 distance calcula-

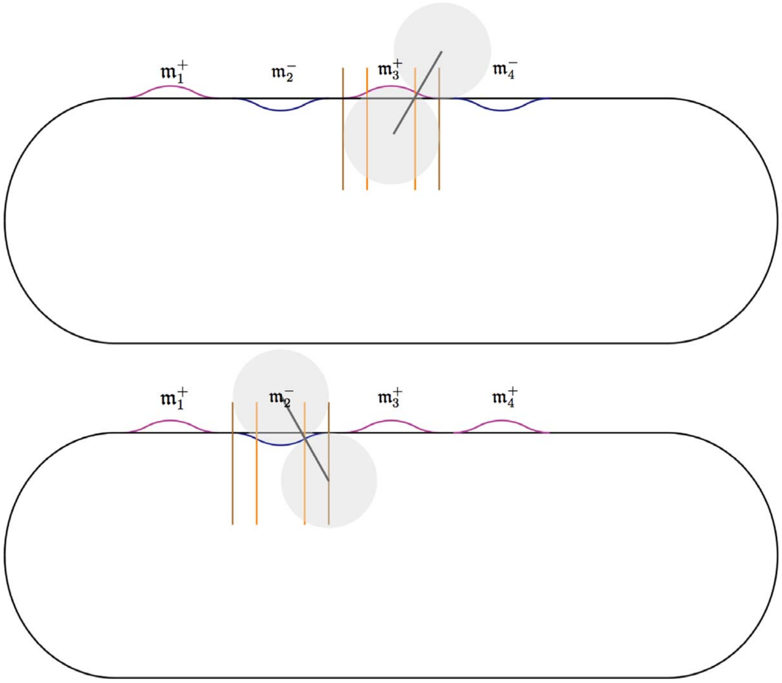

Fig 1. *Constructed manifolds for case $d = 1, m = 2$: M_{0j} (top), $\mathsf{M}_{1j'}$ (bottom). Here we used larger base manifold for illustration.*

tion is that for observations generated from either class there will be several bumps without any observation lying there, thus it seems to be impossible to distinguish one convex hull from the other from those observations. See Section 3 for details of a rigorous justification.

Combining the upper bound in Theorem 3 in Genovese et al. (2012b) and the improved lower bound in Theorem 1, we have the following corollary.

Corollary 2. *Let $Y_1, \dots, Y_n$ be i.i.d. sample from a distribution Q where Q is supported on a manifold $\mathsf{M} \in \mathcal{M}(\kappa)$, and $Q \in \mathcal{G}(\mathsf{M})$. Then*

$$\inf_{\hat{\mathsf{M}}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \left[H(\hat{\mathsf{M}}, \mathsf{M}(Q)) \right] \asymp \left(\frac{\log n}{n} \right)^{2/d} .$$

Theorem 1 can be easily extended to a so called clutter model, for which we observe i.i.d. observations $Y_1, \dots, Y_n \in \mathbb{R}^D$ from a mixture distribution $(1-\pi)U(\mathcal{K}) + \pi G$, where $U(\mathcal{K})$ is a uniform distribution on $\mathcal{K}$ and $\pi \in (0,1)$. Construct the same set of manifolds which are used in the proof of Theorem 1. Clearly the Hausdorff distance between any manifold in each set is at least $(\log n/n)^{2/d}$ (up to a constant) as in the proof of Theorem 1, thus it suffices to show that the testing affinity in (2) is bounded away from zero in the clutter model. See the proof of Theorem 3 for calculation.

Theorem 3. *Let $Y_1, \dots, Y_n$ be i.i.d. sample from a distribution $Q = (1 - \pi)U(\mathcal{K}) + \pi G$, where $G \in \mathcal{G}(\mathsf{M})$ is supported on a manifold $\mathsf{M} \in \mathcal{M}(\kappa)$. Then there is a constant c, not depending on n but depending on d, π, and b and B defined in Equation (3), such that*

$$\inf_{\hat{\mathsf{M}}} \sup_{Q} \mathbb{E}_{Q^n} \left[H(\hat{\mathsf{M}}, \mathsf{M}(Q)) \right] \geq c \left(\frac{\log n}{n} \right)^{2/d}.$$

Combining the upper bound in Theorem 5 by Genovese et al. (2012b), we obtain the following optimal rates of convergence.

Corollary 4. *Let $Y_1, \dots, Y_n$ be i.i.d. sample from a distribution $Q = (1 - \pi)U(\mathcal{K}) + \pi G$, where $G \in \mathcal{G}(\mathsf{M})$ is supported on a manifold $\mathsf{M} \in \mathcal{M}(\kappa)$. Then*

$$\inf_{\hat{\mathsf{M}}} \sup_{Q} \mathbb{E}_{Q^n} \left[H(\hat{\mathsf{M}}, \mathsf{M}(Q)) \right] \asymp \left(\frac{\log n}{n} \right)^{2/d}.$$

3. Proof of main results

In this section we derive the lower bounds in Theorems 1 and 3. The key technique is Le Cam's method for testing two convex hulls. Consider a set of distributions $\mathcal{Q}$, supported on a manifold $\mathsf{M} \in \mathcal{M}$. Let $\hat{\mathsf{M}}$ be the estimator of $\mathsf{M} = \mathsf{M}(Q)$ based on i.i.d. sample $Y_1, \dots, Y_n$ from Q. Let H be the Hausdorff distance. Le Cam (1973) establishes a minimax lower bound as follows. See also Yu (1997).

Lemma 5 (Le Cam's method)**.** *Suppose that there are subsets $\mathcal{M}_0$ and $\mathcal{M}_1$ of $\mathcal{M}$ that are 2γ separated, in the sense that $H(\mathsf{M}_0, \mathsf{M}_1) \geq 2\gamma$ for all $\mathsf{M}_0 \in \mathcal{M}_0$ and $\mathsf{M}_1 \in \mathcal{M}_1$. Suppose that $\mathcal{Q}_0$ and $\mathcal{Q}_1$ are subsets of $\mathcal{Q}$ for which $\mathsf{M}(Q_0) \in \mathcal{M}_0$ for $Q_0 \in \mathcal{Q}_0$ and $\mathsf{M}(Q_1) \in \mathcal{M}_1$ for $Q_1 \in \mathcal{Q}_1$. Denote the cardinality of the set $\mathcal{Q}$ by $|\mathcal{Q}|$. Then*

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} H(\hat{\mathsf{M}}, \mathsf{M}(Q)) \geq \gamma \left| \left(\frac{1}{|\mathcal{Q}_0|} \sum_{Q_0 \in \mathcal{Q}_0} Q_0^n \right) \wedge \left(\frac{1}{|\mathcal{Q}_1|} \sum_{Q_1 \in \mathcal{Q}_1} Q_1^n \right) \right|.$$

The proof of Theorem 1 consists of the following 4 steps: (i) construction of two finite sub-parameter spaces $\mathcal{M}_0$ and $\mathcal{M}_1$ which are separated of an order of $(\frac{\log n}{n})^{2/d}$, (ii) simplification of the L_1 representation, (iii) reduction of the L_1 distance to a combinatorial counting, and finally (iv) bounding the L_1 distance by studying combinatorics through the tail probability bounds for the traditional occupancy problems and hypergeometric distribution.

3.1. Construction of finite sub-parameter spaces

The construction extends the manifold with one bump in Genovese et al. (2012a) to the case of multiple bumps. In particular, $\mathcal{M}_0$ corresponds to the set of manifolds with m outward bumps among $2m$ possible perturbations while $\mathcal{M}_1$ corresponds to the set of manifolds with $m+1$ or $m-1$ outward bumps. These bumps are disjoint and congruent, and the volume of each bump is of order $\log n/n$. Lemma 6 shows that there exist suitable constructions for these sets $\mathcal{M}_0$ and $\mathcal{M}_1$ in order to use Le Cam's method.

Lemma 6. *Let* $2m := n/(t \log n) \asymp \gamma^{-d/2}$ *for some* $t \in (0, 1/2)$*,* $N_0 = \binom{2m}{m}$*, and* $N_1 = \binom{2m}{m+1} + \binom{2m}{m-1}$*. Then there exists two sets of compact* C^1 *Riemannian manifold* $\mathcal{M}_0 = \{\mathsf{M}_{0j}, j = 1, \ldots, N_0\}$ *and* $\mathcal{M}_1 = \{\mathsf{M}_{1j'}, j' = 1, \ldots, N_1\}$ *such that all the manifolds in* $\mathcal{M}_0$ *and* $\mathcal{M}_1$ *satisfy the condition number assumption, that is,* $(\mathcal{M}_0 \cup \mathcal{M}_1) \subseteq \mathcal{M}(\kappa)$*, and*

$$\inf_{j'=1,\ldots,N_1} \inf_{j=1,\ldots,N_0} H(\mathsf{M}_{0j}, \mathsf{M}_{1j'}) \geq 2\gamma.$$

3.2. A simplified L_1 representation

In order to consider the L_1 distance between distributions on these manifolds, we introduce some more notation. Part of manifolds without bumps is denoted as $\mathfrak{m} := \cap_{j=1}^{N_0} \mathsf{M}_{0j} \cap_{j'=1}^{N_1} \mathsf{M}_{1j'}$. Part of manifolds with bumps are denoted with $\mathfrak{m}_l^+$ and $\mathfrak{m}_l^-$ for $l = 1, \ldots, 2m$ where $\mathfrak{m}_l^+$ means the outward bump ($\mathfrak{m}_l^-$ meaning the inward bump) where the order l does not play a crucial role because of the symmetry. For instance, by defining $\mathfrak{m}_{0j} := \cup_{l \in \mathcal{R}_{0j}} \mathfrak{m}_l^+ \cup_{l \in \mathcal{R}_{0j}^c} \mathfrak{m}_l^-$,

$$\mathsf{M}_{0j} = \mathfrak{m} \cup_{l \in \mathcal{R}_{0j}} \mathfrak{m}_l^+ \cup_{l \in \mathcal{R}_{0j}^c} \mathfrak{m}_l^- =: \mathfrak{m} \cup \mathfrak{m}_{0j}. \tag{4}$$

For later use, we also denote $\mathfrak{m}_l := \mathfrak{m}_l^+ \cup \mathfrak{m}_l^-$. Then we suppose μ as a dominating uniform measure on $\cup_{j=1}^{N_0} \mathsf{M}_{0j} \cup_{j'=1}^{N_1} \mathsf{M}_{1j'}$. We let Q_{0j} be the uniform probability measure on M_{0j} (i.e. $Q_{0j} := U(\mathsf{M}_{0j})$) with a density q_{0j} respect to μ and similarly define $Q_{1j'}$ and $q_{1j'}$ on $\mathsf{M}_{1j'}$.

By construction, $\mu(\mathfrak{m}) = C_0$ (where C_0 is a constant only depending on d) and $\mu(\mathfrak{m}_l^+) = \mu(\mathfrak{m}_l^-) = c\gamma^{d/2}$ for all $l = 1, \ldots, 2m$ (see the proof of Theorem 2 by Genovese et al. (2012b)). Then $\mu(\mathsf{M}_{0j}) = \mu(\mathsf{M}_{1j'}) = C_0 + (2m)c\gamma^{d/2} =: C$ for all $j = 1, \ldots, N_0$ and $j' = 1, \ldots, N_1$, where C only depends on the dimension d by the choice of $m \asymp \gamma^{-d/2}$. Accordingly, for $j = 1, \ldots, N_0$ and $j' = 1, \ldots, N_1$,

$$\begin{aligned} \frac{dQ_{0j}}{d\mu}(x) &:= q_{0j}(x) = \frac{1}{C} \mathbb{1}_{\{x \in \mathsf{M}_{0j}\}} = \frac{1}{C} \mathbb{1}_{\{x \in \mathfrak{m} \cup \mathfrak{m}_{0j}\}} \\ \frac{dQ_{1j'}}{d\mu}(x) &:= q_{1j'}(x) = \frac{1}{C} \mathbb{1}_{\{x \in \mathsf{M}_{1j'}\}} = \frac{1}{C} \mathbb{1}_{\{x \in \mathfrak{m} \cup \mathfrak{m}_{1j'}\}}. \end{aligned}$$

Lemma 7 gives an upper bound for the L_1 distance between two mixtures of distributions on the constructed manifolds as an expression with two functions $\bar{f}_0$ and $\bar{f}_1$ which take nonzero values only on the part of manifolds with bumps.

Lemma 7. *Let* $\bar{f}_0(\underline{x}) := \frac{1}{N_0} \sum_{j=1}^{N_0} \prod_{i=1}^n f_{0j}(x_i) := \frac{1}{N_0} \sum_{j=1}^{N_0} \prod_{i=1}^n (\mathbb{1}_{\{x_i \in \mathfrak{m}_{0j}\}} / \mu(\mathfrak{m}_{0j}))$ *and* $\bar{f}_1(\underline{x}) := \frac{1}{N_1} \sum_{j'=1}^{N_1} \prod_{i=1}^n f_{1j'}(x_i) := \frac{1}{N_1} \sum_{j'=1}^{N_1} \prod_{i=1}^n (\mathbb{1}_{\{x_i \in \mathfrak{m}_{1j'}\}} /$

$\mu(\mathfrak{m}_{1j'})$). *Then we have*

$$L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} Q_{0j}^n, \frac{1}{N_1}\sum_{j'=1}^{N_1} Q_{1j'}^n\right) \leq \int \left|\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})\right| d\mu^n. \tag{5}$$

3.3. From L_1 to combinatorics

Before starting the detailed calculations of (5), let us emphasize that the density values $\prod_{i=1}^n f_{0j}(x_i)$ are determined (as a nonzero fixed value $1/(C-C_0)^n$) only by the specified perturbations in $\mathfrak{m}_{0j}$. In order to use some combinatorial ideas, we divide the whole integral region $\cup_{i=1}^n \left\{x_i \in (\cup_{l=1}^{2m}\mathfrak{m}_l)\right\}$ into $2m$ disjoint regions $S_1, \dots, S_{2m}$ for which

$$\cup_{i=1}^n \left\{x_i \in (\cup_{l=1}^{2m}\mathfrak{m}_l)\right\} = \cup_{u=1}^{2m}\{(x_1,\dots,x_n) \in S_u\},$$

and each S_u, $1 \leq u \leq 2m$, is composed of disjoint union of u unique $\mathfrak{m}_l$'s among $2m$ number of possible $\mathfrak{m}_l$'s. In other words, each disjoint region in S_u has the shape $\mathfrak{m}_{l_1} \times \mathfrak{m}_{l_2} \times \ldots \times \mathfrak{m}_{l_n}$ with u unique $\mathfrak{m}_l$'s so that $|\cup_k \{l_k\}| = u$. Accordingly, we let $S_1 = \mathfrak{m}_1^n \cup \mathfrak{m}_2^n \cup \ldots \cup \mathfrak{m}_{2m}^n$ and $S_2 = (\mathfrak{m}_1 \times \mathfrak{m}_2^{n-1}) \cup (\mathfrak{m}_1 \times \mathfrak{m}_3^{n-1}) \cup \ldots \cup (\mathfrak{m}_{2m-1}^{n-1} \times \mathfrak{m}_{2m})$,...,$S_{2m} = (\mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \mathfrak{m}_{2m}^{n-2m+1}) \cup \ldots \cup (\mathfrak{m}_1^{n-2m+1} \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_{2m})$.

Now, we evaluate the integral (5). First note that each S_u are disjoint, which gives

$$\int |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| = \sum_{u=1}^{2m}\int_{S_u} |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})|.$$

For notational convenience, we define a representative disjoint region in S_u as $\mathfrak{s}_u := \mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_u \times \ldots \times \mathfrak{m}_u = \mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_{u-1} \times \mathfrak{m}_u^{n-u+1}$. More precisely, let $\mathfrak{s}_1 = \mathfrak{m}_1^n$, $\mathfrak{s}_2 = \mathfrak{m}_1 \times \mathfrak{m}_2^{n-1}$, $\mathfrak{s}_3 = \mathfrak{m}_1 \times \mathfrak{m}_2 \times \mathfrak{m}_3^{n-2}$, ..., and $\mathfrak{s}_{2m} = \mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_{2m}^{n-2m+1}$.

By Lemma 8, a result for the traditional occupancy problem, we have the total number Υ_u of disjoint regions in each S_u satisfies,

$$\Upsilon_u = \binom{2m}{u}\left[\sum_{l=0}^{u}\binom{u}{l}(-1)^l(u-l)^n\right], \quad u = 1,\dots,2m. \tag{6}$$

Lemma 8. *Consider the distribution of n balls in $2m$ bins, assuming that each ball has the equal probability n^{-2m} of being placed in each bin. Suppose Υ_u be the total number of cases with u unique bins (i.e. $2m-u$ empty bins). Then (6) holds.*

Proof. For the proof, see page 2 by Kolchin et al. (1978). □

Using the Equation (6), we simplify the right side of (5) as follows in Lemma 9.

Lemma 9. *Let* $I_1 = \{l : \max(u-m, 0) \le l \le \min(m, u)\}$, $I_2 = \{l : \max(u-m+1, 0) \le l \le \min(m+1, u)\}$, *and* $I_3 = \{l : \max(u-m-1, 0) \le l \le \min(m-1, u)\}$. *Define*

$$\varrho_l := \binom{u}{l} \left| \frac{\binom{2m-u}{m-l}\mathbb{1}_{I_1}}{\binom{2m}{m}} - \frac{\binom{2m-u}{m+1-l}\mathbb{1}_{I_2} + \binom{2m-u}{m-1-l}\mathbb{1}_{I_3}}{2\binom{2m}{m-1}} \right|.$$

Then we have

$$\begin{aligned} \int |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| &= \sum_{u=1}^{2m} \Upsilon_u \int_{\mathbf{s}_u} |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| \\ &= \sum_{u=1}^{2m} \frac{\Upsilon_u}{(2m)^n} \sum_{l=0}^{u} \varrho_l. \end{aligned} \tag{7}$$

3.4. Bounding the L_1 distance

In this final step, we shall prove that

$$(7) := \sum_{u=1}^{2m} \frac{\Upsilon_u}{(2m)^n} \sum_{l=0}^{u} \varrho_l = O(1/\log n).$$

We start to consider bounds for the quantity $\Upsilon_u/(2m)^n$, whose limiting distribution is found in the traditional occupancy problems. Let Ψ_n be the random variable corresponding to the number of nonempty bins when we throw n balls into $2m$ bins. Then $\Upsilon_u/(2m)^n$ is the probability of having $\Psi_n = u$ in this regime. Define $\alpha_n := (2m)\exp(-n/(2m)) = n^{1-t}/(t \log n) \to \infty$ (by recalling $2m = n/(t \log n)$ with $0 < t < 1/2$). By applying Theorem 2 of Kamath et al. (1994), for a large n satisfying $n^{(1-2t)}/\log n \ge (t/\theta^2)\log(2n^2)$, for any $\theta > 0$,

$$\begin{aligned} \mathbb{P}(|\Psi_n - (2m - \alpha_n)| \ge \theta\alpha_n) &\le 2\exp\left(-\theta^2 \frac{n^{(1-2t)}}{t \log n}\right) \\ &\le \frac{1}{n^2} = o\left(\frac{1}{n}\right), \end{aligned}$$

which implies that we only need to calculate $\sum_{l=0}^{u} \varrho_l$ for the range of $\Psi_n \in [2m - \alpha_n - \theta\alpha_n, 2m - \alpha_n + \theta\alpha_n] =: [u.l, u.u]$ since we know that $\sum_l \varrho_l \le 2$. Observing this with (7),

$$\int |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| = \sum_{u=u.l}^{u.u} \mathbb{P}(\Psi_n = u) \sum_{l=0}^{u} \varrho_l + o\left(\frac{1}{n}\right). \tag{8}$$

Now we focus on the first term in (8). Since $\alpha_n = o(m)$, for the range of $u \in [u.l, u.u]$, we have simpler regions for the indices, e.g. $I_1 = \{l : u - m \le l \le m\}$, $I_2 = \{l : u - m + 1 \le l \le m + 1\}$, and $I_3 = \{l : u - m - 1 \le l \le m - 1\}$. Hence we need to treat 4 cases outside of the common region $I_1 \cap I_2 \cap I_3 = \{l :$

$u-m+1 \leq l \leq m-1\}$ separately. When $l = m$ or $l = u-m$, for a sufficiently large n,

$$\varrho_l = \binom{u}{u-m} \left| \frac{1}{\binom{2m}{m}} \left(1 - \frac{1}{2}\frac{m+1}{m}(2m-u)\right) \right| = O\left(\frac{\alpha_n}{2^{\alpha_n}}\right) = o\left(\frac{1}{n}\right).$$

When $l = u-m-1$ or $l = m+1$,

$$\varrho_l = \frac{\binom{u}{m+1}}{2\binom{2m}{m+1}} = O\left(\frac{1}{2^{\alpha_n}}\right) = o\left(\frac{1}{n}\right).$$

Substituting these into the calculations, we have

$$\begin{aligned}\sum_{u=u.l}^{u.u} \mathbb{P}(\Psi_n = u) \sum_{l=0}^{u} \varrho_l &= \sum_{u=u.l}^{u.u} \mathbb{P}(\Psi_n = u) \left(2\varrho_m + 2\varrho_{m+1} + \sum_{l=u-m+1}^{m-1} \varrho_l\right) \\ &= \sum_{u=u.l}^{u.u} \mathbb{P}(\Psi_n = u) \sum_{l=u-m+1}^{m-1} \varrho_l + o\left(\frac{1}{n}\right). \qquad (9)\end{aligned}$$

Now, for the range of $u-m+1 \leq l \leq m-1$ (where $u.l \leq u \leq u.u$), we can further simplify ϱ_l as follows,

$$\begin{aligned}\sum_{l=u-m+1}^{m-1} \varrho_l &= \sum_{l=u-m+1}^{m-1} \frac{\binom{u}{l}\binom{2m-u}{m-l}}{\binom{2m}{m}} \left|1 - \frac{1}{2}\frac{m+1}{m}\left(\frac{m-u+l}{m+1-l} + \frac{m-l}{m-u+1+l}\right)\right| \\ &=: \sum_{l=u-m+1}^{m-1} p_{2m,u,m}(l) \left|1 - \frac{1}{2}\frac{m+1}{m}\left(\frac{m-u+l}{m+1-l} + \frac{m-l}{m-u+1+l}\right)\right| \end{aligned} \qquad (10)$$

where $p_{2m,u,k}(x) = \frac{\binom{u}{x}\binom{2m-u}{k-x}}{\binom{2m}{k}}$ is the hypergeometric probability with parameters $(2m, u, k)$.

Then let us consider the random variable Z with $\mathbb{P}(Z = l) = p_{2m,u,m}(l)$. From the property of the hypergeometric distribution, we know $\mathbb{E}Z = \frac{u}{2}$.

Lemma 10. *Let Z be the Hypergeometric random variable with parameters (m, u, k). That is,*

$$\mathbb{P}(Z = z) = \frac{\binom{u}{z}\binom{m-u}{k-z}}{\binom{m}{u}}.$$

Denote $p = u/m$. Then, for some $\eta \geq 0$,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq \eta k) \leq 2\left(\left(\frac{p}{p+\eta}\right)^{p+\eta}\left(\frac{1-p}{1-p-\eta}\right)^{1-p-\eta}\right)^k.$$

In addition, if $0 \leq \eta \leq 1-p$, then

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq \eta k) \leq 2\exp(-2\eta^2 k).$$

Proof. For the one side inequality proof, see Hoeffding (1963) or Chvátal (1979). Then we obtain the other side of the inequality using symmetry. □

By the tail probability provided in Lemma 10, with replacing $\mathbb{E}Z$ with its expectation $u/2$ and k with m, we note that for $0 \le \eta \le u/(2m)$

$$\mathbb{P}\left(|Z - \frac{u}{2}| \ge \eta m\right) \le 2\exp(-2\eta^2 m). \tag{11}$$

Here we take $\eta := \alpha_n/(m \log n) \to 0$ so that for $0 < t < 1/2$

$$\eta^2 m = \alpha_n^2/(m(\log n)^2) = 4n^{1-2t}/(\log n)^2 \to \infty.$$

Based on the small tail probabilities, we divide the summation in (10) into two regions, $\{l : |l - u/2| \le \alpha_n/\log n\}$ and $\{l : |l - u/2| > \alpha_n/\log n\}$. For the second region, by bounding the absolute term in (10) by $O(1/m)$ (since the absolute term has the largest value for the smallest $l = u-m+1$ or the largest $l = m-1$),

$$\sum_{|l-u/2|>\alpha_n/\log n} \varrho_l \le O(1/m) \sum_{|l-u/2|>\alpha_n/\log n} p_{2m,u,m}(l) \le o(1/n) \tag{12}$$

where the last inequality is followed by the tail probability (11). For the first region, we bound the absolute term with the largest index $l = u/2 + \alpha_n/\log n$ (this absolute term has the smallest value at $u/2$) where $u \in [2m-\alpha_n-\theta\alpha_n, 2m-\alpha_n+\theta\alpha_n]$:

$$\begin{aligned}&\left|1 - \frac{1}{2}\frac{m+1}{m}\left(\frac{m-u+l}{m+1-l} + \frac{m-l}{m-u+1+l}\right)\right| \\ &= \left|1 - \frac{1}{2}\frac{m+1}{m}\left(\frac{m - \frac{u}{2} - \frac{\alpha_n}{\log n}}{m - \frac{u}{2} + \frac{\alpha_n}{\log n} + 1} + \frac{m - \frac{u}{2} + \frac{\alpha_n}{\log n}}{m - \frac{u}{2} - \frac{\alpha_n}{\log n} + 1}\right)\right| = O\left(\frac{1}{\log n}\right).\end{aligned}$$

Hence,

$$\sum_{|l-u/2|\le\alpha_n/\log n} \varrho_l \le O\left(\frac{1}{\log n}\right) \sum_{|l-u/2|\le\alpha_n/\log n} p_{2m,u,m}(l) \le O\left(\frac{1}{\log n}\right). \tag{13}$$

Considering both regions, $\sum_{l=u-m+1}^{m-1} \varrho_l = O(1/\log n) \to 0$, which implies $\int |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| = O(1/\log n)$ via (8) and (9). Therefore, the claim is proved. Now we combine all of the above ideas into the proof.

3.5. Proof of main results

Proof of Theorem 1. From the first step, we construct $\mathcal{M}_0 := \{\mathsf{M}_{0j},\ j = 1, \dots, N_0\}$ and $\mathcal{M}_1 = \{\mathsf{M}_{1j'},\ j' = 1, \dots, N_1\}$ where $H(\mathsf{M}_{0j}, \mathsf{M}_{1j'}) \ge 2\gamma$ for all $j = 1, \dots, N_0$ and $j' = 1, \dots, N_1$. Then we simplified the L_1 distance between mixture densities from these two groups via equations (5) and (7). Finally in the last step, via equations (8), (9), (12), and (13), we have shown that

$L_1(\frac{1}{N_1}\sum_j Q_{0j}^n, \frac{1}{N_1}\sum_{j'} Q_{1j'}^n) \to 0$ where $Q_{0j} = U(\mathsf{M}_{0j})$ and $Q_{1j'} = U(\mathsf{M}_{1j'})$ with the choice of $\gamma^{d/2} \asymp m \asymp n/\log n$. By Lemma 5,

$$\begin{aligned}\sup_{Q\in\mathcal{Q}} \mathbb{E}_{Q^n} H(\hat{\mathsf{M}}, \mathsf{M}(Q)) &\geq \gamma\left(1 - \frac{1}{2}L_1\left(\frac{1}{N_1}\sum_j Q_{0j}^n, \frac{1}{N_1}\sum_{j'} Q_{1j'}^n\right)\right)\\ &\to \gamma \asymp \left(\frac{\log n}{n}\right)^{d/2},\end{aligned}$$

which proves the theorem. □

Proof of Theorem 3. We construct the same set of manifolds $\mathcal{M}_0$ and $\mathcal{M}_1$ as in the noiseless model. By construction, $H(\mathsf{M}_{0j}, \mathsf{M}_{1j'}) \geq 2\gamma$. Again we claim that $L_1\left(\frac{1}{N_0}\sum_j Q_{0j}^n, \frac{1}{N_1}\sum_{j'=1}^{N_1} Q_{1j'}^n\right) \to 0$ where $Q_{0j} = (1-\pi)U(\mathcal{K}) + \pi U(\mathsf{M}_{0j})$ and $Q_{1j'} = (1-\pi)U(\mathcal{K}) + \pi U(\mathsf{M}_{1j'})$ with the choice $\gamma^{d/2} \asymp m \asymp n/\log n$, which gives the lower bound $(\log n/n)^{2/d}$ by Lemma 5.

Here we let the dominating measure $\mu := U(\mathcal{K}) + U(\mathfrak{m} \cup (\cup_{l=1}^{2m}\mathfrak{m}_l))$. By symmetry and singular property of $U(\mathcal{K})$ and $U(\mathsf{M}_{0j})$ or $U(\mathsf{M}_{1j'})$,

$$\begin{aligned}&L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} Q_{0j}^n, \frac{1}{N_1}\sum_{j'=1}^{N_1} Q_{1j'}^n\right)\\ &\quad= \sum_{k=1}^{n} \beta_{n,k,\pi} L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^k(\mathsf{M}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^k(\mathsf{M}_{1j'})\right).\end{aligned}$$

Now, using the exact same idea in Lemma 11, we have

$$\begin{aligned}&L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^k(\mathsf{M}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^k(\mathsf{M}_{1j'})\right)\\ &\quad\leq L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^n(\mathsf{M}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^n(\mathsf{M}_{1j'})\right),\end{aligned}$$

which implies by Theorem 1 and $\sum_{k=1}^{n} \beta_{n,k,\pi} = 1 - (1-\pi)^n \leq 1$,

$$\begin{aligned}L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} Q_{0j}^n, \frac{1}{N_1}\sum_{j'=1}^{N_1} Q_{1j'}^n\right) &\leq L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^n(\mathsf{M}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^n(\mathsf{M}_{1j'})\right)\\ &\to 0.\end{aligned}$$

□

Remark 3 (Continuation of Remark 1)**.** As explained in Remark 1, we use a base manifold M_0 defined in (14) as the null. Then, we construct the alternatives having one inward bump:

$$\mathsf{M}_j = \{(u, b_j(u), 0_{D-d-1}) : u \in \mathcal{B}_d\} \cup \{(u, -a(u), 0_{D-d-1}) : u \in \mathcal{B}_d\}$$

for $j = 1, \ldots, m$ where $m \asymp n/\log n$ such that

$$b_j(u) = \begin{cases} a(u) & \text{if} \quad u \in \mathcal{B}_d \setminus \Big(\{u : ||u_j|| \leq \sqrt{4\gamma\kappa + 3\gamma^2}\} \Big) \\ 2(\kappa+\gamma) - \tilde{b}_j(u) & \text{if} \quad u \in \{u : ||u_j|| \leq \sqrt{4\gamma\kappa + 3\gamma^2}\}, \end{cases}$$

where $\tilde{b}_j(u)$ is defined exactly the same as before in Step 1. By construction, $\Delta(\mathsf{M}_0) \geq \kappa$, $\Delta(\mathsf{M}_j) \geq \kappa$, and $H(\mathsf{M}_0, \mathsf{M}_j) \geq 2\gamma$ for all $j = 1, \ldots, m$. Then using the same counting method, we can evaluate the L_1 distance between $Q_0 := U(\mathsf{M}_0)$ and mixtures of $Q_j := U(\mathsf{M}_j)$. Denote $\tilde{\Psi}_n$ as the random variable of the number of nonempty bins when we throw n balls into m bins, then $\mathbb{E}[\tilde{\Psi}_n] = m - \tilde{\alpha}_n$ (where $\tilde{\alpha}_n = m\exp(-n/m) = o(m)$). Then,

$$L_1\big(Q_0, \frac{1}{m}\sum_{j=1}^{m} Q_j\big) = \frac{2}{m}\sum_{j=1}^{m} j\frac{\Upsilon_j}{m^n} = \frac{2}{m}\mathbb{E}[\tilde{\Psi}_n] \to 2,$$

which proves that Q_0 and mixtures of Q_j are distinguishable.

4. Proofs of auxiliary lemmas

4.1. Proof of Lemma 6

We define a base manifold for $u \in \mathbb{R}^d$ with the notation $\mathcal{B}_d := B_d(0, 1+\kappa+\gamma)$ where $\gamma \leq \kappa/3$,

$$\mathsf{M}_0 = \{(u, a(u), 0_{D-d-1}) : u \in \mathcal{B}_d\} \cup \{(u, -a(u), 0_{D-d-1}) : u \in \mathcal{B}_d\} \tag{14}$$

where

$$a(u) = \begin{cases} \kappa + \gamma & \text{for} \quad ||u|| \leq 1 \\ \sqrt{(\kappa+\gamma)^2 - (||u|| - 1)^2} & \text{for} \quad 1 < ||u|| \leq 1 + \kappa + \gamma. \end{cases}$$

The radius $1 + \kappa + \gamma$ of this base manifold is larger than the radius $1 + \kappa$ appeared in Genovese et al. (2012a). Larger radius is chosen such that manifold M with bumps (which will be constructed on M_0) satisfies the curvature condition $\Delta(\mathsf{M}) \geq \kappa$.

We consider $2m := n/(t\log n) \asymp \gamma^{-d/2}$ (for some $t \in (0, 1/2)$) number of bumps on M_0. Let $\mathcal{R}_{0j}(\mathcal{R}_{1j'})$ be a set of m ($m+1$ or $m-1$) integers out of $1, \ldots, 2m$ for $j = 1, \ldots, N_0$ ($j' = 1, \ldots, N_1$). For instance, we let $\mathcal{R}_{01} = \{1, 2, \ldots, m\}$, $\mathcal{R}_{02} = \{1, 3, \ldots, m+1\}$,.., $\mathcal{R}_{0N_0} = \{m+1, \ldots, 2m\}$. Similarly, $\mathcal{R}_{11} = \{1, 2, \ldots, m-1\}, \mathcal{R}_{12} = \{1, 3, \ldots, m\}, \ldots, \mathcal{R}_{1N_1} = \{m, \ldots, 2m\}$. By construction, $N_0 = \binom{2m}{m}$ and $N_1 = \binom{2m}{m-1} + \binom{2m}{m+1}$.

For the first group $\mathcal{M}_0$, we consider m outward bumps and m inward bumps. For $j = 1, \ldots, N_0$,

$$\mathsf{M}_{0j} = \{(u, b_{0j}(u), 0_{D-d-1}) : u \in \mathcal{B}_d\} \cup \{(u, -a(u), 0_{D-d-1}) : u \in \mathcal{B}_d\}$$

where $b_{0j}(u)$ is equal to $a(u)$ except $2m$ regions such that

$$b_{0j}(u)=\begin{cases} a(u) & \text{if} \quad u\in\mathcal{B}_d\backslash\left(\cup_{l=1}^{2m}\{u:||u_l||\le\sqrt{4\gamma\kappa+3\gamma^2}\}\right) \\ \tilde{b}_l(u) & \text{if} \quad u\in\cup_{l\in\mathcal{R}_{0j}}\{u:||u_l||\le\sqrt{4\gamma\kappa+3\gamma^2}\} \\ 2(\kappa+\gamma)-\tilde{b}_l(u) & \text{if} \quad u\in\cup_{l\in\mathcal{R}_{0j}^c}\{u:||u_l||\le\sqrt{4\gamma\kappa+3\gamma^2}\}, \end{cases}$$

and

$$\tilde{b}_l(u)=\begin{cases} \gamma+\sqrt{(\kappa+\gamma)^2-||u_l||^2} \\ \qquad\qquad \text{if} \quad ||u_l||\le\frac{1}{2}\sqrt{4\gamma\kappa+3\gamma^2} \\ 2(\kappa+\gamma)-\sqrt{(\kappa+\gamma)^2-\left(||u_l||-\sqrt{4\gamma\kappa+3\gamma^2}\right)^2} \\ \qquad\qquad \text{if} \quad \frac{1}{2}\sqrt{4\gamma\kappa+3\gamma^2}<||u_l||\le\sqrt{4\gamma\kappa+3\gamma^2}. \end{cases}$$

where $\mathcal{R}_{0j}^c:=\{1,\dots,2m\}\backslash\mathcal{R}_{0j}$ and $u_l=u-\iota_l=(u_1,\dots,u_d)-(\iota_{l1},\dots,\iota_{ld})$ denoting $(\iota_{l1},\dots,\iota_{ld})$ as the center point of the bumps. By construction, there exist $2m$ number of centers, namely $\iota_1,\dots,\iota_{2m}$.

For the second group, $\mathcal{M}_1$ consists of manifolds $\mathsf{M}_{1j'}$, $j'=1,\dots,N_1$ similar to M_{0j} but with $m+1$ or $m-1$ outward bumps which in turn means $m-1$ or $m+1$ inward bumps. That is, for $j'=1,\dots,N_1$,

$$\mathsf{M}_{1j'}=\{(u,b_{1j'}(u),0_{D-d-1}):u\in\mathcal{B}_d\}\cup\{(u,-a(u),0_{D-d-1}):u\in\mathcal{B}_d\}$$

where $b_{1j'}(u)$ is equal to $a(u)$ except $2m$ regions

$$b_{1j'}(u)=\begin{cases} a(u) & \text{if} \quad u\in\mathcal{B}_d\backslash\left(\cup_{l=1}^{2m}\{u:||u_l||\le\sqrt{4\gamma\kappa+3\gamma^2}\}\right) \\ \tilde{b}_l(u) & \text{if} \quad u\in\cup_{l\in\mathcal{R}_{1j'}}\{u:||u_l||\le\sqrt{4\gamma\kappa+3\gamma^2}\} \\ 2(\kappa+\gamma)-\tilde{b}_l(u) & \text{if} \quad u\in\cup_{l\in\mathcal{R}_{1j'}^c}\{u:||u_l||\le\sqrt{4\gamma\kappa+3\gamma^2}\}, \end{cases}$$

where $\tilde{b}_l(u)$ is defined exactly the same as before.

Genovese et al. (2012b) constructed one bump on the similar kind of base manifold $\tilde{\mathsf{M}}_0$ whose uniform measure on the manifold with that bump is about $\gamma^{d/2}$. The shape of the bump is a union of two portions of spheres, and the bump is centered at $(0,\dots,0)\in\mathbb{R}^d$ and defined on $||u||\le\sqrt{4\gamma\kappa-\gamma^2}$. And also the Hausdorff distance from $\tilde{\mathsf{M}}_0$ is γ. Here we consider slightly modified bumps, located not only on the one region but located as many disjoint regions as possible on M_0. In other words, we seek the maximal number of disjoint bumps which also guarantees the Hausdorff distance from M_0 being γ. For $d=1$ case, we can construct those bumps on each disjoint interval with length $\sqrt{4\gamma k+3\gamma^2}$ (which is upper bounded by $\sqrt{5\gamma\kappa}$ since $\gamma\le\kappa/3$). For general d dimensional manifolds, by using grid points separated by $\sqrt{\gamma}$ in each dimensions, there exist at least $\gamma^{-d/2}$ (of order) number of disjoint bumps on the region $\{||u||\le 1\}$. Thus, we let $2m\asymp\gamma^{-d/2}$.

Then we need to check if these satisfy the condition for the model. Note that each outward bump is just magnified version of the bump used by Genovese et al. (2012b). Indeed, these bumps are constructed with parts of sphere with radius $\kappa+\gamma$ located on different regions of M_0. Each inward bump is just the reflected version of the outward bump. Thus, constructed manifolds have no manifold boundary and $\Delta(\mathsf{M}_{0j}) \geq \kappa$ and $\Delta(\mathsf{M}_{1j'}) \geq \kappa$ for all $j, j' = 1, \ldots, 2m$. Also we check $H(\mathsf{M}_{0j}, \mathsf{M}_{1j'}) \geq 2\gamma$ for all $j = 1, \ldots, N_0$, and $j' = 1, \ldots, N_1$ because there is always at least one different spot in M_{0j} and $\mathsf{M}_{1j'}$.

4.2. Proof of Lemma 7

To conveniently express the L_1 distance between mixtures, we use the notation $\underline{x} := (x_1, \ldots x_n)$ and $\underline{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Then we expand the product term as follows,

$$\begin{aligned}\prod_{i=1}^{n} \left(\mathbb{1}_{\{x_i \in \mathsf{M}_{0j}\}}\right) &= \prod_{i=1}^{n} \left(\mathbb{1}_{\{x_i \in \mathfrak{m}\}} + \mathbb{1}_{\{x_i \in \mathfrak{m}_{0j}\}}\right) \\ &= \mathbb{1}_{\{\underline{x} \in \mathfrak{m}\}} + \sum_{i=1}^{n} \mathbb{1}_{\{\underline{x}_{-i} \in \mathfrak{m}, x_i \in \mathfrak{m}_{0j}\}} + \sum_{i \neq j} \mathbb{1}_{\{\underline{x}_{-(i,j)} \in \mathfrak{m}, x_i, x_j \in \mathfrak{m}_{0j}\}} \\ &\qquad \ldots + \sum_{i=1}^{n} \mathbb{1}_{\{x_i \in \mathfrak{m}, \underline{x}_{-i} \in \mathfrak{m}_{0j}\}} + \mathbb{1}_{\{\underline{x} \in \mathfrak{m}_{0j}\}}.\end{aligned}$$

Using the above expression,

$$\begin{aligned}C^n\Big(\frac{1}{N_0}\sum_{j=1}^{N_0} q_{0j}^n - \frac{1}{N_1}\sum_{j'=1}^{N_1} q_{1j'}^n\Big) &= \frac{1}{N_0}\sum_{j=1}^{N_0}\prod_{i=1}^{n} \mathbb{1}_{\{x_i \in \mathsf{M}_{0j}\}} - \frac{1}{N_1}\sum_{j'=1}^{N_1}\prod_{i=1}^{n} \mathbb{1}_{\{x_i \in \mathsf{M}_{1j'}\}} \\ &= \sum_{i=1}^{n} \mathbb{1}_{\{\underline{x}_{-i} \in \mathfrak{m}\}} \left(\frac{1}{N_0}\sum_{j=1}^{N_0} \mathbb{1}_{\{x_i \in \mathfrak{m}_{0j}\}} - \frac{1}{N_1}\sum_{j'=1}^{N_1} \mathbb{1}_{\{x_i \in \mathfrak{m}_{1j'}\}}\right) \\ &+ \sum_{i \neq j} \mathbb{1}_{\{\underline{x}_{-(i,j)} \in \mathfrak{m}\}} \left(\frac{1}{N_0}\sum_{j=1}^{N_0} \mathbb{1}_{\{x_i, x_j \in \mathfrak{m}_{0j}\}} - \frac{1}{N_1}\sum_{j'=1}^{N_1} \mathbb{1}_{\{x_i, x_j \in \mathfrak{m}_{1j'}\}}\right) \\ &+ \ldots + \sum_{i=1}^{n} \mathbb{1}_{\{x_i \in \mathfrak{m}\}} \left(\frac{1}{N_0}\sum_{j=1}^{N_0} \mathbb{1}_{\{\underline{x}_{-i} \in \mathfrak{m}_{0j}\}} - \frac{1}{N_1}\sum_{j'=1}^{N_1} \mathbb{1}_{\{\underline{x}_{-i} \in \mathfrak{m}_{1j'}\}}\right) \\ &+ \left(\frac{1}{N_0}\sum_{j=1}^{N_0} \mathbb{1}_{\{\underline{x} \in \mathfrak{m}_{0j}\}} - \frac{1}{N_1}\sum_{j'=1}^{N_1} \mathbb{1}_{\{\underline{x} \in \mathfrak{m}_{1j'}\}}\right).\end{aligned}$$

By symmetry, and also using the disjoint property between $\mathfrak{m}$ and $\mathfrak{m}_{0j}$, and $\mathfrak{m}$ and $\mathfrak{m}_{1j'}$, L_1 distance is actually equal to the following,

$$
\begin{aligned}
&L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} Q_{0j}^n, \frac{1}{N_1}\sum_{j'=1}^{N_1} Q_{1j'}^n\right)\\
&= \int \left| \frac{1}{N_0}\sum_{j=1}^{N_0}\prod_{i=1}^{n}\left(\frac{1}{C}\mathbb{1}_{\{x_i\in \mathsf{M}_{0j}\}}\right) - \frac{1}{N_1}\sum_{j'=1}^{N_1}\prod_{i=1}^{n}\left(\frac{1}{C}\mathbb{1}_{\{x_i\in \mathsf{M}_{1j'}\}}\right)\right| d\mu^n\\
&= \sum_{k=1}^{n}\beta_{n,k,\frac{C_0}{C}}\int \left| \frac{1}{N_0}\sum_{j=1}^{N_0}\prod_{i=1}^{k}\left(\frac{\mathbb{1}_{\{x_i\in \mathfrak{m}_{0j}\}}}{C-C_0}\right) - \frac{1}{N_1}\sum_{j'=1}^{N_1}\left(\prod_{i=1}^{k}\frac{\mathbb{1}_{\{x_i\in \mathfrak{m}_{1j'}\}}}{C-C_0}\right)\right| d\mu^k\\
&= \sum_{k=1}^{n}\beta_{n,k,\frac{C_0}{C}} L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^k(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^k(\mathfrak{m}_{1j'})\right), \qquad (15)
\end{aligned}
$$

where the second equality is followed by $\mu(\mathfrak{m}) = C_0$, $\mu(\mathfrak{m}_{0j}) = \mu(\mathfrak{m}_{1j'}) = C - C_0$ with the binomial coefficient notation $\beta_{n,k,p} := \binom{n}{k} p^{n-k}(1-p)^k$, and the last equality is obtained by definition.

By Lemma 11, the L_1 expression in (15) can be further upper bounded,

$$
\begin{aligned}
&L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^k(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^k(\mathfrak{m}_{1j'})\right)\\
&\qquad \le L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^n(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^n(\mathfrak{m}_{1j'})\right),
\end{aligned}
$$

which gives the desired upper bound

$$
\begin{aligned}
&L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} Q_{0j}^n, \frac{1}{N_1}\sum_{j'=1}^{N_1} Q_{1j'}^n\right)\\
&\qquad \le L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^n(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^n(\mathfrak{m}_{1j'})\right).
\end{aligned}
$$

Lemma 11. *Let $U(\mathfrak{m}_{0j})$ be the uniform measure on $\mathfrak{m}_{0j}$ defined in (4). Define $U(\mathfrak{m}_{1j'})$ similarly. Then for $k < n$,*

$$
\left|\frac{\sum_{j=1}^{N_0} U^k(\mathfrak{m}_{0j})}{N_0} \wedge \frac{\sum_{j'=1}^{N_1} U^k(\mathfrak{m}_{1j'})}{N_1}\right| \ge \left|\frac{\sum_{j=1}^{N_0} U^n(\mathfrak{m}_{0j})}{N_0} \wedge \frac{\sum_{j'=1}^{N_1} U^n(\mathfrak{m}_{1j'})}{N_1}\right|,
$$

and

$$\begin{aligned} &L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^k(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^k(\mathfrak{m}_{1j'})\right) \\ &\quad \le L_1\left(\frac{1}{N_0}\sum_{j=1}^{N_0} U^n(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} U^n(\mathfrak{m}_{1j'})\right). \end{aligned}$$

Proof. Note that $L_1(P,Q) = 2(1-|P\wedge Q|)$. Thus it is enough to prove the first claim. Let $\mathfrak{m}_{0j^*}$ and $\mathfrak{m}_{1j'^*}$ be the pair satisfying $\mu(\mathfrak{m}_{0j^*}\cap\mathfrak{m}_{1j'^*}) = (2m-1)c\gamma^{d/2}$ which share all the bumps except one location. By construction $\mu(\mathfrak{m}_{0j}\cap\mathfrak{m}_{1j'}) \le (2m-1)c\gamma^{d/2}$ for all $j = 1,\dots,N_0$ and $j' = 1,\dots,N_1$. Then,

$$\begin{aligned} &\int \min\left(\frac{1}{N_0}\sum_{j=1}^{N_0} dU^n(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} dU^n(\mathfrak{m}_{1j'})\right) \\ &\quad = \int \min\left(\frac{1}{N_0}\sum_{j=1}^{N_0} dU^{n-1}(\mathfrak{m}_{0j})dU(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} dU^{n-1}(\mathfrak{m}_{1j'})dU(\mathfrak{m}_{1j'})\right) \\ &\quad \le \int \min\left(\frac{1}{N_0}\sum_{j=1}^{N_0} dU^{n-1}(\mathfrak{m}_{0j})dU(\mathfrak{m}_{0j^*}), \frac{1}{N_1}\sum_{j'=1}^{N_1} dU^{n-1}(\mathfrak{m}_{1j'})dU(\mathfrak{m}_{1j'^*})\right) \\ &\quad = \frac{2m-1}{2m}\int \min\left(\frac{1}{N_0}\sum_{j=1}^{N_0} dU^{n-1}(\mathfrak{m}_{0j}), \frac{1}{N_1}\sum_{j'=1}^{N_1} dU^{n-1}(\mathfrak{m}_{1j'})\right). \end{aligned}$$

We can continue the same calculation for a smaller value for k, which proves the claim. □

4.3. Proof of Lemma 9

First, we consider the simplest case $u = 1$ with the integral region $\mathfrak{m}_1^n$. Then the only one perturbation region $\mathfrak{m}_1$ in the constructed manifold will affect the density. Suppose $\mathfrak{m}_1$ takes the outward perturbation $\mathfrak{m}_1^+$. Then,

$$\int_{(\mathfrak{m}_1^+)^n}\left|\frac{1}{N_0}\sum_{j=1}^{N_0}\prod_{i=1}^{n}\mathbb{1}\{x_i\in\mathfrak{m}_{0j}\} - \frac{1}{N_1}\sum_{j'=1}^{N_1}\prod_{i=1}^{n}\mathbb{1}\{x_i\in\mathfrak{m}_{1j'}\}\right|$$

becomes the comparison problem between counting how many $\mathfrak{m}_{0j}$ can take $(\mathfrak{m}_1^+)$ and counting how many $\mathfrak{m}_{1j'}$ can take $(\mathfrak{m}_1^+)$. If $\mathfrak{m}_1$ takes the inward perturbation $\mathfrak{m}_1^-$, then we can ask similar question with replacing $\mathfrak{m}_1^+$ to $\mathfrak{m}_1^-$ from the previous sentence. Again, we can ask the same question for other region $\mathfrak{m}_l$ for $l = 2,\dots,2m$. By symmetry, by defining the representative region in S_1 as $\mathfrak{s}_1 := \mathfrak{m}_1^n$,

$$\int_{S_1} |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| = \Upsilon_1 \int_{\mathbf{s}_1} |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})|,$$

where Υ_1 is the total number of unique region in S_1 as defined in (6).

We extend the same idea to S_2. Then only two perturbation regions on the constructed manifold will affect the joint density. First consider the region $\mathfrak{m}_1$ and $\mathfrak{m}_2$. Irrelevant to whether the integral region as $\mathfrak{m}_1^l \times \mathfrak{m}_2^{n-l}$ or $\mathfrak{m}_1^{n-l} \times \mathfrak{m}_2^l$ where l $(1 \leq l < n)$ is some arbitrary integer, the density $\prod_{i=1}^n (\mathbb{1}\{x_i \in \mathfrak{m}_{0j}\}/\mu(\mathfrak{m}_{0j}))$ becomes nonzero as long as $\mathfrak{m}_{0j}$ contains perturbations defined on $\mathfrak{m}_1$ and $\mathfrak{m}_2$. Suppose $\mathfrak{m}_1$ takes $\mathfrak{m}_1^+$ and $\mathfrak{m}_2$ takes $\mathfrak{m}_2^-$. Then

$$\int_{(\mathfrak{m}_1^+)^l \times (\mathfrak{m}_2^-)^{n-l}} \left| \frac{1}{N_0} \sum_{j=1}^{N_0} \prod_{i=1}^n \mathbb{1}\{x_i \in \mathfrak{m}_{0j}\} - \frac{1}{N_1} \sum_{j'=1}^{N_1} \prod_{i=1}^n \mathbb{1}\{x_i \in \mathfrak{m}_{1j'}\} \right|$$

becomes the comparison problem between counting how many $\mathfrak{m}_{0j}$ can take $\mathfrak{m}_1^+, \mathfrak{m}_2^-$, and counting how many $\mathfrak{m}_{1j'}$ can take $\mathfrak{m}_1^+, \mathfrak{m}_2^-$ for any $l = 1, \ldots, n-1$. Again, it would not make any difference if we change the region of the perturbations as long as the unique number u of regions is 2.

In general, since it is more complicate, we first only consider the region without specifying perturbation shape. With the same intuition as before, we have the same density value on any disjoint regions in S_u, say, $\mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \mathfrak{m}_u \times \mathfrak{m}_u^{n-u}$ or $\mathfrak{m}_1^2 \times \mathfrak{m}_2^2 \times \ldots \times \mathfrak{m}_u^2 \times \mathfrak{m}_u^{n-2u}$. Indeed, we only need to count how many $\mathfrak{m}_{0j}$ and $\mathfrak{m}_{1j'}$ would contain each specific perturbation (combinations of outward and inward on these regions $\mathfrak{m}_1, \ldots, \mathfrak{m}_u$; we will explain how to calculate these in detail later). Similarly as before, by symmetry, considering the region $\mathfrak{m}_1, \ldots, \mathfrak{m}_u$ or $\mathfrak{m}_2, \ldots, \mathfrak{m}_{u+1}$ would not make any difference.

For notational convenience, we define the representative disjoint region in S_u as $\mathbf{s}_u := \mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_u \times \ldots \times \mathfrak{m}_u = \mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_{u-1} \times \mathfrak{m}_u^{n-u+1}$. More precisely, $\mathbf{s}_2 = \mathfrak{m}_1 \times \mathfrak{m}_2^{n-1}$, $\mathbf{s}_3 = \mathfrak{m}_1 \times \mathfrak{m}_2 \times \mathfrak{m}_3^{n-2}$, …, and $\mathbf{s}_{2m} = \mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_{2m}^{n-2m+1}$. Then, by symmetry as explained,

$$\int \left|\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})\right| = \sum_{u=1}^{2m} \int_{S_u} \left|\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})\right| = \sum_{u=1}^{2m} \Upsilon_u \int_{\mathbf{s}_u} \left|\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})\right|. \tag{16}$$

Now, we evaluate the above integrals. First, we consider $\int_{\mathbf{s}_1} |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})|$. As explained before, on $(\mathfrak{m}_1^-)^n$, $\prod_{i=1}^n f_{0j}(x_i)$ are either zero or $1/\mu(\mathfrak{m}_{0j})^n = 1/(2mc\gamma^{d/2})^n$, and $\int_{\mathfrak{m}_1^-} \ldots \int_{\mathfrak{m}_1^-} d\mu^n = (c\gamma^{d/2})^n$. Then

$$\begin{aligned}
&\int_{\mathbf{s}_1} \left|\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})\right| \\
&\quad = \frac{1}{\mu(\mathfrak{m}_{0j})^n} \int_{(\mathfrak{m}_1^-)^n} \left| \frac{1}{N_0} \sum_{j=1}^{N_0} \mathbb{1}_{\{\underline{x} \in (\mathfrak{m}_{0j} \cap \mathfrak{m}_1^-)^n\}} - \frac{1}{N_1} \sum_{j'=1}^{N_1} \mathbb{1}_{\{\underline{x} \in (\mathfrak{m}_{1j'} \cap \mathfrak{m}_1^-)^n\}} \right| \\
&\qquad + \frac{1}{\mu(\mathfrak{m}_{0j})^n} \int_{(\mathfrak{m}_1^+)^n} \left| \frac{1}{N_0} \sum_{j=1}^{N_0} \mathbb{1}_{\{\underline{x} \in (\mathfrak{m}_{0j} \cap \mathfrak{m}_1^+)^n\}} - \frac{1}{N_1} \sum_{j'=1}^{N_1} \mathbb{1}_{\{\underline{x} \in (\mathfrak{m}_{1j'} \cap \mathfrak{m}_1^+\}^n)} \right|
\end{aligned}$$

$$
= \left(\frac{1}{2m}\right)^n \left| \frac{1}{N_0} \sum_{j=1}^{N_0} \mathbb{1}_{\{\mathfrak{m}_1^- \subset \mathfrak{m}_{0j}\}} - \frac{1}{N_1} \sum_{j'=1}^{N_1} \mathbb{1}_{\{\mathfrak{m}_1^- \subset \mathfrak{m}_{1j'}\}} \right|
+ \left(\frac{1}{2m}\right)^n \left| \frac{1}{N_0} \sum_{j=1}^{N_0} \mathbb{1}_{\{\mathfrak{m}_1^+ \subset \mathfrak{m}_{0j}\}} - \frac{1}{N_1} \sum_{j'=1}^{N_1} \mathbb{1}_{\{\mathfrak{m}_1^+ \subset \mathfrak{m}_{1j'}\}} \right|.
$$

Thus we need to count how many of $\mathfrak{m}_{0j}$s contain $\mathfrak{m}_1^-$ (and $\mathfrak{m}_1^+$) for the first group and how many of $\mathfrak{m}_{1j'}$s contain $\mathfrak{m}_1^-$ (and $\mathfrak{m}_1^+$) for the second group. In fact, there exist $\binom{2m-1}{m}$ number of $\mathfrak{m}_{0j}$s contain $\mathfrak{m}_1^-$. This is since after fixing $\mathfrak{m}_1^-$, there are $2m-1$ number of regions $\mathfrak{m}_2, \ldots, \mathfrak{m}_{2m}$ left for choosing m number of $\mathfrak{m}^+$s. Recall that the first group of manifolds has m number of $+$ (outward)s, then automatically the regions not choosing $+$'s will take values on $-$'s. Similarly there exist $\binom{2m-1}{m-1}$ number of $\mathfrak{m}_{0j}$s that contain $\mathfrak{m}_1^+$ (fixing $\mathfrak{m}_1^+$, there are $2m-1$ number of $\mathfrak{m}_2, \ldots, \mathfrak{m}_{2m}$ left for choosing $m-1$ number of $+$'s). For the second group $\mathfrak{m}_{1j'}$, there exist $\binom{2m-1}{m+1} + \binom{2m-1}{m-1}$ number of manifolds with $\mathfrak{m}_1^-$ and $\binom{2m-1}{m} + \binom{2m-1}{m-2}$ number of manifolds with $\mathfrak{m}_1^+$.

We extend the same idea for a general $\mathbf{s}_u = (\mathfrak{m}_1 \times \mathfrak{m}_2 \times \ldots \times \mathfrak{m}_{u-1} \times \mathfrak{m}_u^{n-u+1})$ consisting of first u unique regions. For now, suppose $u \leq m-1$. The counting ideas are still working. First, count separately for the case of l number of $+$'s on $\mathfrak{m}_1 \times \ldots \times \mathfrak{m}_u$. Clearly l can take values from 0 to u. For choosing the location of this l number of $+$'s, there exist $\binom{u}{l}$ number of possible cases. Consider for the case of 0 number of $+$ (that is, $l = 0$ case) on the region $\mathfrak{m}_1 \times \ldots \times \mathfrak{m}_u$. Then we need to pick m number of $+$'s out of $2m-u$ number of regions $\mathfrak{m}_{u+1}, \ldots, \mathfrak{m}_{2m}$. Thus, for $l = 0$, there exist $\binom{u}{0}\binom{2m-u}{m}$ number of $\mathfrak{m}_{0j}$s. For more general l number of $+$'s, there exist $\binom{u}{l}\binom{2m-u}{m-l}$ number of $\mathfrak{m}_{0j}$s, and with the exact same ideas there exist $\binom{u}{l}\left(\binom{2m-u}{m+1-l} + \binom{2m-u}{m-1-l}\right)$ number of $\mathfrak{m}_{1j'}$s. Thus for the case $u \leq m-1$, we have the exact expression

$$
\int_{\mathbf{s}_u} |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| d\mu^n = \left(\frac{1}{2m}\right)^n \sum_{l=0}^{u} \binom{u}{l} \left| \frac{\binom{2m-u}{m-l}}{N_0} - \frac{\binom{2m-u}{m+1-u} + \binom{2m-u}{m-1-l}}{N_1} \right|.
$$

Now, we consider the case where $u \geq m$. Then we need to be more careful in deciding the possible range of l. For the extreme case, if $u = 2m$, then we know that $+$ numbers should be fixed as m among $2m$ regions for the first group, which restricts the range of l as $l = m$. After choosing this location, there is no freedom left, since there should be a unique $2m$ regions, which determines the exact form of the manifolds. Also in this case, there does not exist manifolds in the second group (since those cannot have m number of $+$'s and m number of $-$'s). Similarly for the first group case, either $m+1$ or $m-1$ number of $+$'s should be fixed with no freedom. This yields

$$
\int_{\mathbf{s}_{2m}} |\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})| = \left(\frac{1}{2m}\right)^n \left[\left| \frac{\binom{2m}{m}}{N_0} - 0 \right| + \left| 0 - \frac{\binom{2m}{m+1}}{N_1} \right| + \left| 0 - \frac{\binom{2m}{m-1}}{N_1} \right| \right]
= 2\left(\frac{1}{2m}\right)^n.
$$

The other cases can be considered in the similar way by restricting the range of the indices.

Combining these ideas, the exact evaluation of the L_1 distance between $\frac{1}{N_0}\sum_{j=1}^{N_0} U^n(\mathfrak{m}_{0j})$ and $\frac{1}{N_1}\sum_{j'=1}^{N_1} U^n(\mathfrak{m}_{1j'})$ is obtained as follows,

$$\begin{aligned}
&\int \left|\bar{f}_0(\underline{x}) - \bar{f}_1(\underline{x})\right| \qquad (17)\\
&= \sum_{u=1}^{2m} \frac{\Upsilon_u}{(2m)^n} \sum_{l=0}^{u} \binom{u}{l} \left| \frac{\binom{2m-u}{m-l}\mathbb{1}_{I_1}}{\binom{2m}{m}} - \frac{\binom{2m-u}{m+1-l}\mathbb{1}_{I_2} + \binom{2m-u}{m-1-l}\mathbb{1}_{I_3}}{2\binom{2m}{m-1}} \right| \\
&=: \sum_{u=1}^{2m} \frac{\Upsilon_u}{(2m)^n} \sum_{l=0}^{u} \varrho_l
\end{aligned}$$

where $I_1 = \{l : \max(u-m, 0) \le l \le \min(m, u)\}$, $I_2 = \{l : \max(u-m+1, 0) \le l \le \min(m+1, u)\}$, and $I_3 = \{l : \max(u-m-1, 0) \le l \le \min(m-1, u)\}$, and by defining

$$\varrho_l := \binom{u}{l} \left| \frac{\binom{2m-u}{m-l}\mathbb{1}_{I_1}}{\binom{2m}{m}} - \frac{\binom{2m-u}{m+1-l}\mathbb{1}_{I_2} + \binom{2m-u}{m-1-l}\mathbb{1}_{I_3}}{2\binom{2m}{m-1}} \right|.$$

Note that for $u \le m-1$, all these range becomes $0 \le l \le u$; but for $u \ge m$, we will have certain differences. For instance, if $u > m$, then $l = u-m-1 \in I_3$ (but not in I_1 and I_2), $l = u - m \in I_1$ and I_3 but not in I_2. $l = m \in I_1$ and I_2 but not in I_3, $l = m+1 \in I_2$ but not in I_1 and I_3.

Acknowledgement

The authors are grateful to David Pollard for valuable comments and suggestions.

References

Cai, T. T. and M. G. Low (2011). Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *Annals of Statistics 39*(2), 1012–1041. MR2816346

Chvátal, V. (1979). The tail of the hypergeometric distribution. *Discrete Mathematics 25*(3), 285–287. MR0534946

Genovese, C., M. Perone-Pacifico, I. Verdinelli, and L. Wasserman (2012a). Minimax manifold estimation. *Journal of machine learning research* (3), 1263–1291. MR2930639

Genovese, C. R., M. Perone-Pacifico, I. Verdinelli, and L. Wasserman (2012b). Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics 40*(2), 941–963. MR2985939

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *The Annals of Statistics 58*(301), 13–30. MR0144363

Kamath, A., R. Motwani, K. Palem, and P. Spirakis (1994). Tail bounds for occupancy and the satisfiability threshold conjecture. In *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on*, pp. 592–603. MR1346284

Kolchin, V. F., B. A. Sevast'yanov, and V. P. Chistyakov (1978). *Random allocations*. Washington, D.C.: V. H. Winston & Sons. Translated from the Russian, Translation edited by A. V. Balakrishnan, Scripta Series in Mathematics. MR0471016

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics 1*, 38–53. MR0334381

Niyogi, P., S. Smale, and S. Weinberger (2006). Finding the homology of submanifolds with high confidence from random samples. *Discrete and computational geometry 39*, 419–441. MR2383768

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer-Verlag. MR2724359

Yu, B. (1997). Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. L. Yang (Eds.), *A Festschrift for Lucien Le Cam*, pp. 423–435. New York: Springer-Verlag. MR1462963