GENERALIZED CLUSTER TREES AND SINGULAR MEASURES

By Yen-Chi Chen

Department of Statistics, University of Washington

October 12, 2016

In this paper, we study the α-cluster tree (α-tree) under both singular and nonsingular measures. The α tree is to use probability contents within a level set to construct a cluster tree so that it is well-defined for singular measures. We first derive the rate of convergence for a density level set around critical points, which leads to the convergence rate of estimating an α-tree under nonsingular measures. For singular measures, we study how the kernel density estimator (KDE) behaves and prove that KDE is not uniformly consistent but pointwisely consistent after rescaling. We further prove that the estimated α-tree fails to converge in the L_∞ metric but is still consistent under the integrated distance. We also observe a new type of critical points–the dimensional critical points (DCPs)–of a singular measure. These DCPs occur only at singular measures and similar to the usual critical points, DCPs contribute to the topology of a cluster tree as well. Building upon the analysis for the KDE and the DCPs, we prove a topological consistency of the estimated α-tree.

1. Introduction. Given a function f defined on a smooth manifold $\mathcal{M}$, the cluster tree of f is a tree structure representing the creation and merging of connected components of the level set $\{x : f(x) \geq \kappa\}$ when we move down the level κ (Chen et al., 2016b). Because a cluster tree keeps track of the connected components of the level sets, the shape of a cluster tree contains topological information about the underlying function f. And a cluster tree can be displayed on a two-dimensional plane regardless of the dimension of $\mathcal{M}$; this makes it an attractive method for visualizing the function f under multivariate case. In this paper, we focus on the case where $f \equiv f_P$. Namely, f is some functional of the underlying distribution P. In this context, the cluster tree of f reveals information about the distribution P.

In most of the cluster tree literatures, the cluster trees being studied are the λ-tree of a distribution (Balakrishnan et al., 2012; Chaudhuri and Dasgupta, 2010; Chaudhuri et al., 2014; Chen et al., 2016b; Kpotufe and Luxburg, 2011; Stuetzle, 2003). The λ-tree of a distribution is to choose f to be p, the underlying density function. In this case, the tree structure contains the topological information of the underlying density function and we can use the λ-tree to visualize a multivariate density function; when we use the λ-tree for visualization purposes, the λ-tree is also called a density tree (Klemelä, 2004, 2006, 2009).

In Kent (2013), the author proposed a new type of cluster tree of a distribution–the α-tree. The α-tree is to use the function $\alpha(x) = P(\{y : p(y) \leq p(x)\})$ to construct a cluster tree. When the distribution is nonsingular, the α-tree and the λ-tree are topologically equivalent (Lemma 1) so they both provide similar topological information of the underlying distribution. To estimate the α-tree, we use the cluster tree of the function estimator $\widehat{\alpha}_n(x) = \widehat{P}_n(\{y : \widehat{p}_n(y) \leq \widehat{p}_n(x)\})$ where $\widehat{P}_n$ is the empirical measure and

MSC 2010 subject classifications: Primary 62G20; secondary 62G05, 62G07

Keywords and phrases: cluster tree, kernel density estimator, level set, singular measure, critical points, topological data analysis

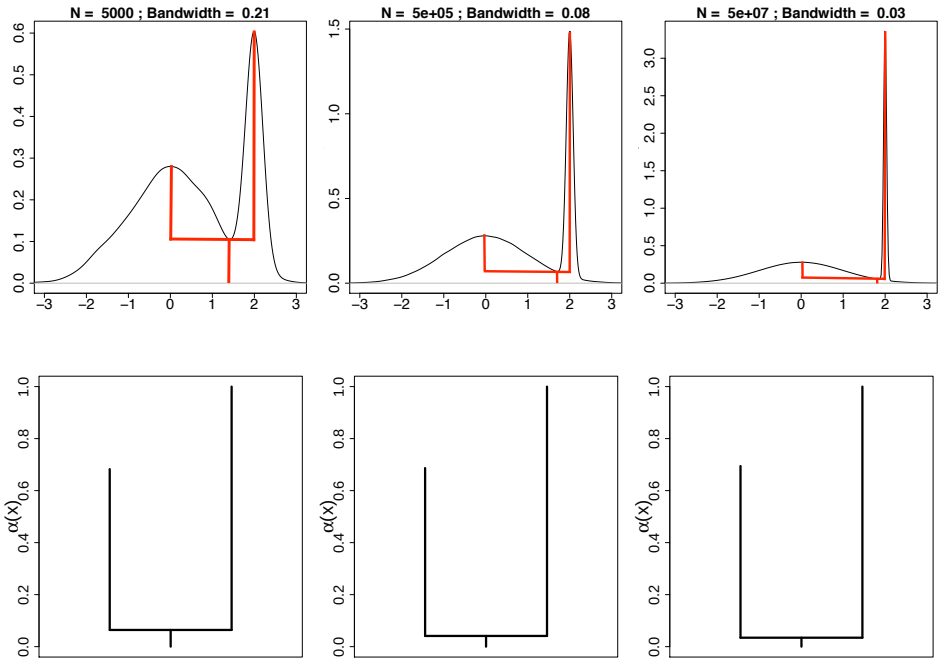

FIG 1. *An example of the estimated λ-tree and α-tree of a singular distribution. This is a random sample from a singular distribution that with probability* 0.3 *we obtain a point mass at $x = 2$ and with probability* 0.7 *we sample from a standard normal. The top panel shows the density estimated by the kernel density estimator (KDE) and the red tree structure corresponds to the estimated λ-tree. The bottom panel displays the estimated α-tree. From left to right, we increase the sample size from 5×10^3, 5×10^5, to 5×10^7. Because the distribution is singular, there is no population λ-tree so when the smoothing bandwidth decreases (when the sample size increases), the estimate λ-tree is getting degenerated. On the other hand, the estimated α-trees remain stable regardless of the smoothing bandwidth.*

$\widehat{p}_n$ is the kernel density estimator (KDE). Namely, we first use the KDE to estimate the density of each data points and then count the number of data points whose density is below the given point.

When the distribution is singular, the λ-tree is ill-defined due to the lack of density but the α-tree is still well-defined under a mild modification. As an illustrating example, consider Figure 1. These are random samples from a distribution that mixed with a point mass at $x = 2$ with probability 0.3 and a standard normal distribution with probability 0.7. Thus, these samples are from a singular distribution. We generate $n = 5 \times 10^3$ (left), 5×10^5 (middle) and 5×10^7 (right) data points and estimate the density using the KDE. The estimated density along with the estimated λ-trees (red trees) are displayed in the top row. It can be seen easily that when the sample size increases, the λ-trees become degenerated. This is because there is no population λ-tree for this distribution. However, the α-trees are stable in all three panels (see the bottom row of Figure 1); this shows the power of α-trees.

Main Results. The main results of this paper are summarized as follows:

- When the distribution is nonsingular,
  1. we derived the rate of convergence for the estimated level set when the level equals to the density value of a critical point (Theorem 3).
  2. we derived the rate of convergence of $\widehat{\alpha}_n$ (Theorem 4).

- When the distribution is singular,
  3. we propose a framework that generalizes $\alpha(x)$ to define the α-tree (Section 4);
  4. we show that after rescaling, the KDE is pointwisely consistent but not uniformly consistent (Theorem 8);
  5. we prove that $\widehat{\alpha}_n$ is inconsistent under the L_∞ metric (Corollary 7) but is consistent under the integrated distance and probability-weighted integrated distance (Theorem 10);
  6. we identify a new type of critical points, the dimensional critical points (DCPs), that also contributes to the change of cluster tree topology and analyze their properties (Lemma 11, 13, and 14);
  7. we demonstrate that the the estimated α-tree $T_{\widehat{\alpha}_n}$ is topological equivalent to the population α-tree with probability exponentially converging to 1 (Theorem 15).

Related Work. There are a vast literatures about theoretical aspects of the λ-tree; notions of consistency are analyzed in Hartigan (1981); Chaudhuri and Dasgupta (2010); Chaudhuri et al. (2014); Eldridge et al. (2015); the rate of convergence and the minimax theory are studied in Chaudhuri and Dasgupta (2010); Balakrishnan et al. (2012); Chaudhuri et al. (2014); in Chen et al. (2016b), the authors study how to perform statistical inference for a λ-tree. The cluster tree is also related to the topological data analysis (Carlsson, 2009; Edelsbrunner and Morozov, 2012); in particular, a cluster tree contains the information about the zeroth order homology groups (Cohen-Steiner et al., 2007; Fasy et al., 2014; Bobrowski et al., 2014; Bubenik, 2015). The theory of estimating a cluster tree is closely related to the theory of estimating a level set; an incomplete list of literatures is as follows: Polonik (1995); Tsybakov (1997); Walther (1997); Mason and Polonik (2009); Singh et al. (2009); Rinaldo and Wasserman (2010); Steinwart (2011).

Outline. We begin with an introduction about cluster trees and geometric concepts used in this paper in Section 2. We derive the convergence rate for the α-tree estimator under nonsingular measures in Section 3. We study the behavior of the kernel density estimator and the stability of the estimated α-tree under singular measures in Section 4. In Section 5, we investigate critical points of singular measures and derive topological consistency of the estimated α-tree. We summarize this paper and discuss possible future directions in Section 6.

2. Backgrounds.

2.1. *Cluster Trees.* Here we recalled the definition of cluster tree in Chen et al. (2016b). Let $\mathbb{K} \subset \mathbb{R}^d$ and $f : \mathbb{K} \mapsto [0, \infty)$ be a function with support $\mathbb{K}$. The cluster tree of f is defined as follows.

DEFINITION 1 (Definition 1 in Chen et al. (2016b)). *For any $f : \mathbb{K} \mapsto [0, \infty)$ the* cluster tree *of f is a function $T_f : \mathbb{R} \mapsto 2^{\mathbb{K}}$, where $2^{\mathbb{K}}$ denotes the set of all subsets of $\mathbb{K}$, and $T_f(\lambda)$ is the set of the connected components of the upper-level set $\{x \in \mathbb{K} : f(x) \geq \lambda\}$. We define the collection of connected components $\{T_f\}$, as $\{T_f\} = \bigcup_\lambda T_f(\lambda)$. Thus, $\{T_f\}$ is a collection of subsets of $\mathbb{K}$ indexed by λ.*

It is easy to see that the cluster tree T_f has a tree structure, because for every pair $C_1, C_2 \in T_f$, either $C_1 \subset C_2$, $C_2 \subset C_1$, or $C_1 \cap C_2 = \phi$ holds.

To get some geometric understanding of the cluster tree in Definition 1, we identify edges that constitute the cluster tree. Intuitively, edges correspond to either leaves or internal branches. An edge is roughly defined as a set of clusters whose inclusion relationship with respect to clusters outside an edge are

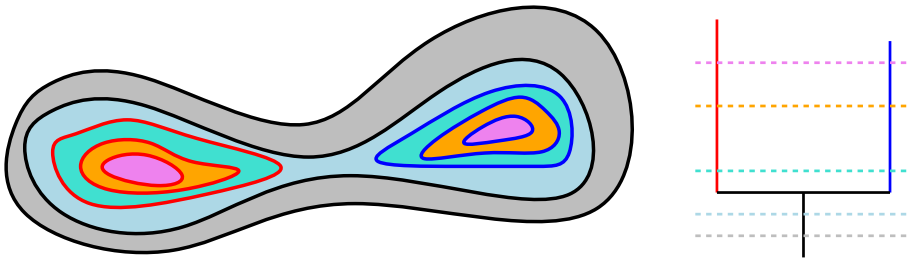

FIG 2. *Connected components, edges, and edge set of a cluster tree. Left: we display connected components of level sets under five different levels (indicated by the colors: magenta, orange, seagreen, skyblue, and gray). The color of boundaries of each connected component denotes the edge they corresponds to. Right: the cluster tree; we color the three edges (vertical lines) by red, blue, and black. The edge set* $E(T_f) = \{\mathbb{C}_{\mathsf{red}}, \mathbb{C}_{\mathsf{blue}}, \mathbb{C}_{\mathsf{black}}\}$ *and we have the ordering* $\mathbb{C}_{\mathsf{red}} \leq \mathbb{C}_{\mathsf{black}}$ *amd* $\mathbb{C}_{\mathsf{blue}} \leq \mathbb{C}_{\mathsf{black}}$*. Note that the solid black horizontal line is not an edge set; it is just a visual representation to connect the blue and red edges to the black edge. The horizontal dashed lines indicates the five levels corresponds to the the left panel. In the left panel, the three connected components with red boundaries are elements of the edge* $\mathbb{C}_{\mathsf{red}}$*.*

equivalent, so that when the collection of connected components is divided into edges, we observe the same inclusion relationship between representative clusters whenever any cluster is selected as representative for each edge.

To formally define edges, we define an interval in the cluster tree, and the equivalence relation in the cluster tree. For any two clusters $A, B \in \{T_f\}$, the interval $[A, B] \subset \{T_f\}$ is defined as a set clusters that contain A and are contained in B, i.e.

$$[A, B] := \{C \in \{T_f\} : A \subset C \subset B\},$$

The equivalence relation $\sim$ is defined as $A \sim B$ if and only if their inclusion relationship with respect to clusters outside $[A, B]$ and $[B, A]$. Namely, $A \sim B$ if and only if

$$\begin{aligned} &\forall C \in \{T_f\} \text{ such that } C \notin [A, B] \cup [B, A], \\ &\qquad\qquad C \subset A \Leftrightarrow C \subset B, \quad A \subset C \Leftrightarrow B \subset C. \end{aligned}$$

It is easy to see that the relation $\sim$ is reflexive($A \sim A$), symmetric($A \sim B$ implies $B \sim A$), and transitive ($A \sim B$ and $B \sim C$ implies $A \sim C$). Hence the relation $\sim$ is indeed an equivalence relation, and we can consider the set of equivalence classes $\{T_f\}/_\sim$. We define the edge set (the collection of edges) $E(T_f)$ as $E(T_f) := \{T_f\}/_\sim$. Each element in the edge set $\mathbb{C} \in E(T_f)$ is called an edge and and edge contains many nested connected components of the cluster tree $\{T_f\}$ (i.e. if $C_1, C_2 \in \mathbb{C}$, then either $C_1 \subset C_2$ or $C_2 \subset C_1$). Note that every element in an edge corresponds to a connected component of an upper level set of f.

To associate the edge set $E(T_f)$ to a tree structure, we define a partial order on the edge set as follows: let $\mathbb{C}_1, \mathbb{C}_2 \in E(T_f)$ be two edges, we write $\mathbb{C}_1 \leq \mathbb{C}_2$ if and only if for all $A \in \mathbb{C}_1$ and $B \in \mathbb{C}_2$, $A \subset B$. Then the shape of the cluster tree (topology of the cluster tree) is completely determined by the edge set $E(T_f)$ and the partial order among them. Figure 2 provides an example about the connected components, the edges, and the edge set of a cluster tree along with a tree representation.

Based on the above definitions, we define the topological equivalence between two cluster trees.

DEFINITION 2. *For two functions* $f : \mathbb{K} \mapsto [0, \infty)$ *and* $g : \mathbb{K} \mapsto [0, \infty)$*, we say* T_f *and* T_g *are topological equivalent, denoted as* $T_f \overset{T}{\approx} T_g$*, if there exists a bijective mapping* $S : E(T_f) \mapsto E(T_g)$ *such that for any*

$\mathbb{C}_1, \mathbb{C}_2 \in E(T_f)$,

$$\mathbb{C}_1 \leq \mathbb{C}_2 \Longleftrightarrow S(\mathbb{C}_1) \leq S(\mathbb{C}_2).$$

For each $\mathbb{C} \in E(T_f)$, we define

$$U(\mathbb{C}) = \sup\{\lambda : \exists C \in T_f(\lambda), C \in \mathbb{C}\}$$

to be the maximal level of an edge $\mathbb{C}$. We define the *critical tree-levels* of f as

$$\mathcal{A}_f = \{U(\mathbb{C}) : \mathbb{C} \in E(T_f)\}. \tag{1}$$

It is easy to see $\mathcal{A}_f$ is collection levels of f where a creation of new connected component occurs or a merging of two connected components occurs.

In most of the cluster tree literatures (Balakrishnan et al., 2012; Chaudhuri and Dasgupta, 2010; Chaudhuri et al., 2014; Chen et al., 2016b; Eldridge et al., 2015), the cluster tree is referred to the λ-tree, which is to use the probability density function p to build a cluster tree. Namely, the λ-tree is T_p.

In this paper, we focus on the α-tree (Kent, 2013), which is to use the function

$$\alpha(x) = P(\{y : p(y) \leq p(x)\})1 - P(\{y : p(y) > p(x)\}) = 1 - P(L_{p(x)}) \tag{2}$$

to build the cluster tree T_α. The set $L_\lambda = \{x : p(x) \geq \lambda\}$ is the upper level set of p (note that $P(\{y : p(y) > p(x)\}) = P(\{y : p(y) \geq p(x)\})$ when the density function p is bounded). The cluster tree T_α is called the *α-tree*. A feature for α-tree is that the function $\alpha(x)$ depends only on the 'ordering' of points within $\mathbb{K}$. Namely, any function that assigns the same ordering to points within $\mathbb{K}$ as the density function p can be used to construct the function $\alpha(x)$. To be more specific, let ℓ be an ordering such that for any $x_1, x_2 \in \mathbb{K}$,

$$\begin{aligned}
\ell(x_1) > \ell(x_2) &\Leftrightarrow p(x_1) > p(x_2),\\
\ell(x_1) < \ell(x_2) &\Leftrightarrow p(x_1) < p(x_2),\\
\ell(x_1) = \ell(x_2) &\Leftrightarrow p(x_1) = p(x_2).
\end{aligned}$$

Then

$$\alpha(x) = 1 - P(\{y : p(y) \geq p(x)\}) = 1 - P(\{y : \ell(y) \geq \ell(x)\}). \tag{3}$$

For instance, $\ell(x) = 2p(x)$, or $\ell(x) = \log p(x)$ both yield the same $\alpha(x)$. Later we will use this feature to generalize equation (2) to singular measures.

A feature for the α-tree is that it is topological equivalent to the λ-tree.

LEMMA 1. *Assume the distribution P has density p. Then the λ-tree and α-tree are topological equivalent. Namely,*

$$T_p \overset{T}{\approx} T_\alpha.$$

The proof is simple so we ignore it; the main idea is that by equation (2), $\alpha(x)$ is just a monotonic transformation of the density p the topology are preserved.

When we use the α-tree, the induced upper level set

$$\mathbb{A}_a = \{x : \alpha(x) \geq a\}$$

is called an *α-level set*.

REMARK 1 (κ-tree). In Kent (2013), the author also proposed another cluster tree–the κ-tree–which is to use the probability content within each edge set defined by an α-tree (or a λ-tree) to compute the function $\kappa(x)$. Because it is just a rescaling from the α-tree, the theory of α-tree also works for κ-tree. For simplicity, we only study the theory of α-tree in this paper.

2.2. *Singular Measure.* When the probability measure is singular, the λ-tree is no longer well-defined since there is no density function. However, the α-tree can still be defined.

A key feature for constructing the α-tree is the ordering function $\ell(x)$. Here we will use a generalized density function, the Hausdorff density (Preiss, 1987; Mattila, 1999), to define the α-tree under singular measures. Given a probability measure P, the s-density (s dimensional Hausdorff density) is

$$\mathcal{H}_s(x) = \lim_{r \to 0} \frac{P(B(x,r))}{C_s \cdot r^s},$$

where C_s is the volume of an unit s-dimensional sphere and $B(x,r) = \{y : \|y - x\| \leq r\}$.

For a given point x, we define the notion of a generalized density using two quantities $\tau(x)$ and $\rho(x)$:

$$\begin{aligned} \tau(x) &= \mathsf{argmax}_{s \leq d} \mathcal{H}_s(x) < \infty \\ \rho(x) &= \mathcal{H}_{\tau(x)}(x). \end{aligned}$$

Namely, $\tau(x)$ is the 'dimension' of the probability measure at x and $\rho(x)$ is the corresponding Hausdorff density at that dimension. Note that the function $\rho(x)$ is well-defined for every x. For any two points $x_1, x_2 \in \mathbb{K}$, we define an ordering ℓ by $\ell(x_1) > \ell(x_2)$ if

$$\tau(x_1) < \tau(x_2), \qquad \text{or} \quad \tau(x_1) = \tau(x_2), \quad \rho(x_1) > \rho(x_2).$$

That is, for any pair of points, we first compare the their 'dimensions' $\tau(x)$. The point with lower dimensional value τ will be ranked higher than the other point. If two points have the same dimensions, then we compare their corresponding Hausdorff density. When the distribution is non-singular, $\tau(x) = d$ for every $x \in \mathbb{K}$ and $\rho(x) = p(x)$ is the usual density function. So the ordering $\ell(x)$ can be chosen simply as the density function $p(x)$.

To define the α-tree, we use equation (3):

$$\alpha(x) = P\left(\{y : \ell(y) \geq \ell(x)\}\right). \tag{4}$$

Namely, $\alpha(x)$ is the probability content of regions where the ordering function ℓ is lower than or equal to $\ell(x)$. As is shown in equation (3), when P is non-singular, equation (4) is the same as equation (2). Note that by equation (4), the α-level set $\mathbb{A}_a = \{x : \alpha(x) \geq a\}$ is well-defined in singular measure.

2.3. *Geometric Concepts.* Based on the definition of $\tau(x)$, we decompose the support $\mathbb{K}$ into

$$\mathbb{K} = \mathbb{K}_d \bigcup \mathbb{K}_{d-1} \bigcup \cdots \bigcup \mathbb{K}_0, \tag{5}$$

where $\mathbb{K}_s = \{x : \tau(x) = s\}$. Thus, $\{\mathbb{K}_0, \cdots, \mathbb{K}_d\}$ forms a partition of the entire support $\mathbb{K}$. We call each $\mathbb{K}_s$ an s-dimensional support (structure). When we analyze the support $\mathbb{K}_s$, any $\mathbb{K}_{s'}$ with $s' > s$ is called a higher dimensional support (with respect to $\mathbb{K}_s$) and $s' < s$ will be called a lower dimensional support.

To regularize the behavior of $\rho(x)$ on each support $\mathbb{K}_s$, we assume that the closure of the support $\overline{\mathbb{K}}_s$ is an s-dimensional smooth manifold (properties about a smooth manifold can be found in Lee 2012; Tu 2010). For a s-dimensional smooth manifold $\mathcal{M}$, the tangent space on each point of $\mathcal{M}$ changes smoothly (Tu, 2010; Lee, 2012). Namely, for $x \in \mathcal{M}$, we can find an orthonormal basis $\{v_1(x), \cdots, v_s(x) : v_\ell(x) \in \mathbb{R}^d, \ell = 1, \cdots, s\}$ such that the tangent space of $\mathcal{M}$ at x is spanned by $\{v_1(x), \cdots, v_s(x)\}$ and each $v_\ell(x)$ is a smooth (multivalued) function on $\mathcal{M}_s$. For simplicity, for $x \in \mathbb{K}_s$, we denote $T_s(x)$ as the tangent space of $\mathbb{K}_s$ at x and $N_s(x)$ as the normal space of $\mathbb{K}_s$ at x. And we define $\nabla_{T_s(x)}$ to be taking derivative in the tangent space.

For a function $f : \mathcal{M} \mapsto \mathbb{R}$ defined on a smooth manifold $\mathcal{M}$, the function f is a *Morse function* (Milnor, 1963; Morse, 1925, 1930) if all critical points of f are non-degenerate. Namely, the eigenvalues of the Hessian matrix of f at each critical point is away from zero. Being a Morse function is essential for a density function to have a stable λ-tree (Chazal et al., 2014; Chen et al., 2016b).

To link the concept of Morse function to the Hausdorff density $\rho(x)$, we introduce a generalized density

$$\rho_s^\dagger : \overline{\mathbb{K}}_s \mapsto [0, \infty)$$

such that $\rho_s^\dagger(x) = \lim_{x_n \in \mathbb{K}_s : x_n \to x} \rho(x_n)$. It is easy to see that $\rho_s^\dagger(x) = \rho(x)$ when $x \in \mathbb{K}_s$ but now it is defined on a smooth manifold $\overline{\mathbb{K}}_s$. We say $\rho(x)$ is a *generalized Morse function* if the corresponding $\rho_s^\dagger(x)$ is a Morse function for $s = 1, \cdots, d$. Later we will show that this generalization leads to a stable α-tree for a singular measure.

For $\overline{\mathbb{K}}_s$, let $\mathcal{C}_s = \{x \in \overline{\mathbb{K}} : \nabla_{T_s(x)} \rho_s^\dagger(x) = 0\}$ be the collection of its critical points. Then the fact that $\rho_s^\dagger(x)$ is a Morse function implies the eigenvalues of the Hessian matrix $\nabla_{T_s(c)} \nabla_{T_s(c)} \rho_s^\dagger(c)$ are non-zero for every $c \in \mathcal{C}_s$. Note that we called $g_s(x) = \nabla_{T_s(x)} \rho_s^\dagger(x)$ the generalized gradient and $H_s(x) = \nabla_{T_s(x)} \nabla_{T_s(x)} \rho_s^\dagger(x)$ the generalized Hessian. For the case $s = 0$ (point mass), we define $\mathcal{C}_0 = \mathbb{K}_0$. The the collection $\mathcal{C} = \bigcup_{s=1,\cdots,d} \mathcal{C}_s$ is called the collection of *generalized critical points* of $\rho(x)$. And each element $c \in \mathcal{C}$ is called a generalized critical point.

Finally, we introduce the concept of *reach* (Federer, 1959; Chen et al., 2015a) for a smooth manifold $\mathcal{M}$. The reach of $\mathcal{M}$ is defined as

$$\mathsf{reach}(\mathcal{M}) = \sup\{r \geq 0 : \text{every point in } \mathcal{M} \oplus r \text{ has an unique projection onto } \mathcal{M}\},$$

where $A \oplus r = \{x : d(x, A) \leq r\}$. One can view the reach as the radius of the largest ball that can roll freely outside $\mathcal{M}$. More details about reach can be found in Federer (1959); Chen et al. (2015a). Reach plays a key role in the stability of a level set; see Chen et al. (2015a) for more details.

2.4. *Estimating the α function and the α-tree.* In this paper, we focus on estimating the α-trees via the kernel density estimator (KDE):

$$\widehat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\|x - X_i\|}{h}\right).$$

Specifically, we first estimate the density by $\widehat{p}_n$ and then construct the estimator $\widehat{\alpha}_n$:

$$\widehat{\alpha}_n(x) = \widehat{P}_n\left(\{y : \widehat{p}_n(y) \leq \widehat{p}_n(x)\}\right) \tag{6}$$

where $\widehat{P}_n(A)$ is the empirical measure and $\widehat{L}_\lambda = \{x : \widehat{p}_n(x) \geq \lambda\}$. Note that when x does not contain any point mass of P, $\widehat{\alpha}_n(x) = 1 - \widehat{P}_n\left(\widehat{L}_{\widehat{p}_n(x)}\right)$.

To quantify the uncertainty in the estimator $\widehat{\alpha}_n$, we consider three error measurements. The first error measurement is the L_∞ *error*, which is defined as

$$\|\widehat{\alpha}_n - \alpha\|_\infty = \sup_x |\widehat{\alpha}_n(x) - \alpha(x)|.$$

The L_∞ error has been used in several cluster tree literatures; see, e.g., Eldridge et al. (2015); Chen et al. (2016b). An appealing feature of L_∞ error is that this quantity is the same (up to some constant) as some other tree error metrics such as the merge distortion metric (Eldridge et al., 2015). And the convergence in the merge distortion metric implies the Hartigan consistency (Eldridge et al., 2015), a notion of consistency of a cluster tree estimator described in Hartigan (1981); Chaudhuri and Dasgupta (2010); Chaudhuri et al. (2014). Thus, due to the equivalence between L_∞ error and the merge distortion metric, convergence in L_∞ implies the Hartigan consistency of an estimated cluster tree.

The other two errors are the integrated error and the probability error (probability-weighted integrated error). Both are common error measurements for evaluating the quality of a function estimator (Wasserman, 2006; Scott, 2015). The *integrated error* is

$$\|\widehat{\alpha}_n - \alpha\|_\mu = \int |\widehat{\alpha}_n(x) - \alpha(x)| dx,$$

which is also known as the integrated distance or L_1 distance. The *probability error (probability-weighted integrated error)* is

$$\|\widehat{\alpha}_n - \alpha\|_P = \int |\widehat{\alpha}_n(x) - \alpha(x)| dP(x),$$

is the integrated distance weighted by the probability measure, which is also known as $L_1(P)$ distance. The integrated error and the probability error are more robust than the L_∞ error–a large difference in a small region will not have much impact on these errors.

To quantify the uncertainty in the topology of α-tree, we introduce the notion of *topological error*, which is defined as

$$P\left(T_{\widehat{\alpha}_n} \overset{T}{\not\approx} T_\alpha\right) = 1 - P\left(T_{\widehat{\alpha}_n} \overset{T}{\approx} T_\alpha\right).$$

Namely, the topological error is the probability that the estimated α-tree is not topological equivalent to the population α-tree.

Finally, we define the following notations. For a smooth function p, we define $\|p\|_{\ell,\infty}$ as the supremum maximal norm of ℓ-th derivative of p. For instance, $\|p\|_{0,\infty} = \sup_{x\in\mathbb{K}} p(x)$, $\|p\|_{1,\infty} = \sup_{x\in\mathbb{K}} \|g\|_{\max}$, and $\|p\|_{2,\infty} = \sup_{x\in\mathbb{K}} \|H\|_{\max}$, where $g(x) = \nabla p(x)$ and $H(x) = \nabla\nabla p(x)$ are the gradient and Hessian matrix. For sets A and B, define $A \triangle B = (A \backslash B) \cup (B \backslash A)$ to be their symmetric difference. A vector $\alpha = (\alpha_1, \ldots, \alpha_d)$ of non-negative integers is called a multi-index with $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ and the corresponding derivative operator is

$$D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}, \tag{7}$$

where $D^\alpha f$ is often written as $f^{(\alpha)}$.

3. Theory for Nonsingular Measures. To study the theory for nonsingular measures, we make the following assumptions.

Assumptions.

(P1) p has a compact support $\mathbb{K}$ and is a Morse function with $\|p\|_{\ell,\infty} < \infty$ for $\ell = 0, 1, 2$.

(K1) $K(x)$ has compact support and is non-increasing on $[0, 1]$, and $\int K(\|x\|)dx = 1$ and third order partial derivatives of $\mathbb{K}(\|x\|)$ exists.

(K2) Let

$$\mathcal{K}_r = \left\{ y \mapsto K^{(\alpha)}\left(\frac{\|x-y\|}{h}\right) : x \in \mathbb{R}^d, |\alpha| = r \right\},$$

where $K^{(\alpha)}$ is defined in (7) and let $\mathcal{K}_l^* = \bigcup_{r=0}^{l} \mathcal{K}_r$. We assume that $\mathcal{K}_2^*$ is a VC-type class. i.e. there exists constants A, v and a constant envelope b_0 such that

$$\sup_Q N(\mathcal{K}_2^*, \mathcal{L}^2(Q), b_0\epsilon) \leq \left(\frac{A}{\epsilon}\right)^v, \tag{8}$$

where $N(T, d_T, \epsilon)$ is the ϵ-covering number for an semi-metric set T with metric d_T and $\mathcal{L}^2(Q)$ is the L_2 norm with respect to the probability measure Q.

Assumption (P1) is a common condition to guarantee the stability of critical points (Chazal et al., 2014; Chen et al., 2016b). Assumption (K1) is a standard condition on kernel function (Wasserman, 2006; Scott, 2015). Assumption (K2) is to regularize the complexity of kernel functions so that we have uniform bounds on the density, gradient, and Hessian estimation; it was first proposed in Giné and Guillou (2002) and Einmahl and Mason (2005) and later was used in various literatures such as Genovese et al. (2009, 2014); Chen et al. (2015b).

We first study the error rates under nonsingular measures. For the case of λ-tree, the error rates are well-studied and here we summarize them in the following theorem.

THEOREM 2. *Assume (P1, K1–2). Then*

$$\begin{aligned}
\|\widehat{p}_n - p\|_\infty &= O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right) \\
\|\widehat{p}_n - p\|_\mu &= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) \\
\|\widehat{p}_n - p\|_P &= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) \\
P\left(T_{\widehat{p}_n} \overset{T}{\approx} T_p\right) &\geq 1 - e^{-C_0 \cdot nh^{d+4}},
\end{aligned}$$

for some $C_0 > 0$.

The rate of consistency under L_∞ error can be found in Chen et al. (2015a); Giné and Guillou (2002); Einmahl and Mason (2005); the integrated error and probability error can be seen in Scott (2015); and

topological error bound follows from Lemma 2 in Chen et al. (2016b) and the concentration of L_∞ distance, see, e.g., Theorem 9 in Chen et al. (2015a).

Now we turned to the consistency for α-tree. To derive the rate for the α-tree, we need to study the convergence rate of estimating a level set when the level is the density value of a critical point (also known as a critical level). The reason is that the quantity $\alpha(x) = 1 - P(L_{p(x)})$ is the probability content of upper level set $L_{p(x)} = \{y : p(y) \geq p(x)\}$. When $p(x) = p(c)$ for some critical point c of p, we face the problem of analyzing the stability of level sets at a critical level.

THEOREM 3 (Level set error at a critical value). *Assume (P1) and (K1–2) and $d \geq 2$. Let λ be a density level corresponds to the density of a critical point. Then*

$$\mu\left(\widehat{L}_\lambda \triangle L_\lambda\right) = O_P\left(\|\widehat{p}_n - p\|_\mu^{\frac{d}{d+1}}\right),$$

where μ is the Lebesgue measure.

The rate in Theorem 3 is slower than the usual density estimation rate. This is because when λ equals to the density of a critical point, the boundary of L_λ hits a critical point. The regions around a critical point has very low gradient, which lead to a slower rate of convergence. Note that it is well-known (Wasserman, 2006; Scott, 2015) that under assumption (P) and (K1),

$$\|\widehat{p}_n - p\|_\mu = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right).$$

REMARK 2. The Hausdorff distance when λ is the density of a saddle point is $\mathsf{Haus}\left(\widehat{L}_\lambda, L_\lambda\right) = O_P\left(\|\widehat{g}_n - g\|_\infty\right)$, where

$$\mathsf{Haus}(A, B) = \max\left\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\right\}$$

and $d(x, A) = \inf_{y\in A} \|x - y\|$. Thus, the symmetric distance and the Hausdorff distance have different rate of convergence. The reason is that the rate of Hausdorff distance is dominated by the uncertainty around critical points, which is at the gradient error rate. When we consider the symmetric distance $\mu(A \triangle B)$, the uncertainty around critical point is a minor contribution that can be ignored.

REMARK 3. Theorem 3 complements to many existing level set estimation theories. To our knowledge, no literature has worked on the situation where λ equals to the density of a critical point. Theories of level sets mostly focus on one of the following three cases: (i) the gradient on the boundary of level set $\partial L_\lambda = \{x : p(x) = \lambda\}$ is bounded away from 0 (Molchanov, 1991; Tsybakov, 1997; Walther, 1997; Cadre, 2006; Laloe and Servien, 2013; Mammen and Polonik, 2013; Chen et al., 2015a), (ii) a lower bound on the density changing rate around the level λ (Singh et al., 2009; Rinaldo et al., 2012), (iii) an (ϵ, σ) conditions for density (Chaudhuri and Dasgupta, 2010; Chaudhuri et al., 2014). When λ equals to a critical level, none of these assumptions holds.

Based on Theorem 3, we derive the convergence rate of $\widehat{\alpha}_n$.

THEOREM 4. *Assume (P1) and (K1–2) and $d \geq 2$. Let $\mathcal{C} = \{x : \nabla p(x) = 0\}$ be the collection of critical points and let a_n be a sequence of n such that $\|\widehat{p}_n - p\|_\infty = o(a_n)$. Then uniformly for all x,*

$$\widehat{\alpha}_n(x) - \alpha(x) = \begin{cases} O_P\left(\|\widehat{p}_n - p\|_\mu\right) & \text{, if } \ |p(x) - p(c)| > a_n \text{ for all } c \in \mathcal{C}, \\ O_P\left(\|\widehat{p}_n - p\|_\mu^{\frac{d}{d+1}}\right) & \text{, otherwise.} \end{cases}$$

Theorem 4 shows the uniform error rates for $\widehat{\alpha}_n$. When the given point whose density is away from critical levels, the rate follows the usual density estimation rate. When the given point has density value close to some critical points, the rate is slow down by the low gradient areas around critical points. Note that the sequence a_n is to make the bound uniformly for all x; to obtain an integrated error rate (and the probability error rate) of $\widehat{\alpha}_n$, we can choose $a_n = \frac{1}{\log n}\left(O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right)\right)$ which leads to the following result.

COROLLARY 5. *Assume (P1) and (K1–2) and $d \geq 2$. Then*

$$\begin{aligned} \|\widehat{\alpha}_n - \alpha\|_\infty &= O\left(h^{\frac{2d}{d+1}}\right) + O_P\left(\left(\frac{\log n}{nh^d}\right)^{\frac{d}{2(d+1)}}\right) \\ \|\widehat{\alpha}_n - \alpha\|_\mu &= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) \\ \|\widehat{\alpha}_n - \alpha\|_P &= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right) \\ P\left(T_{\widehat{\alpha}_n} \overset{T}{\approx} T_\alpha\right) &\geq 1 - e^{-C_0 \cdot nh^{d+2}}, \end{aligned}$$

for some $C_0 > 0$.

By comparing Corollary 5 to Theorem 2, the only difference is the L_∞ error rate. This is because Theorem 4 proves that only at level of a critical point we will have a slower rate of convergence. Thus L_∞ error will be slowed down by these points. However, the collection of points $\{x : p(x) = p(c)$ for some $c \in \mathcal{C}\}$ has Lebesgue measure 0 so the slow rate of convergence does not translate to the integrated error and probability error. The topological error follows from Theorem 2 and Lemma 1: $T(\widehat{p}_n) \overset{T}{\approx} T_{\widehat{\alpha}_n}, T(p) \overset{T}{\approx} T_\alpha$.

4. Singular Measures: Error Rates. Now we study the error rates under singular measures. When the measure is singular, the usual (Radon-Nikodym) density cannot be defined. Thus, we cannot define the λ-tree. However, as is discussed in Section 4, we are still able to define the α-tree. Thus, in this section, we will focus on the error rates for the α-tree.

4.1. *Analysis of the KDE under singular measures.* To study the rate of convergence, we first investigate the 'bias' of smoothing in the singular measure. Let $p_h(x) = \mathbb{E}(\widehat{p}_n)$, which is also known as the smoothed density.

Assumption.

(S) For all $s < d$, $\overline{\mathbb{K}}_s$ is a smooth manifold with positive reach.

(P2) $\rho(x)$ is a generalized Morse function and there exists some $\rho_{\min}, \rho_{\max} > 0$ such that $0 < \rho_{\min} \leq \rho(x) \leq \rho_{\max} < \infty$ for all x; moreover, for $s > 0$, $\rho_s^\dagger$ has bounded continuous derivatives up to the third order.

Assumption (S) is to regularize the lower dimensional support to make sure the Hausdorff density $\rho(x)$ behaves well within each $\mathbb{K}_s$. Assumption (P2) is a generalization of (P1) to singular distributions.

LEMMA 6 (Bias of the smoothed density). *Assume (S, P2). Let $x \in \mathbb{K}_s$ and define $m(x) = \min\{\ell \geq s : x \in \overline{K}_\ell\} - s$. Let $C_\ell^\dagger = (\int_{B_\ell} K(\|x\|)dx)^{-1}$, where $B_\ell = \{x : \|x\| \leq 1, x_{\ell+1} = x_{\ell+2} = \cdots = x_d = 0\}$ for $\ell = 1, \cdots, d$ and $C_0^\dagger = 1/K(0)$. Then for a fixed x, when $h \to 0$ and $m(x) > 0$,*

$$C_{\tau(x)}^\dagger h^{d-\tau(x)} \cdot p_h(x) = \rho(x) + \begin{cases} O(h^2) + O(h^{m(x)}), & if \quad m(x) > 0 \\ O(h^2), & if \quad m(x) = 0 \end{cases}.$$

Moreover, if $\overline{\mathbb{K}}_\ell \cap \mathbb{K}_s \neq \phi$, for some $s < \ell$, then there exists $\epsilon > 0$ such that

$$\limsup_{h \to 0} \sup_{x \in \mathbb{K}} |C_{\tau(x)} h^{d-\tau(x)} \cdot p_h(x) - \rho(x)| > \epsilon > 0.$$

Lemma 6 is a key result about the bias of the KDE. The scaling factor $C_{\tau(x)}^\dagger h^{d-\tau(x)}$ is to rescale the smoothed density to make it comparable to the generalized density. The first assertion is a pointwise convergence of smoothed density. In the case of $m(x) > 0$, the bias contains two components, the first one $O(h^2)$ is the usual smoothing bias and the second component $O(h^{m(x)})$ is the bias from 'higher' dimensional support. This is because the KDE is isotropic so the probability content outside $\mathbb{K}_s$ will also be included, which causes this additional bias. The second assertion states that the smoothed density does not uniformly converge to the generalized density $\rho(x)$; together with the first assertion, we conclude that the smoothing bias converges pointwisely but not uniformly. In what follows, we provide a concrete example showing the failure of uniform convergence of a singular measure.

EXAMPLE 1 (Failure of the uniform convergence). *Here is an example for the failure of the uniform convergence. We consider X from the same distribution as Figure 1: with probability 0.3, $X = 2$ and with probability 0.7, X follows a standard normal. For simplicity, we assume that the kernel function is the spherical kernel $K(x) = \frac{1}{2} I(0 \leq x \leq 1)$ and consider the smoothing bandwidth $h \to 0$. This choice of kernel yields $C_1^\dagger = 1$. Now consider a sequence of points $x_h = 1 + \frac{h}{2}$. Then the smoothed density at each x_h is*

$$\begin{aligned} p_h(x_h) &= \frac{1}{h} P(x_h - h < X < x_h + h) \\ &= \frac{1}{h} P\left(1 - \frac{h}{2} < X < 1 + \frac{3h}{2}\right) \\ &\geq \frac{1}{h} P\left(X = 1\right) = \frac{3}{10h}, \end{aligned}$$

which diverges when $h \to 0$. However, it is easy to see that $\tau(x_h) = 1$ and $\rho(x_h) = \frac{7}{10}\phi(x_h) \to \frac{7}{10}\phi(1)$ which is a finite number. Thus, $|\mathbb{E}(p_h(x_h) - \rho(x_h)|$ does not converge.

REMARK 4. The scaling factor in Theorem 6 $C^{\dagger}_{\tau(x)}h^{d-\tau(x)}$ depends on the support $\mathbb{K}_s$ where x resides in. In practice we do not know $\tau(x)$ so we cannot properly rescale $\widehat{p}_n(x)$ to estimate $\rho(x)$. However, we are still able to 'rank' pairs of data points based on Theorem 6. To see this, assume we want to recover the ordering of x_1 and x_2 using $\ell(x)$ (i.e. $\ell(x_1) > \ell(x_2)$ or $\ell(x_2) > \ell(x_1)$ or $\ell(x_1) = \ell(x_2)$). When x_1 and x_2 are both in $\mathbb{K}_s$ for some s, the scaling does not affect the ranking between them so the sign of $\rho(x_1) - \rho(x_2)$ is the same as the sign of $p_h(x_1) - p_h(x_2)$. When x_1 and x_2 are in different supports (i.e. $x_1 \in \mathbb{K}_{s_1}, x_2 \in \mathbb{K}_{s_2}$, where $s_1 \neq s_2$), $p_h(x_1)$ and $p_h(x_2)$ diverges at different rates so that eventually we can distinguish them. Thus, the ordering of points (for most points) can still be recovered under singular measure. This is an important property that leads to the consistency of $\widehat{\alpha}_n$ under other error measurements.

Due to the failure of uniform convergence in the bias, the L_∞ error of $\widehat{\alpha}_n$ does not converge under singular measures.

COROLLARY 7 (L_∞ error for singular meausres). *Assume (S, P2). When $\overline{\mathbb{K}}_d \cap \overline{\mathbb{K}}_s \neq \phi$, for some $s < d$, $\|\widehat{\alpha}_n - \alpha\|_\infty$ does not converges to 0. Namely, there exists $\epsilon > 0$ such that*

$$\liminf_{n,h} P(\|\widehat{\alpha}_n - \alpha\|_\infty > \epsilon) > 0.$$

The proof of Corollary 7 is a direct application of the failure of uniform convergence in smoothing bias in Theorem 6. This corollary shows that for a singular measure, the L_∞ error of the estimator $\widehat{\alpha}_n$ does not converges in general. Thus, there is no guarantee for the Hartigan consistency of the estimated α-tree.

4.2. *Error measurements.* Although Corollary 7 presents a negative result on estimating the α-tree, in this section we will show that the estimator $\widehat{\alpha}_n$ is still consistent under other error measurements. A key observation is that there is a 'good region' where we have the uniform convergence and the ordering of p_h is consistent to the ordering of ℓ.

Define $\mathbb{K}_s(h) = \mathbb{K}_s \backslash (\bigcup_{\ell<s} \mathbb{K}_\ell \oplus h)$ be the set $\mathbb{K}_s(h)$ that are away from lower dimensional support. Define further $\mathbb{K}(h) = \bigcup_{s \leq d} \mathbb{K}_s(h)$, which is the union of each $\mathbb{K}_s(h)$. Later we will show that the set $\mathbb{K}(h)$ is the 'good region'.

In Lemma 6, the quantity

$$m(x) = \min\{\ell \geq \tau(x) : x \in \overline{K}_\ell\} - \tau(x)$$

plays a key role in determining the rate of smoothing bias. Only when $m(x) = 1$ we have a slower rate for the bias. Thus, to obtain a uniform rate on the bias, we introduce the quantity

$$m_{\min} = \inf_{x \in \mathbb{K}, m(x) > 0} m(x). \tag{9}$$

If $m(x) = 0$ for all $x \in \mathbb{K}$, we define $m_{\min} = 2$. We define the following quantities:

$$\begin{aligned}
\delta_{n,h,s} &= O(h^{2 \wedge m_{\min}}) + O_P\left(\sqrt{\frac{\log n}{nh^s}}\right), \\
\delta^{(1)}_{n,h,s} &= O(h^{2 \wedge m_{\min}}) + O_P\left(\sqrt{\frac{\log n}{nh^{s+2}}}\right) \\
\delta^{(2)}_{n,h,s} &= O(h^{2 \wedge m_{\min}}) + O_P\left(\sqrt{\frac{\log n}{nh^{s+4}}}\right).
\end{aligned} \tag{10}$$

Later we will see that these quantities act as the density estimation rate, the gradient estimation rate, and the Hessian estimation rate.

THEOREM 8 (Consistency of the KDE under singular measures). *Assume (S, P2, K1–2). Let $x \in \mathbb{K}_s$ and $m(x)$ be defined in equation (9). Let $C_\ell^\dagger$ be the constants in Lemma 6. Let $\delta_{n,h,s}, \delta_{n,h,s}^{(1)}, \delta_{n,h,s}^{(2)}$ be defined in equation (10). Then when $h \to 0, \frac{nh^{d+4}}{\log n} \to \infty$,*

$$\sup_{x \in \mathbb{K}_s(h)} \|C_s^\dagger h^{s-d} \widehat{p}_n(x) - \rho(x)\| = \delta_{n,h,s}$$
$$\sup_{x \in \mathbb{K}_s(h)} \|C_s^\dagger h^{s-d} \nabla_{T_s(x)} \widehat{p}_n(x) - \nabla_{T_s(x)} \rho(x)\|_{\max} = \delta_{n,h,s}^{(1)}$$
$$\sup_{x \in \mathbb{K}_s(h)} \|C_s^\dagger h^{s-d} \nabla_{T_s(x)} \nabla_{T_s(x)} \widehat{p}_n(x) - \nabla_{T_s(x)} \nabla_{T_s(x)} \rho(x)\|_{\max} = \delta_{n,h,s}^{(2)},$$

where $\nabla_{T_s(x)}$ is taking gradient with respect to the tangent space of $\mathbb{K}_s$ at x.

Theorem 8 shows that after rescaling, the KDE is uniformly consistent within the good region $\mathbb{K}_s(h)$ for the density, gradient, and Hessian estimation. A more interesting result is that, after rescaling, the error rate is the same as the usual L_∞ error rate in the s-dimensional case with a modified bias term (bias is affected by the higher dimensional support).

REMARK 5. (Non-convergence of the integrated distance of the KDE) One may wonder if the scaled KDE ($C_{\tau(x)}^\dagger h^{\tau(x)-d} \cdot \widehat{p}_n(x)$) converges to the generalized density $\rho(x)$ under the integrated distance. In general, the answer is false:

$$\int \|C_{\tau(x)}^\dagger h^{\tau(x)-d} \cdot \widehat{p}_n(x) - \rho(x)\| dx = O_P(1).$$

To see this, consider a point $x \in \mathbb{K}_s$ and let $\mathbb{K}_\ell$ be a higher order support ($\ell > s$) with $x \in \overline{\mathbb{K}}_\ell$. Then the region $B(x,h) \cap \mathbb{K}_\ell$ has ℓ-dimensional volume at rate $O(h^{\ell-s})$. For any point $y \in B(x,h) \cap \mathbb{K}_\ell$, $\rho(y) = \ell$ but the KDE $\widehat{p}_n(y)$ is at rate $O_P(h^{s-d})$. Thus, the difference between scaled KDE and the generalized density

$$C_\ell^\dagger h^{d-\ell} \cdot \widehat{p}_n(y) - \rho(y) = O_P(h^{s-\ell}).$$

Such y has ℓ-dimensional volume at rate $O(h^{\ell-s})$ so the integrated error is at rate $O_P(h^{s-\ell}) \times O(h^{\ell-s}) = O_P(1)$, which does not converge.

Based on Theorem 8, we derive a nearly uniform convergence rate of $\widehat{\alpha}_n$.

THEOREM 9 (Nearly uniformly consistency of α-trees). *Assume (S, P2, K1–2). Let $\mathcal{C}_s$ be the collection of generalized critical points of $\mathbb{K}_s$. Let $\delta_{n,h,s}$ be defined in equation (10) and $r_{n,h,s}$ be a quantity such that $\frac{\delta_{n,h,s}}{r_{n,h,s}} = o_P(1)$. Then when $h \to 0, \frac{nh^{d+2}}{\log n} \to \infty$, unifmormly for every $x \in \mathbb{K}_s(h)$,*

$$\widehat{\alpha}_n(x) - \alpha(x) = \begin{cases} \delta_{n,h,s} & \text{if} \quad \inf_{c \in \mathcal{C}_s} |\rho(x) - \rho(c)| > r_{n,h,s}, \\ (\delta_{n,h,s})^{\frac{s}{s+1}} & \text{otherwise.} \end{cases}$$

In Theorem 9, the convergence rate haves similarly to the rate in Theorem 8: for a given point x when the $\alpha(x)$ is away from α value of a generalized critical point (a critical α level); when the $\alpha(x)$ is close to a critical α level, we have a slower rate of convergence. The quantity $r_{n,h,s}$ behaves like the quantity a_n in Theorem 4 which is introduced to guarantee the uniform convergence. To derive the consistency of $\widehat{\alpha}_n$ under the integrated error (and the probability error), we will choose $r_{n,h,s} = \frac{\delta_{n,h,s}}{\log n}$, which leads to the following theorem.

THEOREM 10 (Consistency of α-trees). *Assume (S, P2 , K1–2). Let $m(x)$ be the quantity in equation (9). Then*

$$\|\widehat{\alpha}_n - \alpha\|_P = \delta_{n,h,s},$$
$$\|\widehat{\alpha}_n - \alpha\|_\mu = \delta_{n,h,s}.$$

Namely, Theorem 9 shows that the quantity $\widehat{\alpha}_n(x)$ is stable for majority points–this implies that the ordering of points in $\mathbb{K}$ from $\widehat{p}_n$ is consistent to the ordering from $\ell(x)$ in general.

REMARK 6. In Theorem 10, $\widehat{\alpha}_n$ converges under the integrated distance but in Remark 5, the scaled KDE fails to converge. Both the scaled KDE and $\widehat{\alpha}_n$ rescale the original KDE to adjust to the singular measure. The rescaling in $\widehat{\alpha}_n$ is with respect to the probability, which is bounded by 1 so the bad regions does not contribute too much to the integrated error. On the other hand, the scaled KDE is unbounded so the contribution from the bad regions is huge, causing the failure of convergence in the integrated error.

5. Singular Measures: Critical Points and Topology. Recalled from Section 2.1 that the topology of an α-tree T_α is determined by its edge set $E(T_\alpha)$ and the relation among edges $\mathbb{C} \in E(T_\alpha)$. And the set $\mathcal{A}_\alpha$ contains the levels where the upper level set $\mathbb{A}_a = \{x : \alpha(x) \geq a\}$ changes its shape. For a nonsingular measure, it is well-known that $\mathcal{A}_\alpha$ corresponds to the density value of some critical points. For a singular measure, this is not true even when $\rho(x)$ is a generalize Morse function.

Consider the example in Figure 3, the solid box in the left panel indicates a new type of 'critical points', where a merge between elements in different edge sets occurs (change of the topology of level sets occurs); by the definition of $\mathcal{A}_\alpha$, this corresponds an element in $\mathcal{A}_\alpha$ but it is clearly not a generalized critical point. We called this type of critical points the *dimensional critical points (DCPs)*. In Figure 3, the dimension $d = 2$ and we have a $2D$ spherical distribution mixed with a $1D$ singular measure that distributed on the red curves (this red curve is $\mathbb{K}_1$). The bluish contours are density contours of the $2D$ spherical distribution; the crosses are locations of local modes; and the solid box is the location of a DCP. To see how the solid box changes the topology of level sets, we display two level sets in the middle and right panels. In the middle panel, the level is high and there are two connected components (the gray area and the solid curve) . In the right panel, we lower the level and now the two connected components are merged at the location of the solid box. Although the location of the solid box does not belong to $\mathcal{C}$, the collection of generalized critical points, this point does corresponds to mergings of connected components in the level sets. So this point corresponds to a element in $\mathcal{A}_a$.

Here is the formal definition of the DCP. Recalled that $\mathcal{C}$ is the collection of generalized critical points of $\rho(x)$ and recalled from equation (1) that $\mathcal{A}_\alpha$ is the collection of levels of $\alpha(x)$ such that the creation of a new connected component or a merging of connected components occurs. For simplicity, we denote $\mathcal{A} = \mathcal{A}_\alpha$. For $a \in \mathcal{A}$, define $\xi(a)$ to be the integer such that $\mathbb{K}_s \subset \mathbb{A}_a$ for all $s \leq \xi(a)$ and $\mathbb{K}_{\xi(a)+1} \not\subset \mathbb{A}_a$.

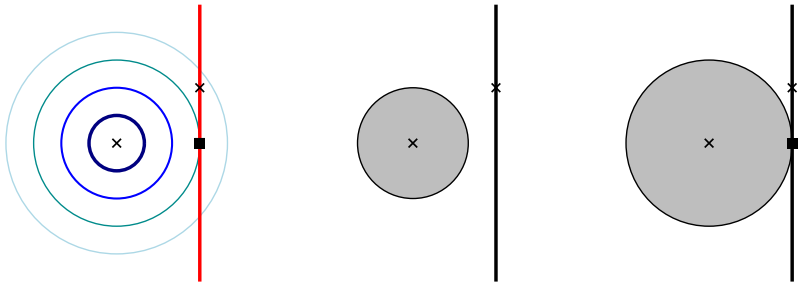

FIG 3. *An example for the dimensional critical points (DCPs). This is a $d = 2$ cases; there is a 2-dimensional spherical distribution mixed with a 1-dimensional distribution on a line segment. Left: the blue contours are density contours of the 2D spherical distribution and the red line segment is, $\mathbb{K}_1$, the support of the 1-dimensional singular distribution. The two crosses are the density maxima at the 2D distribution and the 1D singular distribution. The black square indicates a DCP. To see how DCP merges two connected compoments, we consider the middle and the right panel, which are level sets of $\alpha(x)$ at two different levels. Middle: the level set $\mathbb{A}_a$ where the level a is high; we can see that there are two connected components (left gray-black disk and the right line segment). Right: we move down the level a little bit; now the two connected components merged so there is only one connected component. The merging point is the square point, which is defined as a DCP.*

DEFINITION 3. *For $a \in \mathcal{A}$, we say x is a dimensional critical point (DCP) if the following holds*

(1) $x \in \mathbb{K}_\ell$ for some $\ell \leq \xi(a)$.
(2) there exist an edge $\mathbb{C} \in E(T_\alpha)$ such that
  (i) $x \notin C$ for all $C \in \mathbb{C}$,
  (ii) $d(x, C_\epsilon) \to 0$ when $\epsilon \to 0$, where $C_\epsilon = \mathbb{C} \cap T_\alpha(a + \epsilon)$.

Note that x may not exist, in such a case, there is not DCP for the level a.

The first requirement is to ensure that x is on a lower dimensional support ($\mathbb{K}_\ell : \ell < \xi(a)$). The second requirement shows that the DCP x is not in the same edge $\mathbb{C}$ but the distance to the elements (connected components) of the edge is shrinking to 0. By the definition of $\alpha(x)$, the first requirement implies that x is contained in $\mathbb{A}_\alpha(a + \epsilon)$ for sufficiently small ϵ. Therefore, we can find $\mathbb{C}' \in E(T_\alpha)$ such that every element $C \in \mathbb{C}'$ contains x. Because x is (i) in the elements of edge $\mathbb{C}'$, (ii) not in any element of edge $\mathbb{C}$, and (iii) the distance from x to the element of $\mathbb{C}$ converges to 0 when the level decreases to the level a, this indicates that x is a merging point of edges $\mathbb{C}'$ and $\mathbb{C}$ and the level a is their merging level.

Let $\mathcal{C}^D$ be the collection of DCPs. For a point $c \in \mathcal{C}^D$, we denote $\alpha^\dagger(c)$ as the level of α function corresponds to the DCP at c.

REMARK 7 (Relation to the usual critical points). The definition of DCPs is very similar to those saddle points (or local minima) contributing to the merging of level sets. A saddle point (or local minimum) that contributes to a merging of level sets can be defined as a point x with the following properties

(1) $x \in \mathbb{K}_{\xi(a)+1}$.
(2) there exist two different edges $\mathbb{C}_1, \mathbb{C}_2 \in E(T_\alpha)$ such that
  (i) $x \notin C_1, x \notin C_2$ for all $C_1 \in \mathbb{C}_1$ and $C_2 \in \mathbb{C}_2$,
  (ii) $d(x, C_{1,\epsilon}) \to 0$ when $\epsilon \to 0$, where $C_{1,\epsilon} = \mathbb{C}_1 \cap T_\alpha(a + \epsilon)$,

(iii) $d(x, C_{2,\epsilon}) \to 0$ when $\epsilon \to 0$, where $C_{2,\epsilon} = \mathbb{C}_2 \cap T_\alpha(a+\epsilon)$.

It is easy to see that for a Morse function, such a point x must be a saddle point or a local minimum. Comparing this definition to the definition of DCPs, the main difference is the support where x lives in– if x lives in a lower dimensional support $\mathbb{K}_\ell$, $\ell \le \xi(a)$, then it is a DCP; if x lives in the support $\mathbb{K}_{\xi(a)+1}$, then it is a saddle point (or a local minimum).

REMARK 8. Note that a DCP might be located at the same position as a critical point. Consider the example in Figure 1 or Example 1. In this case, there is a DCP located at $x = 2$, which coincides with a local mode. However, the DCP and the local mode corresponds to different elements in $\mathcal{A}_\alpha = \{1, 0.7, 0.0319\}$; the local mode corresponds to the level 1 and the DCP corresponds to 0.0319 because it represents the merging of the two connected components. Note that the number $0.0319 = 0.7 \times (\Phi_0(-2) + 1 - \Phi_0(2))$, where $\Phi_0(x)$ is the cumulative distribution of a standard normal.

To analyze the properties of DCPs and their estimators, we consider the following assumptions.

Assumptions.

(A) The elements in the collection $\mathcal{A}$ are distinct and each element corresponds exactly to one critical point or one DCP (but not both); and all DCPs are distinct.
(B) For every $x \in \partial\mathbb{K}_s$ ($s > 0$) and $r > 0$, there is $y \in B(x, r) \cap K_s$ such that $\rho(y) > \rho(x)$.
(C) There exists $\eta_0 > 0$ such that

$$\inf_{c \in \mathcal{C}_s} d(c, \mathbb{K}_\ell) \ge \eta_0,$$

for all $\ell < s$ and $s = 1, 2, \cdots, d$.

The assumption (A) is to ensure that there will be no multiple topological changes occurring at the same level so each level corresponds to only a merging or a creation. The assumption (B) is to guarantee that there will be no creation of a new connected component at the boundary of a lower dimensional manifold. Thus, any creation of a new connected component of the level set $\mathbb{A}_a$ occurs only at a (generalized) local mode. The assumption (C) is to regularize (generalized) critical points so that they are away from lower dimensional supports. This implies that when h is sufficiently small, all critical points will be in the 'good region' $\mathbb{K}(h)$.

LEMMA 11 (Properties of DCPs). *Assume (S, P2, B). The DCPs have the following properties*

- *If a new connected component of $\mathbb{A}_a$ is created at $a \in \mathcal{A}$, then there is a local mode c of ρ or an element in $\mathbb{K}_0$ such that $a = \alpha(c)$. Namely, DCPs only merge connected components.*
- *For any value $a \in \mathcal{A}$, either $a = \alpha(c)$ for some $c \in \mathcal{C}$ or there is a DCP associated to a.*

Lemma 11 provides two basic properties of DCPs. First, DCPs only merge connected component. And secondly, when the topology of connected components of α-level sets changes (when we decrease the level), there must be either a critical point or a DCP that causes this. Namely, as long as we control the stability of generalized critical points and DCPs, we control the topology of an α-tree. Thus, in what follows we will study the stability of generalized critical points and DCPs.

LEMMA 12 (Stablility of generalized critical points). *Assume (S, P2, K1–2, C). Let $c \in \mathbb{K}_s$ be a generalized critical point with $n(s)$ negative eigenvalues of its generalized Hessian matrix. Let $\widehat{\mathcal{C}}$ be the*

collection of critical points of $\widehat{p}_n$. *Let* $m(c) = \min\{\ell \geq s : c \in \overline{K}_\ell\} - s$. *Then when* $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$ *and* $m(c) > 1$, *there exists a point* $\widehat{c} \in \widehat{\mathcal{C}}$ *such that*

$$\|\widehat{c} - c\| = O\left(h^{2 \wedge m(c)}\right) + O_P\left(\sqrt{\frac{1}{nh^{s+2}}}\right)$$
$$\|\widehat{\alpha}_n(\widehat{c}) - \alpha(c)\| = (\delta_{n,h,s})^{\frac{s}{s+1}}$$

and the estimated Hessian matrix at $\widehat{c}$ *has* $n(c) + d - s$ *negative eigenvalues. The quantity* $\delta_{n,h,s}$ *is defined in equation* (10).

Lemma 12 is a generalization of the stability Theorem of critical points given in Lemma 16 of Chazal et al. (2014). The rate is similar to the ones in Theorem 8 and 9; we have a modified bias due to the lower dimensional support and adaptive stochastic error rate to the dimension of the support where c resides in. Due to the adaptive rate, generalized critical points are more stable when they are on a lower dimensional support.

LEMMA 13 (Properties of the estimated critical points). *Assume (S, P2, K1–2, A, B). Assume there are* k *DCPs. Let* $\widehat{\mathcal{C}}$ *be the critical points of* $\widehat{p}_n$. *Define* $\widehat{\mathcal{G}} \subset \widehat{\mathcal{C}}$ *be the collection of estimated critical points corresponding to the generalized critical points. Let* $\widehat{\mathcal{D}} = \widehat{\mathcal{C}} \backslash \widehat{\mathcal{G}}$ *be the remaining estimated critical points. Then when* $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$,

- $\widehat{\mathcal{D}} \subset \mathbb{K}^C(h)$,
- $\left|\widehat{\mathcal{D}}\right| \geq k$, *where* $|A|$ *for a set* A *is the cardinality,*
- $\widehat{\mathcal{D}}$ *contains no local mode of* $\widehat{p}_n$.

Lemma 13 provides several useful properties about the estimated critical points (critical points of the $\widehat{p}_n$). First, estimated critical points are all in the bad region $\mathbb{K}^C(h)$ except those converging to generalized critical points. Second, the number of estimated critical points will (asymptotically) not be less than the total number of DCPs. Third, all estimated local modes are estimators of generalized critical points.

LEMMA 14 (Stability of critical tree-levels from DCPs). *Assume (S, P2, K1–2, A, B). Let* c *be a DCP and* $\alpha^\dagger(c) \in \mathcal{A}$ *be the associated level. Let* $\widehat{\mathcal{D}}$ *be defined in Lemma 13. Then when* $h \to 0$, $\frac{nh^{d+2}}{\log n} \to \infty$, *there exists a point* $\widehat{c} \in \widehat{\mathcal{D}}$ *such that*

$$\|\widehat{\alpha}_n(\widehat{c}) - \alpha^\dagger(c)\| = \delta_{n,h,\xi(\alpha_0(c))+1},$$

where $\delta_{n,h,s}$ *is defined in* (10). *Moreover, the* $\widehat{\mathbb{A}}_{\widehat{\alpha}_n(\widehat{c})+\epsilon}$ *and* $\widehat{\mathbb{A}}_{\widehat{\alpha}_n(\widehat{c})}$ *are not topological equivalent.*

Lemma 14 illustrates the stability of critical levels from DCPs– for every DCP, there will be an estimated critical point that corresponds to this DCP and this estimated critical point also represents a merging of estimated level sets.

Note that in Lemma 12, we derive the rate of convergence of the estimated (generalized) critical points versus the population critical points but here in Lemma 14, we only derive the rate for the critical tree-levels. The reason is that the critical points are solution to a certain function (gradient equals to 0) so we can perform a Talyor expansion to obtain the convergence rate. But for the DCPs, they are not solutions to some functions so it is unclear how to derive the convergence rate for the locations.

EXAMPLE 2 (An example of a DCP and its estimator). *Consider again the example in Figure 1 and Example 1; we have a singular distribution mixed with a point mass at $x = 2$ with probability 0.3 and a standard normal with probability 0.7. In this case, as indicated, a DCP is located at $x = 2$ with level 0.0319 (see Remark 8). In every panel of the top row of Figure 1, there is a local minimum located in the region $x \in [1.5, 2]$. Moreover, when we increases the sample size (from left to right), this local minimum is moving toward $x = 2$. This local minimum is an estimated critical point $\widehat{c} \in \widehat{\mathcal{D}}$ described in Lemma 14 whose estimated α-level is approaching the α-level of the DCP at $x = 2$.*

THEOREM 15 (Topological error of α-trees). *Assume (S, P2, K1–2, A, B, C). Then when $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$,*

$$P\left(T_{\widehat{\alpha}_n} \overset{T}{\approx} T_\alpha\right) \geq 1 - c_0 \cdot e^{-c_1 \cdot nh^{d+4}},$$

for some $c_0, c_1 > 0$.

Theorem 15 quantifies the topological error of the estimated α-tree under singular measures and the error rate is the same as 'nonsingular' measures (Theorem 5). Compared to Corollary 5, the topological error bound is very similar; both are exponential concentration bound with a factor of nh^{d+4}, which is the Hessian estimation error rates. The two concentration bounds are similar because as is shown in Theorem 8, the main difference between singular and nonsingular measures is in the bias part, which will not contribute to the concentration inequality as long as $h \to 0$. Note that the Hessian error rate is because we need to make sure the signs of eigenvalues of Hessian matrices around critical points remain unchanged.

REMARK 9. Theorem 15 also implies that, under singular measures, the cluster tree of the KDE $\widehat{p}_n$ (estimated λ-tree) converges topologically to a population cluster tree defined by the function $\alpha(x)$. To see this, recalled that by Lemma 1, $T_{\widehat{p}_n} \overset{T}{\approx} T_{\widehat{\alpha}_n}$. This together with Theorem 15 implies

$$P\left(T_{\widehat{p}_n} \overset{T}{\approx} T_\alpha\right) \geq 1 - c_0 \cdot e^{-c_1 \cdot nh^{d+4}} \to 1$$

under suitable choice of h. This shows that even when the population distribution is singular, the estimated λ-tree still converges topologically to the population α-tree.

6. Discussion. In this paper, we study how the α-tree behaves under both singular and nonsingular measures. In the nonsingular case, due to the slow rate of level set estimating around saddle points, the error rate under the L_∞ metric is slower than other metrics. But other error rates are the same as estimating the λ-tree.

When the distribution is singular, we obtain many fruitful results for both the KDE and the estimated α-tree. In terms of the KDE, we prove that

1. the KDE is a pointwisely consistent estimator after rescaling;
2. the KDE is a uniformly consistent estimator after rescaling for the majority part of the support;
3. the cluster tree from the KDE (estimated λ-tree) converges topologically to a population cluster tree defined by α.

For the estimator $\widehat{\alpha}_n(x)$ and the estimated α-tree, we show that

1. $\widehat{\alpha}_n$ is a pointwisely consistent estimator of α;
2. $\widehat{\alpha}_n$ is a uniformly consistent estimator for the majority part of the support;
3. $\widehat{\alpha}_n$ is a consistent estimator of α under the integrated distance and probability distance;
4. the estimated α-tree converges topologically to the population α-tree.

Moreover, we observe a new type of critical points–the DCPs–that also contribute to the merging of level sets for singular measures. We study the properties of DCPs and show that estimated critical points from the KDE approximate these DCPs.

Finally, we point out some possible future directions.

- **Persistence homology.** The cluster tree is highly related to the persistent homology of level sets (Fasy et al., 2014; Bobrowski et al., 2014). In the persistent homology, a common metric for evaluating the quality of an estimator is the bottleneck distance (Cohen-Steiner et al., 2007; Edelsbrunner and Morozov, 2012). Because the bottleneck distance is bounded by the L_∞ metric (Cohen-Steiner et al., 2007; Edelsbrunner and Morozov, 2012), many bounds on the bottleneck distance is derived via bounding the L_∞ metric (Fasy et al., 2014; Bobrowski et al., 2014). However, for α-trees under singular measures, the L_∞ does not converge (Corollary 7) but we do have topological consistency (Theorem 15), which implies the convergence in the bottleneck distance. This provides an example where we have consistency under the bottleneck distance and inconsistency of the L_∞ metric. How this phenomenon affects the persistence homology is unclear and we leave the study along this line as a future work.
- **Higher order homology groups.** Our definition of DCPs is for connected components, which are the zeroth order homology groups (Cohen-Steiner et al., 2007; Fasy et al., 2014; Bubenik, 2015). For analyzing cluster trees, zeroth order homology groups are sufficient. However, critical points also contribute to the creation and elimination of higher order homology groups such as loops and voids which are not covered in this paper. Thus, a future direction is to study if the KDE is also consistent in recovering higher order homology groups under singular measures.
- **Minimax theory.** In the seminal work of Chaudhuri and Dasgupta (2010); Chaudhuri et al. (2014), they derived the minimax theory for estimating the λ-tree under nonsingular measures and they prove that the the k-nearest neighbor estimator is minimax. When the distribution is nonsingular, the α-tree and λ-tree are very similar so we expect the minimax theory to be the same. For singular measures, however, it is unclear how to derive the minimax theory so we plan to investigate this in the future.

APPENDIX A: PROOFS

Proof of Theorem 3. Let c be a critical point with $\lambda = p(c)$. To prove this theorem, we partition the difference $\widehat{L}_\lambda \triangle L_\lambda$ into three regions:

$$A_n = (\widehat{L}_\lambda \triangle L_\lambda) \bigcap B^C(c, R_0),$$
$$B_n = (\widehat{L}_\lambda \triangle L_\lambda) \bigcap (B(c, R_0) \backslash B(c, R_n)),$$
$$C_n = (\widehat{L}_\lambda \triangle L_\lambda) \bigcap B(c, R_n),$$

where R_0 is some small but fixed constant and $R_n < R_0$ is a value that shrinks to 0 when $n \to \infty, h \to 0$. Later we will provide an explicit expression for R_n. We only consider the case where c is a saddle point because when c is a local mode or local minimum, the set B_n will be an empty set because C_n will cover the fluctuations around c.

Overview of the proof. Here is an overview for the proof. For region A_n, it is away from the saddle point so there will be a minimal gradient bound on the boundary of L_λ. Thus, we can apply the existing result from the literatures to bound the rate. Later we will show that this has the usual rate of convergence $O_P(\|\widehat{p}_n - p\|_\mu)$. For region B_n, the gradient lower bound on the boundary of L_λ is decreasing at rate $O(R_n)$ because we are moving close to the saddle point. However, as long as R_n does not shrink too fast, we can still approximate the difference between $\widehat{L}_\lambda$ and L_λ by existing method and the rate will be $O_P(\frac{1}{R_n}\|\widehat{p}_n - p\|_\mu)$. The final part C_n is just to control the variation of saddle point; we need to pick R_n large enough so that the estimated saddle point $\widehat{c}$ will still be inside $B(c, R_n)$. And this part contributes to the error less than $O(R_n^d)$.

Thus, to sum up, the total error rate is bounded by

$$\mu(\widehat{L}_\lambda \triangle L_\lambda) = O_P(\|\widehat{p}_n - p\|_\mu) + O_P\left(\frac{1}{R_n}\|\widehat{p}_n - p\|_\mu\right) + O(R_n^d),$$

so the optimal choice is $R_n = O\left(\|\widehat{p}_n - p\|_\mu^{\frac{1}{d+1}}\right)$, which yields the desired result. For simplicity, we use notation $D_\lambda = \partial L_\lambda$ and $\widehat{D}_n = \partial \widehat{L}_\lambda$.

Part 1: analysis for A_n. Recalled that $g(x) = \nabla p(x)$. For L_λ inside $B^C(c, R_0)$, there exists a constant $g_0 > 0$ such that

$$\inf_{x \in D_\lambda} \|g(x)\| \geq g_0 > 0.$$

Namely, every point on the boundary of L_λ has nonzero gradient. This is true when R_0 is sufficiently small (the only case the gradient can be close to 0 is when $x \in D_\lambda$ is close to the saddle point c because we assume all critical values are distinct).

The difference inside A_n can be approximated by integrating the local difference $d(x, \widehat{D}_\lambda)$ over the set $D_\lambda \bigcap B^C(c, R_0)$. By Lemma 2 in Chen et al. (2015a),

$$d(x, \widehat{D}_\lambda) = \frac{1}{\|g(x)\|}\|\widehat{p}_n(x) - p(x)\|(1 + O_P(\|\widehat{p}_n - p\|_\infty)).$$

Because the gradient $\|g(x)\|$ is lower bounded for D_λ, we conclude that

$$\mu(A_n) \leq O_P\left(\int_{D_\lambda \cap B^C(c,R_0)} \frac{1}{\|g(x)\|}\|\widehat{p}_n(x) - p(x)\| dx\right) = O_P\left(\|\widehat{p}_n - p\|_\mu\right).$$

Part 2: analysis for B_n. An important observation for the region $B(c, R_n)$ is that the gradient lower bound behaves at rate $O(R_n)$. This is because $g(c) = 0$ (definition of saddle point) and the eigenvalues of Hessian matrix are bounded away from 0. Thus, the gradient has to increase at least linearly when we are moving away from a saddle point. Therefore, there exists a constant $\eta_0 > 0$ such that

$$\inf_{x \in B(c,R_0)\setminus B(c,R_n)} \|g(x)\| \geq \eta_0 R_n.$$

Note that a simple choice is η_0 being the half of the smallest eigenvalue of $H(c)$.

Now again, we will apply Lemma 2 in Chen et al. (2015a) to approximate the local difference

$$d(x, \widehat{D}_\lambda) = \frac{1}{\|g(x)\|}\|\widehat{p}_n(x) - p(x)\|(1 + O_P(\|\widehat{p}_n - p\|_\infty)).$$

Let $D'_\lambda = D_\lambda \bigcap (B(c, R_0) \backslash B(c, R_n))$. Because we have a lower bound on the gradient, the quantity

$$\begin{aligned}\int_{D'_\lambda} d(x, \widehat{D}_\lambda) dx &= \int_{D'_\lambda} \frac{1}{\|g(x)\|} \|\widehat{p}_n(x) - p(x)\| dx \\ &\leq O_P\left(\frac{1}{R_n} \int_{D'_\lambda} \|\widehat{p}_n(x) - p(x)\| dx\right) \\ &= O_P\left(\frac{1}{R_n} \|\widehat{p}_n - p\|_\mu\right).\end{aligned}$$

Note that to ensure $\int_{D'_\lambda} d(x, \widehat{D}_\lambda) dx = \mu(B_n)$, we need the normal compatibility property between $\widehat{D}'_\lambda$ and D'_λ (see Section 2.3 in Chen et al. 2015a or Chazal et al. (2007)). And this is guaranteed if their Hausdorff distance is less than the reach of both D'_λ and $\widehat{D}'_\lambda$. The reach of D'_λ decreases at rate $O(R_n)$ by Lemma 1 in Chen et al. 2015a and so is the reach of $\widehat{D}'_\lambda$. Thus, as long as we have

$$\|\widehat{p}_n - p\|_\infty = o(R_n), \tag{11}$$

we have $\int_{D'_\lambda} d(x, \widehat{D}_\lambda) dx = \mu(B_n)$. Later we will show that our choice of R_n is $\|\widehat{p}_n - p\|_\mu^{\frac{1}{d+1}}$ so we do have this result.

Part 3: analysis for C_n. It is well-known that the distance between estimated saddle point and true saddle point follows the rate

$$\|\widehat{c} - c\| = O\left(\|\widehat{g}_n(c) - g(c)\|\right).$$

See, e.g., Theorem 1 in Chen et al. (2016a) and Lemma 16 in Chazal et al. (2014). Thus, all we need is to choose R_n such that

$$\|\widehat{c} - c\| = O\left(\|\widehat{g}_n(c) - g(c)\|\right) = o(R_n). \tag{12}$$

And it is easy to see that

$$\mu(C_n) \leq \mu(B(c, R_n)) = O(R_n^d).$$

Thus, putting altogether, we have

$$\mu(\widehat{L}_\lambda \triangle L_\lambda) = \mu(A_n) + \mu(B_n) + \mu(C_n) = O_P(\|\widehat{p}_n - p\|_\mu) + O_P\left(\frac{1}{R_n} \|\widehat{p}_n - p\|_\mu\right) + O(R_n^d),$$

And the optimal choice of R_n is $R_n = \|\widehat{p}_n - p\|_\mu^{\frac{1}{d+1}}$, which leads to the rate

$$\mu(\widehat{L}_\lambda \triangle L_\lambda) = O_P\left(\frac{1}{R_n} \|\widehat{p}_n - p\|_\mu^{\frac{d}{d+1}}\right).$$

Moreover, this choice of R_n satisfies both equation (11) and (12).

□

PROOF OF THEOREM 4. Recalled that $\widehat{\alpha}_n(x) = \widehat{P}_n\left(\widehat{L}_{\widehat{p}_n(x)}\right)$ and $\alpha(x) = P\left(L_{p(x)}\right)$. The main idea for the proof is the following decomposition:

$$\widehat{\alpha}_n(x) - \alpha(x) = \underbrace{\widehat{P}_n\left(\widehat{L}_{\widehat{p}_n(x)}\right) - P\left(\widehat{L}_{\widehat{p}_n(x)}\right)}_{\text{(A)}} + \underbrace{P\left(\widehat{L}_{\widehat{p}_n(x)}\right) - P\left(\widehat{L}_{p(x)}\right)}_{\text{(B)}} + \underbrace{P\left(\widehat{L}_{p(x)}\right) - P\left(L_{p(x)}\right)}_{\text{(C)}}.$$

Part (A). This is just the difference between empirical measure and probability measure for a given set. Thus, this term has rate $O_P\left(\sqrt{\frac{1}{n}}\right)$.

Part (B). Because

$$|P\left(\widehat{L}_{\widehat{p}_n(x)}\right) - P\left(\widehat{L}_{p(x)}\right)| \leq P\left(\widehat{L}_{\widehat{p}_n(x)} \triangle \widehat{L}_{p(x)}\right), \tag{13}$$

we first investigate a more general bound on the right hand side. Let $\epsilon > 0$ be a small number. For the estimated level set $\widehat{L}_\lambda$ and $\widehat{L}_{\lambda+\epsilon}$, we want to control the quantity $P(\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon})$ when $\epsilon \to 0$. Note that to obtain the bound in equation (13), we pick $\lambda = p(x)$ and $\epsilon = \widehat{p}_n(x) - p(x)$.

An interesting result is that $P(\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon})$ differs if λ is a critical value (i.e. density value of a critical point) or not. When λ is not a critical value, the gradient $g(x)$ has a non-zero lower bound on $\partial\widehat{L}_\lambda$ and we have the local approximation

$$d(x, \widehat{L}_{\lambda+\epsilon}) = \frac{\epsilon}{\|g(x)\|} + o(\epsilon)$$

for each $x \in \partial\widehat{L}_\lambda$. Thus, this implies

$$P(\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon}) = O(\epsilon). \tag{14}$$

When λ coincides with a critical value, then we need to split the region $\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon}$ into three subregions:

$$\begin{aligned} A_n &= (\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon}) \bigcap B^C(c, R_0), \\ B_n &= (\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon}) \bigcap (B(c, R_0) \backslash B(c, R_\epsilon)), \\ C_n &= (\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon}) \bigcap B(c, R_\epsilon), \end{aligned}$$

where R_0 is a non-zero constant and R_ϵ is a constant converging to 0 when $\epsilon \to 0$.

The idea is similar to the proof of Theorem 3. In A_n we apply the same local approximation for the case where λ is not a critical value. Thus, $P(A_n) = O_P(\epsilon)$. For B_n, we still have the local approximation but now the rate is slowed down by the gradient. Using the same calculation as the Part 2 of the proof of Theorem 3, we obtain the rate

$$P(B_n) = O_P\left(\frac{\epsilon}{R_\epsilon}\right).$$

The final part C_n is to control the area around critical points. This part contributes $P(C_n) = O_P(\epsilon^d)$. Thus, putting altogether, we have

$$\begin{aligned} P(\widehat{L}_\lambda \triangle \widehat{L}_{\lambda+\epsilon}) &= P(A_n) + P(B_n) + P(C_n) \\ &= O_P(\epsilon) + O_P\left(\frac{\epsilon}{R_\epsilon}\right) + O_P(R_\epsilon^d) \\ &= O_P\left(\epsilon^{\frac{d}{d+1}}\right) \end{aligned} \tag{15}$$

when we choose $R_n = O(\epsilon^{\frac{1}{d+1}})$.

Note that the area around critical point is at rate $\sqrt{\epsilon}$. This is because density at critical point behaves quadratically so difference in density by ϵ result in a difference in distance by $\sqrt{\epsilon}$. And the choice $R_n = O(\epsilon^{\frac{1}{d+1}})$ is obviously slower than $\sqrt{\epsilon}$, so equation (A) is valid.

Plugging-in equation (14) and into equation (13), we conclude that

- if the density at x, $p(x)$, is not a critical value of $\widehat{p}$,

$$P\left(\widehat{L}_{\widehat{p}_n(x)}\right) - P\left(\widehat{L}_{p(x)}\right) = O_P\left(\|\widehat{p}_n(x) - p(x)\|\right);$$

- if the density at x, $p(x)$, is a critical value of $\widehat{p}$,

$$P\left(\widehat{L}_{\widehat{p}_n(x)}\right) - P\left(\widehat{L}_{p(x)}\right) = O_P\left(\|\widehat{p}_n(x) - p(x)\|^{\frac{d}{d+1}}\right).$$

Finally, note that for a point x, the case where its density $p(x)$ might be a critical value of $\widehat{p}_n$ occurs only when $|p(x) - p(c)| \leq \|\widehat{p}_n - p\|_\infty$ for some critical point c of p. Recalled that a_n is a sequence of n such that $\|\widehat{p}_n - p\|_\infty = o(a_n)$. Then the above bound can be rewritten as

$$P\left(\widehat{L}_{\widehat{p}_n(x)}\right) = \begin{cases} O_P\left(\|\widehat{p}_n(x) - p(x)\|\right) & \text{, if } \quad |p(x) - p(c)| > a_n \text{ for all } c \in \mathcal{C}. \\ O_P\left(\|\widehat{p}_n(x) - p(x)\|^{\frac{d}{d+1}}\right) & \text{, otherwise.} \end{cases}$$

Note that this bound is uniformly for all x because the sequence a_n does not depend on x.

Part (C). This part is simply by applying Theorem 3, which shows

- if the density at x, $p(x)$, is not a critical value of p,

$$P\left(\widehat{L}_{p(x)}\right) - P\left(L_{p(x)}\right) = O_P\left(\|\widehat{p}_n - p\|_\mu\right);$$

- if the density at x, $p(x)$, is a critical value of p,

$$P\left(\widehat{L}_{p(x)}\right) - P\left(L_{p(x)}\right) = O_P\left(\|\widehat{p}_n - p\|_\mu^{\frac{d}{d+1}}\right).$$

Thus, putting everything together, we conclude that uniformly for all x,

$$\widehat{\alpha}_n(x) - \alpha(x) = \begin{cases} O_P\left(\|\widehat{p}_n - p\|_\mu\right) & \text{, if } \quad |p(x) - p(c)| > a_n \text{ for all } c \in \mathcal{C}, \\ O_P\left(\|\widehat{p}_n - p\|_\mu^{\frac{d}{d+1}}\right) & \text{, otherwise.} \end{cases}$$

Note that the pointwise rate $\|\widehat{p}_n(x) - p(x)\|$ and the integrated rate $\|\widehat{p}_n - p\|_\mu$ are at the same order so we use the integrated rate.

□

PROOF OF LEMMA 6. **Part 1: pointwise bias.** Without loss of generality, we assume $x \in \mathbb{K}_s$. We first consider the case $m(x) > 0$. In this case, there is higher dimensional support $\mathbb{K}_{s+m(x)}$ such that $x \in \overline{\mathbb{K}}_{s+m(x)}$. Thus, for any $r > 0$, the ball $B(x,r) \bigcap \mathbb{K}_{s+m(x)} \neq \phi$.

Let $\mu_s(x)$ be the s-dimensional Lebesgue measure. Because the kernel function K is supported on $[0,1]$,

$$p_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right) = \int \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y) = \sum_{\ell=0}^{d} \int_{\mathbb{K}_\ell} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y).$$

When h is sufficiently small, $B(x,h) \bigcap \mathbb{K}_\ell = \phi$ for any $\ell < s$ so the above expression can be rewritten as

$$p_h(x) = \sum_{\ell=s}^{d} \int_{\mathbb{K}_\ell} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y).$$

Now by the definition of $m(x)$, $B(x,r) \bigcap \mathbb{K}_\ell = \phi$ for every $\ell > s$ and $\ell < s + m(x)$. Thus, we can again rewrite $p_h(x)$ as

$$\begin{aligned} p_h(x) &= \int_{\mathbb{K}_s} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y) \\ &\quad + \sum_{\ell \geq s+m(x)}^{d} \int_{\mathbb{K}_\ell} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y) \\ &= \underbrace{\int_{\mathbb{K}_s \bigcap B(x,h)} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y)}_{\text{(I)}} \\ &\quad + \underbrace{\sum_{\ell \geq s+m(x)}^{d} \int_{\mathbb{K}_\ell \bigcap B(x,h)} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y)}_{\text{(II)}}. \end{aligned} \tag{16}$$

Using the generalized density on the s-dimensional support, the first term equals to

$$\begin{aligned} (I) &= \int_{\mathbb{K}_s \bigcap B(x,h)} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y) \\ &= \int_{\mathbb{K}_s \bigcap B(x,h)} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) \rho(y) dy \end{aligned} \tag{17}$$

Because we assume $\rho(x)$ is at least three-times bounded differentiable, for any $y\mathbb{K}_s$ that is close to x,

$$\rho(y) = \rho(x) + (y-x)^T g_s(x) + (y-x)^T H_s(x)(y-x) + o(\|x-y\|^2).$$

Plugging this into equation (17) and use the fact that $K(\|x\|)$ is symmetric and $\mathbb{K}_s$ is an s-dimensional

manifold,

$$\begin{aligned}(I) &= \int_{\mathbb{K}_s \bigcap B(x,h)} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) \rho(y) dy \\ &= \int_{x+uh \in \mathbb{K}_s \bigcap B(x,h)} \frac{1}{h^d} K\left(\|u\|\right) \rho(x) du \cdot h^s \\ &\quad + \int_{x+uh \in \mathbb{K}_s \bigcap B(x,h)} \frac{1}{h^d} K\left(\|u\|\right) u^T H_s(x) u du \cdot h^{s+2} + o\left(\frac{h^{s+2}}{h^d}\right) \\ &= \rho(x) \cdot \int_{x+uh \in \mathbb{K}_s \bigcap B(x,h)} \frac{1}{h^d} K\left(\|u\|\right) du \cdot h^s + O\left(\frac{h^{s+2}}{h^d}\right).\end{aligned}$$

Because $\mathbb{K}_s$ has positive reach,

$$\int_{x+uh \in \mathbb{K}_s \bigcap B(x,h)} K\left(\|u\|\right) du = \int_{B(0,1)} K\left(\|u\|\right) du(1+O(h^2)) = \frac{1}{C_s^\dagger}(1+O(h^2)).$$

Using this and the fact that $\tau(x) = s$, the quantity (I) equals

$$(I) = \frac{1}{C_s^\dagger} h^{s-d} \rho(x) + O(h^{s-d+2}) = \frac{1}{C_{\tau(x)}^\dagger} h^{\tau(x)-d} \rho(x) + O(h^{\tau(x)-d+2}).$$

To bound the second quantity (II), note that the set $\mathbb{K}_\ell \bigcap B(x,h)$ has ℓ-dimensional volume $O(h^\ell)$. Thus,

$$\int_{\mathbb{K}_\ell \bigcap B(x,h)} \frac{1}{h^d} K\left(\frac{\|x-y\|}{h}\right) dP(y) \leq \frac{1}{h^d} \int_{\mathbb{K}_\ell \bigcap B(x,h)} K(0) \rho_{\max} dy = O\left(h^{\ell-d}\right).$$

And the smallest possible ℓ is $\ell = s + m(x)$. Using this and the bound on (I), we have

$$\begin{aligned}p_h(x) &= \frac{1}{C_s^\dagger} h^{s-d} \rho(x) + O(h^{s-d+2}) + O(h^{s-d+m(x)}) \\ &= \frac{1}{C_{\tau(x)}^\dagger} h^{\tau(x)-d} \rho(x) + O(h^{\tau(x)-d+2}) + O(h^{\tau(x)-d+m(x)}).\end{aligned}$$

Therefore, multiplying both side by $C_s^\dagger h^{d-\tau(x)}$, we obtain

$$C_{\tau(x)}^\dagger h^{d-\tau(x)} \cdot p_h(x) = \rho(x) + O(h^2) + O(h^{m(x)}),$$

which proves the first assertion. Note that if $m(x) = 0$, the we will not have the second term (II) so there will be no dimensional bias $O(h^{m(x)})$.

Part 2: failure of uniform convergence of the bias. Without loss of generality, let (s, ℓ) be the two lower dimensional support such that $\overline{\mathbb{K}}_\ell \bigcap \mathbb{K}_s \neq \phi$ and $s < \ell$.

Let $x \in \overline{\mathbb{K}}_\ell \bigcap \mathbb{K}_s$ be a point on $\mathbb{K}_s$. Then by the first assertion, we have

$$C_s^\dagger h^{d-s} \cdot p_h(x) = \rho(x) + O(h^2) + O(h^{m(x)}).$$

Now consider a sequence of points when $h \to 0$: $\{x_h \in \mathbb{K}_\ell : \|x_h - x\| = h^2\}$. We can always find such a sequence because $\overline{\mathbb{K}}_\ell \bigcap \mathbb{K}_s \neq \phi$. For such a sequence, the set $B(x_h, h) \bigcap K_s$ converges to the set $B(x,h) \bigcap K_s$ in the sense that

$$\frac{P\left((B(x_h,h) \bigcap K_s) \triangle (B(x,h) \bigcap K_s)\right)}{P\left(B(x,h) \bigcap K_s\right)} \to 0$$

when $h \to 0$. This is because the distance in the center of balls shrinks at rate h^2 but the radius of the ball shrinks at rate h.

Thus, $\frac{p_h(x_h)}{p_h(x)} \to 1$ when $h \to 0$. This implies that $C_s^\dagger h^{d-s} \cdot p_h(x_h) \to \rho(x)$ so

$$C^\dagger_{\tau(x_h)} h^{d-\tau(x_h)} \cdot p_h(x_h) = C^\dagger_\ell h^{d-\ell)} \cdot p_h(x_h) = O(h^{s-\ell})$$

diverges. Thus, we do not have uniform convergence for the bias.

□

Before we proceed to the proof of Theorem 8, we first derive a useful lemma about the variance of the KDE.

LEMMA 16 (Pointwise variance). *Assume (S, P2, K1–2). Then for $x \in \mathbb{K}(h)$,*

$$\mathsf{Var}(\widehat{p}_n(x)) = O\left(\frac{1}{nh^{2d-\tau(x)}}\right),$$
$$\mathsf{Var}(C_{\tau(x)} h^{d-\tau(x)} \cdot \widehat{p}_n(x)) = O\left(\frac{1}{nh^{\tau(x)}}\right).$$

PROOF. Without loss of generality, let $x \in \mathbb{K}_s(h)$. This implies that $B(x,h) \cap \mathbb{K}_\ell = \phi$ for all $\ell < s$. By definition,

$$\begin{aligned}
\mathsf{Var}(\widehat{p}_n(x)) &= \mathbb{E}(\widehat{p}_n(x) - \mathbb{E}(\widehat{p}_n(x)))^2 \\
&= \mathbb{E}\left(\frac{1}{nh^d}\sum_{i=1}^n \left(K\left(\frac{\|X_i - x\|}{h}\right) - \mathbb{E}\left(K\left(\frac{\|X_i - x\|}{h}\right)\right)\right)\right)^2 \\
&= \frac{1}{n^2h^{2d}}\mathbb{E}\left(\sum_{i=1}^n \left(K^2\left(\frac{\|X_i - x\|}{h}\right) - n\left(\mathbb{E}^2\left(K\left(\frac{\|X_i - x\|}{h}\right)\right)\right)\right)\right) \\
&\le \frac{1}{nh^{2d}}\mathbb{E}\left(K^2\left(\frac{\|X_i - x\|}{h}\right)\right) \\
&= \frac{1}{nh^{2d}}\int_{B(x,h)} K^2\left(\frac{\|y - x\|}{h}\right) dP(y) \\
&= \frac{1}{nh^{2d}}\sum_{\ell \le s}\int_{B(x,h)\cap\mathbb{K}_\ell} K^2\left(\frac{\|y - x\|}{h}\right)\rho(y)dy \\
&\le \frac{1}{nh^{2d}}\sum_{\ell \le s}\int_{x+uh\in B(x,h)\cap\mathbb{K}_\ell} K^2(\|u\|)\rho_{\max}du \cdot h^\ell \\
&= O\left(\frac{h^s}{nh^{2d}}\right) = O\left(\frac{1}{nh^{2d-s}}\right).
\end{aligned}$$

Note that in the last inequality, we use the transform $y = x + uh$ and because $x + uh$ has to be on $\mathbb{K}_\ell$, there is only ℓ degree of freedom for u. Thus, the change of variable gives $dy = du \cdot h^s$. The above derivation is for the case $x \in \mathbb{K}_s(h)$ so $\tau(x) = s$. This proves the first assertion. The second assertion follows trivially from the first assertion.

□

PROOF OF THEOREM 8. Without loss of generality, we pick a point $x \in \mathbb{K}_s(h)$. The difference has the following decomposition:

$$C_s^\dagger h^{s-d} \cdot \widehat{p}_n(x) - \rho(x) = C_s^\dagger h^{s-d} \cdot (\widehat{p}_n(x) - p_h(x)) + C_s^\dagger h^{s-d} \cdot p_h(x) - \rho(x). \tag{18}$$

The former part is the stochastic variation and the latter part is the bias. The bias is controlled by Theorem 6:

$$C_s^\dagger h^{s-d} \cdot p_h(x) - \rho(x) = O(h^{2 \wedge m_{\min}}).$$

Thus, in what follows we will control the stochastic variation.

It is well-known that the quantity $|\widehat{p}_n(x) - p_h(x)|$ can be written as an empirical process (Einmahl and Mason, 2005; Giné and Guillou, 2002) and to control the supremum $\sup_{x \in \mathbb{K}_s} C_s h^{d-s} \cdot |\widehat{p}_n(x) - p_h(x)|$, we need to uniformly bound the variance. By Lemma 16, the variance is uniformly bounded at rate $O\left(\frac{1}{nh^s}\right)$. Therefore, by the assumption (K2) and applying Theorem 2.3 in Giné and Guillou (2002), we have

$$\sup_{x \in \mathbb{K}_s(h)} |C_s^\dagger h^{s-d} \cdot (\widehat{p}_n(x) - p_h(x)) = O_P\left(\sqrt{\frac{\log n}{nh^s}}\right),$$

which together with the bias term proves the desired result for density estimation.

The case of the gradient and the Hessian can be proved in a similar way as the density estimation case so we ignore the proof. The only difference is that the stochastic part has variance $\frac{1}{nh^{s+2}}$ and $\frac{1}{nh^{s+4}}$; the extra $+2$ and $+4$ in the power of h come from taking the derivatives.

□

PROOF OF THEOREM 9. Recalled that

$$\widehat{\alpha}_n(x) = 1 - \widehat{P}_n(\{y : \widehat{p}_n(y) \geq \widehat{p}_n(x)\}).$$

Now we consider a modified version

$$\widetilde{\alpha}_n(x) = 1 - P(\{y : \widehat{p}_n(y) \geq \widehat{p}_n(x)\}) = 1 - P(\widehat{\Omega}_n(x)).$$

It is easy to see that $|\widehat{\alpha}_n(x) - \widetilde{\alpha}_n(x)| = O_P\left(\sqrt{\frac{1}{n}}\right)$. Thus, all we need is to compare $\widetilde{\alpha}_n(x)$ to $\alpha(x)$.

By definition of $\alpha(x)$,

$$\begin{aligned} & 1 - \alpha(x) \\ & = P(\{y : \ell(y) \geq \ell(x)\}) \\ & = P\left(\{y : \tau(y) > \tau(x)\} \bigcup \{y : \tau(y) = \tau(x),\ \rho(y) \geq \rho(x)\}\right) \\ & = P\left(\{y : \tau(y) > \tau(x)\}\right) + P\left(\{y : \tau(y) = \tau(x),\ \rho(y) \geq \rho(x)\}\right) \\ & = P(\Omega(x)) + P(D(x)), \end{aligned} \tag{19}$$

where $\Omega(x) = \{y : \tau(y) > \tau(x)\}$ and $D(x) = \{y : \tau(y) = \tau(x),\ \rho(y) \geq \rho(x)\}$. Define $E(x) = \{y : \tau(y) = \tau(x),\ \rho(y) < \rho(x)\}$ and $\Phi(x) = \{y : \tau(y) < \tau(x)\}$ then $\Omega(x), D(x), E(x), \Phi(x)$ form a partition of $\mathbb{K}$.

Using the fact that $\Omega(x), D(x), E(x), \Phi(x)$ is a partition of $\mathbb{K}$, we bound the difference

$$
\begin{aligned}
|\alpha(x) - \widetilde{\alpha}_n(x)| &= P\left(\widehat{\Omega}_n(x) \triangle (\Omega(x) \cup D(x))\right) \\
&= P\left((\Omega(x) \cup D(x)) \backslash \widehat{\Omega}_n(x)\right) + P\left(\widehat{\Omega}_n(x) \backslash (\Omega(x) \cup D(x))\right) \\
&= P\left(\Omega(x) \backslash \widehat{\Omega}_n(x)\right) \\
&\quad + P\left(D(x) \backslash \widehat{\Omega}_n(x)\right) + P\left(\widehat{\Omega}_n(x) \cap (E(x) \cup \Phi(x))\right) \\
&= P\left(\Omega(x) \backslash \widehat{\Omega}_n(x)\right) + P\left(D(x) \backslash \widehat{\Omega}_n(x)\right) \\
&\quad + P\left(\widehat{\Omega}_n(x) \cap E(x)\right) + P\left(\widehat{\Omega}_n(x) \cap \Phi(x)\right) \\
&\leq \underbrace{P\left(\left(\Omega(x) \backslash \widehat{\Omega}_n(x)\right) \cap \mathbb{K}(h)\right)}_{\text{(I)}} + \underbrace{P\left(\left(D(x) \backslash \widehat{\Omega}_n(x)\right) \cap \mathbb{K}(h)\right)}_{\text{(II)}} \\
&\quad + \underbrace{P\left(\widehat{\Omega}_n(x) \cap E(x) \cap \mathbb{K}(h)\right)}_{\text{(III)}} + \underbrace{P\left(\widehat{\Omega}_n(x) \cap \Phi(x) \cap \mathbb{K}(h)\right)}_{\text{(IV)}} + \underbrace{P(\mathbb{K}^C(h))}_{\text{(V)}}.
\end{aligned} \tag{20}
$$

Our approach is to first control (I) and (IV) and then control (II) and (III). Note that Lemma 17 controls the quantity (V):

$$
(V) \leq O(h^{2 \wedge m_{\min}}). \tag{21}
$$

Bounding (I) and (IV). We first focus on the set $\mathbb{K}(h)$ and $\mathbb{K}_s(h)$. For a point $x \in \mathbb{K}_s(h)$, it must be at least h distance away from lower dimensional support. Thus, its estimated density $\widehat{p}_n(x)$ will not contain any additional probability mass from lower dimensional support. Therefore, by Theorem 6 and Lemma 16, the scaled density

$$
C_s^\dagger h^{d-s} \cdot \widehat{p}_n(x) - \rho(x) = O(h^{2 \wedge m(x)}) + O_P\left(\sqrt{\frac{1}{nh^s}}\right). \tag{22}
$$

Moreover, we can easily extend equation (22) to a uniform bound

$$
\sup_{x \in \mathbb{K}_s(h)} |C_s^\dagger h^{d-s} \cdot \widehat{p}_n(x) - \rho(x)| = O(h^{2 \wedge m_{\min}}) + O_P\left(\sqrt{\frac{\log n}{nh^s}}\right) = \delta_{n,h,s}. \tag{23}
$$

This implies that for another point $y \in \mathbb{K}_\ell(h)$ where $\ell < s$,

$$
\sup_{x \in \mathbb{K}(h)} \frac{\widehat{p}_n(x)}{\widehat{p}_n(y)} = O(h \cdot \delta_{n,h,s}). \tag{24}
$$

Namely, eventually we will be abel to separate points in different dimensional support if these points are in the good region $\mathbb{K}(h)$. Thus, the contribution from lower dimensional support $\Omega_n(x) \bigcap \mathbb{K}(h)$ can be well-estimated by $\widehat{\Omega}_n$ in the sense that

$$
P\left(\left(\Omega(x) \backslash \widehat{\Omega}_n(x)\right) \cap \mathbb{K}(h)\right) = O(h \cdot \delta_{n,h,s}). \tag{25}
$$

This bounds the quantity (I). And equation (24) also implies

$$P\left(\widehat{\Omega}_n(x) \cap \Phi(x) \cap \mathbb{K}(h)\right) = O(h \cdot \delta_{n,h,s}), \tag{26}$$

which bounds the quantity (IV). Moreover, since equation (24) is an uniform result in $\mathbb{K}_s(h)$, the bounds in (I) and (IV) are uniform for all $x \in \mathbb{K}(h)$.

Bounding (II) and (III). We consider the case in (II) and (III) together. Again we consider $x \in \mathbb{K}_s$. Recalled the bound (II) and (III) are the probability within the regions

$$\left(D(x) \backslash \widehat{\Omega}_n(x)\right) \cap \mathbb{K}(h), \quad \widehat{\Omega}_n(x) \cap E(x) \cap \mathbb{K}(h).$$

Now define the region $\Psi(x;h) = \{y : \tau(y) = \tau(x)\} \cap \mathbb{K}(h)$. Thus, we have

$$\begin{aligned}(II) + (III) &\leq P\left(\left(\widehat{\Omega}_n(x) \triangle \Omega(x)\right) \cap \Psi(x;h)\right) \\ &= P\left((\widehat{\Omega}_n(x) \cap \Psi(x;h)) \triangle (\Omega(x) \cap \Psi(x;h))\right).\end{aligned} \tag{27}$$

The event

$$\begin{aligned}\widehat{\Omega}_n(x) \cap \Psi(x;h) &= \{y \in \mathbb{K}(h) : \widehat{p}_n(y) \geq \widehat{p}_n(x), \tau(y) = \tau(x), x \in \mathbb{K}(h)\} \\ &= \big\{y \in \mathbb{K}(h) : C_s^\dagger h^{d-s} \cdot \widehat{p}_n(y) \geq C_s^\dagger h^{d-s} \cdot \widehat{p}_n(x), \\ &\qquad \tau(y) = \tau(x) = s, x \in \mathbb{K}(h)\big\}.\end{aligned}$$

which can be viewed as the estimated density upper level set at level $\widehat{p}_n(x)$ of the support $\mathbb{K}_s(h)$ (because $\tau(x) = s$). Similarly, $\Omega(x) \cap \Psi(x;h)$ is just the upper level set at level $\rho(x)$ of the support $\mathbb{K}_s(h)$. Therefore, the difference can be bounded by Theorem 4:

$$(II) + (III) \leq \begin{cases} \delta_{n,h,s}, & \text{if} \quad \inf_{c \in \mathcal{C}_s} |p(x) - p(c)| > r_{n,h,s} \\ (\delta_{n,h,s})^{\frac{s}{s+1}}, & \text{otherwise} \end{cases},$$

where $r_{n,h,s}$ is a deterministic quantity such that $\delta_{n,h,s} = o_P(r_{n,h,s})$.

Thus, putting the above bound and equations (21), (25), and (26) into equation (20), we have

$$\begin{aligned}|\alpha(x) - \widetilde{\alpha}_n(x)| &\leq (I) + (II) + (III) + (IV) + (V) \\ &\leq O(h \cdot \delta_{n,h,s}) + O(h^{2 \wedge m_{\min}}) \\ &\quad + \begin{cases} \delta_{n,h,s}, & \text{if} \quad \inf_{c \in \mathcal{C}_s} |p(x) - p(c)| > r_{n,h,s} \\ (\delta_{n,h,s})^{\frac{s}{s+1}}, & \text{otherwise} \end{cases} \\ &= \begin{cases} \delta_{n,h,s}, & \text{if} \quad \inf_{c \in \mathcal{C}_s} |p(x) - p(c)| > r_{n,h,s} \\ (\delta_{n,h,s})^{\frac{s}{s+1}}, & \text{otherwise} \end{cases},\end{aligned}$$

which along with the fact that $|\widehat{\alpha}_n(x) - \widetilde{\alpha}_n(x)| = O_P\left(\sqrt{\frac{1}{n}}\right)$ proves the desired result.

□

Before we move on to the proof of Theorem 10, we first give a lemma that quantifies the 'size' of bad regions.

LEMMA 17 (Size of good region). *Assume (S, P2, K1–2). Define* $m_{\min}$ *from equation* (9) *and*

$$m^*_{\min} = d - \max\{s < d : x \in \mathbb{K}_s \cap \overline{\mathbb{K}}_d\}.$$

Let μ *be the Lebesgue measure. Then*

$$\mu(\mathbb{K}^C(h)) = O(h^{m^*_{\min}}),$$
$$P(\mathbb{K}^C(h)) = O(h^{m_{\min}}).$$

PROOF. **Case of the Lebesgue measure.** By assumption (S), all supports $\mathbb{K}_s$ have Lebesgue measure $\mu(\mathbb{K}_s) = 0$ except $\mathbb{K}_d$. Thus,

$$\mu(\mathbb{K}^C(h)) = \mu(\mathbb{K}^C_d(h)).$$

By the definition of $m^*_{\min}$, the quantity $d - m^*_{\min} = \max\{s < d : x \in \mathbb{K}_s \cap \overline{\mathbb{K}}_d\}$ denotes the support with largest dimension that intersects the closure of $\mathbb{K}_d$.

Note that for any two compact sets A, B with dimension $d(a)$ and $d(b)$ such that $d(b) > d(a)$ and $\overline{A} \cap \overline{B} \neq \phi$, then the $d(b)$-dimensional Lebesgue measure

$$B \cap (A \oplus r) = O(r^{d(b)-d(a)}) \tag{28}$$

when $r \to 0$. Thus, the set $\mu(\mathbb{K}_d \cap (\mathbb{K}_{d-m^*_{\min}} \oplus h))$ shrinks at rate $O(h^{m^*_{\min}})$.

Recalled that $\mathbb{K}_d(h) = \mathbb{K}_d \backslash (\bigcup_{\ell<d} \mathbb{K}_\ell \oplus h)$ and $\mathbb{K}^C_d(h) = (\bigcup_{\ell<d} \mathbb{K}_\ell \oplus h)$. Because only sets in $\mathbb{K}_d$ has non-zero Lebesgue measure, we have

$$\begin{aligned}
\mu(\mathbb{K}^C_d(h)) &= \mu(\mathbb{K}^C_d(h) \cap \mathbb{K}_d) \\
&= \mu\left(\mathbb{K}_d \cap (\bigcup_{\ell<d} \mathbb{K}_\ell \oplus h)\right) \\
&= \mu\left(\mathbb{K}_d \cap (\bigcup_{\ell \le d-m^*_{\min}} \mathbb{K}_\ell \oplus h)\right) \\
&\le \sum_{\ell=0}^{d-m^*_{\min}} \mu\left(\mathbb{K}_d \cap (\mathbb{K}_\ell \oplus h)\right) \\
&= O(h^{m^*_{\min}}) + o(h^{m^*_{\min}}),
\end{aligned}$$

which proves the first assertion.

Case of the probability measure. The case for the probability measure is very similar to the case of Lebesgue measure but now we also need to consider lower dimensional support because of the singular probability measure.

First we expand the probability of bad regions by the follows:

$$
\begin{aligned}
P(\mathbb{K}^C(h)) &= 1 - P(\mathbb{K}(h)) \\
&= 1 - P(\bigcup_{\ell \le d} \mathbb{K}_\ell(h)) \\
&= 1 - \sum_{\ell=0}^{d} P(\mathbb{K}_\ell(h)) \\
&= 1 - \sum_{\ell=0}^{d} P\left(\mathbb{K}_\ell \backslash \left(\bigcup_{s<\ell} \mathbb{K}_s \oplus h\right)\right) \\
&= \sum_{\ell=0}^{d} P(\mathbb{K}_\ell) - P\left(\mathbb{K}_\ell \backslash \left(\bigcup_{s<\ell} \mathbb{K}_s \oplus h\right)\right) \\
&= \sum_{\ell=0}^{d} P\left(\mathbb{K}_\ell \bigcap \left(\bigcup_{s<\ell} \mathbb{K}_s \oplus h\right)\right) \\
&= \sum_{\ell=0}^{d} P\left(\mathbb{K}_\ell \bigcap \left(\bigcup_{s<\ell} \mathbb{K}_s \oplus h\right)\right)
\end{aligned} \tag{29}
$$

Note that we use the fact that $P(A) - P(A \backslash B) = B(A \cap B)$ in the last equality. We will show that

$$
P\left(\mathbb{K}_\ell \bigcap \left(\bigcup_{s<\ell} \mathbb{K}_s \oplus h\right)\right) = O(h^{m_{\min}}). \tag{30}
$$

for every ℓ. For simplicity, we define $\mathbb{K}_\ell^C(h) = \mathbb{K}_\ell \bigcap \left(\bigcup_{s<\ell} \mathbb{K}_s \oplus h\right)$.

Without loss of generality, we consider the support $\mathbb{K}_\ell$ and $\mathbb{K}_\ell^C(h)$. For simplicity, By the definition of $m_{\min}$, the largest lower dimensional support that intersects the closure $\overline{\mathbb{K}}_\ell$ has dimension lower than or equal to $\mathbb{K}_{\ell - m_{\min}}$. Note that if $\ell < m_{\min}$ then it is easy to see that there will be no other lower dimensional support intersecting $\overline{\mathbb{K}}_\ell$ so $\mathbb{K}_\ell(h) = \mathbb{K}_\ell$ and there is nothing to prove. So the set $\mathbb{K}_\ell^C(h)$ can be rewritten as

$$
\mathbb{K}_\ell^C(h) = \mathbb{K}_\ell \bigcap \left(\bigcup_{s \le \ell - m_{\min}} \mathbb{K}_s \oplus h\right)
$$

Now by equation (28), the ℓ-dimensional Lebesgue measure μ_ℓ on the set $\mathbb{K}_\ell^C(h)$ is at rate

$$
\mu_\ell\left(\mathbb{K}_\ell^C(h)\right) = O(h^{m_{\min}}) + o(h^{m_{\min}}).
$$

By assumption (P2), the ℓ-dimensional Lebesgue measure on $\mathbb{K}_\ell$ implies that bound on probability measure so we have

$$
P\left(\mathbb{K}_\ell^C(h)\right) = O(h^{m_{\min}}).
$$

Because this works for every ℓ, by equation (29) and (30), we have

$$
P(\mathbb{K}^C(h)) = O(h^{m_{\min}}),
$$

which proves the result.

□

PROOF OF THEOREM 10. We first note that it is easy to see $m^*_{\min} \geq m_{\min}$ so the rate of the integrated error is bounded by the rate of the probability error. Thus, here we only prove the case for the probability error.

Because $\mathbb{K}_0, \cdots, \mathbb{K}_d$ form a partition of $\mathbb{K}$. We separately analyze the probability error at each $\mathbb{K}_\ell$ and then joint them to get the final bound.

For a support $\mathbb{K}_\ell$, we partition it into three subregions A, B, C, where

$$\begin{aligned} A &= \mathbb{K}_\ell^C(h) = \mathbb{K}_\ell \backslash \mathbb{K}_\ell(h) \\ B &= \mathbb{K}_\ell(h) \cap \{x : \min_{c \in \mathcal{C}_s} |\rho(x) - \rho(c)| \leq r_{n,h,s}\} \\ C &= \mathbb{K}_\ell(h) \cap \{x : \min_{c \in \mathcal{C}_s} |\rho(x) - \rho(c)| > r_{n,h,s}\}, \end{aligned} \tag{31}$$

where $r_{n,h,s} = \frac{h^2 \wedge m_{\min} + \frac{\log n}{nh^s}}{\log n}$ satisfies the requirement $\frac{\delta_{n,h,s}}{r_{n,h,s}} = o_P(1)$.

Case A. By Lemma 17, $P(A) = O(h^{m_{\min}})$ and $|\widehat{\alpha}(x) - \alpha(x)| \leq 1$. Thus,

$$\int_A \|\widehat{\alpha}_n(x) - \alpha(x)\| dP(x) = O(h^{m_{\min}}). \tag{32}$$

Case B. For set B, note that the generalized density $\rho(x)$ behaves quadratically around critical points. So difference in density level at rate δ results in the difference in the difference in distance at rate $\sqrt{\delta}$. Thus, the ℓ-dimensional Lebesgue measure $\mu_\ell(B) = O(\sqrt{r_{n,h,s}})$, which by assumption (P2) implies $P(B) = O(\sqrt{r_{n,h,s}})$. By Theorem 9, $\|\widehat{\alpha}_n(x) - \alpha(x)\| = \delta_{n,h,s}^{\frac{s}{s+1}}$ uniformly for all $x \in B$. Thus, the error is

$$\int_B \|\widehat{\alpha}_n(x) - \alpha(x)\| dP(x) = O_P\left(\sqrt{r_{n,h,s}} \cdot \delta_{n,h,s}^{\frac{s}{s+1}}\right). \tag{33}$$

Case C. For points in this region, directly applying Theorem 9 yields

$$\int_C \|\widehat{\alpha}_n(x) - \alpha(x)\| dP(x) = O_P\left(\delta_{n,h,s}\right). \tag{34}$$

By adding up equation (32), (33), and (34) and use the fact that $\sqrt{r_{n,h,s}} \cdot \delta_{n,h,s}^{\frac{s}{s+1}} = o(\delta_{n,h,s})$ and $O(h^{m_{\min}})$ is part of $\delta_{n,h,s}$, we obtain

$$\int_{\mathbb{K}_\ell} \|\widehat{\alpha}_n(x) - \alpha(x)\| dP(x) = O_P\left(\delta_{n,h,s}\right). \tag{35}$$

This works for every $\mathbb{K}_\ell$, which proves the desired bound.

□

PROOF OF LEMMA 11. **First assertion.** By assumption (B) and (M), the first assertion is trivially true since when we move down the level α, the only situation to have a new connected component is when α pass through the α-level of a local mode.

Second assertion. For level sets, there are only two situations of change in topology: creation of a new connected component, and merging of two (or more) connected components. By the first assertion, we only need to focus on the merging case.

Assume $a \in \mathcal{A}$ be a level that only merging of connected component occurs. And recall $\xi(a)$ is the integer such that $\mathbb{K}_s \subset \mathbb{A}_a$ for all $s \leq \xi(\alpha)$ and $\mathbb{K}_{\xi(a)+1} \not\subset \mathbb{A}_a$. For ϵ sufficiently small, the difference between $\mathbb{A}_a$ and $\mathbb{A}_{a+\epsilon}$ is in $\mathbb{K}_{\xi(a)+1}$. Thus, when we move ϵ down, only the connected components in $\mathbb{K}_{\xi(\alpha)+1}$ extend. When the merging occurs, there are only two cases: two connected components in $\mathbb{K}_{\xi(a)+1}$ meet each other, or a connected components in $\mathbb{K}_{\xi(a)+1}$ hits a lower dimensional support. Note that there is no higher dimensional support in $\mathbb{A}_a$ because their ordering is less than any point in $\mathbb{K}_{\xi(a)}$. The first case corresponds to a saddle point, which is an element in $\mathcal{C}$. The second case corresponds to a DCP. Thus, a must be either an α-level of a critical point or a DCP.

□

Proof of Lemma 12. Without loss of generality, let c be a critical point of $\rho(x)$ on $\mathbb{K}_s$. Namely, the gradient in the tangent space $g_s(c) = \nabla_{T_s(c)}\rho(c) = 0$. Note that $\nabla_{T_s(c)}$ denotes taking gradient along the tangent space of $\mathbb{K}_s$ at the point $c \in \mathbb{K}_s$. By assumption (C), c is away from lower dimensional support so when h is sufficiently small, $c \in \mathbb{K}(h)$. Thus, by Theorem 8,

$$\begin{aligned} C_s^\dagger h^{s-d} \cdot \nabla_{T_s(c)} \widehat{p}_n(c) &= C_s^\dagger h^{s-d} \cdot \nabla_{T_s(c)} \widehat{p}_n(c) - \nabla_{T_s(c)}\rho(c) \\ &= O(h^{2 \wedge m(c)}) + O_P\left(\sqrt{\frac{1}{nh^{s+2}}}\right). \end{aligned} \tag{36}$$

Note that here we only need a pointwise bound for the gradient so the bias depends on $m(c)$ rather than $m_{\min}$ and we will not have the $\sqrt{\log n}$ in the O_P term.

Let $\widehat{c}$ be an estimator to c. Recalled that $N_s(c)$ denotes the normal space of $\mathbb{K}_s$ at the point $c \in \mathbb{K}_s$ Note that the full gradient ∇ can be decomposed in to the gradient along the tangent space and the normal space, i.e. $\nabla = [\nabla_{T_s(c)}, \nabla_{N_s(c)}]$. Using Taylor expansion and $\nabla_{T_s(c)}\widehat{p}_n(\widehat{c}) = 0$, the left hand side of equation (36) becomes

$$\begin{aligned} C_s^\dagger h^{s-d} \cdot \nabla_{T_s(c)} &= O(h^{2 \wedge m(c)}) + O_P\left(\sqrt{\frac{1}{nh^{s+2}}}\right) \\ &= C_s^\dagger h^{s-d} \cdot \nabla_{T_s(c)}(\widehat{p}_n(c) - \widehat{p}_n(\widehat{c})) \\ &= C_s^\dagger h^{s-d} \cdot \nabla \nabla_{T_s(c)} \widehat{p}_n(c)(c - \widehat{c}) + O(\|\widehat{c} - c\|^2) \\ &= C_s^\dagger h^{s-d} \cdot [\nabla_{T_s(c)} \nabla_{T_s(c)}, \nabla_{N_s(c)} \nabla_{T_s(c)}] \widehat{p}_n(c)(c - \widehat{c}) + O(\|\widehat{c} - c\|^2) \end{aligned}$$

It is easy to see that $h^{s-d}\nabla_{N_s(c)}\nabla_{T_s(c)}\widehat{p}_n(c)$ diverges so the components of $c - \widehat{c}$ in the normal subspace $N_s(c)$ converges faster than in the tangent subspace $T_s(c)$. The other term $h^{s-d}\nabla_{T_s(c)}\nabla_{T_s(c)}\widehat{p}_n(c)$ converges to the generalized Hessian at c whose eigenvalues are bounded away from 0 Thus, the inverse of

$$h^{s-d}\nabla_{T_s(c)}\nabla_{T_s(c)}\widehat{p}_n(c)$$

exists, and thus the component of $c - \widehat{c}$ on the tangent space $T_s(c)$ is at rate $O(h^{2 \wedge m(c)}) + O_P\left(\sqrt{\frac{1}{nh^{s+2}}}\right)$. This proves that $c - \widehat{c} = O(h^{2 \wedge m(c)}) + O_P\left(\sqrt{\frac{1}{nh^{s+2}}}\right)$.

To derive the rate of $|\widehat{\alpha}_n(\widehat{c}) - \alpha(c)|$, note that we can decompose it into

$$|\widehat{\alpha}_n(\widehat{c}) - \alpha(c)| \leq |\widehat{\alpha}_n(\widehat{c}) - \widehat{\alpha}_n(c)| + |\widehat{\alpha}_n(c) - \alpha(c)|.$$

Because the KDE around $\widehat{c}$ behaves quadratically, the difference in distance results in square of difference in density. Thus, the second term dominates the first term. The quantity $|\widehat{\alpha}_n(c) - \alpha(c)|$ can be bounded by Theorem 9, which is at rate $\delta_{n,h,s}^{\frac{s}{s+1}}$.

For the eigenvalues, it is easy to see that when we are moving away from c along a normal direction within $N_s(c)$, the estimated density is going down. Thus, these $(d-s)$ directions must have negative eigenvalues (dimension of the normal subspace is $d-s$). And the original generalized Hessian matrix at c has $n(c)$ negative eigenvalues. So the total number of negative eigenvalue of the Hessian matrix of $\widehat{p}_n$ at c is $n(s) + d - s$. Because $\widehat{c}$ is converging to c, the sign of negative eigenvalues also converges, which proves the lemma.

□

Proof of Lemma 13. **First assertion: location of DCPs.** By Theorem 8, the scaled KDE are uniformly consistent in both density estimation and gradient estimation in the good region $\mathbb{K}(h)$.

The estimated dimensional critical points are points satisfying $\nabla\widehat{p}_n(x) = 0$. Therefore, when $\delta_{n,h,d}^{(2)} \xrightarrow{P} 0$, the only area in $\mathbb{K}(h)$ such that $\nabla\widehat{p}_n(x) = 0$ will be the regions where $\nabla_{T_{\tau(x)}(x)}\rho(x)$ is small. This can only be the regions around generalized critical points. As a result, we cannot have any dimensional critical points within $\mathbb{K}(h)$ when $\delta_{n,h,d}^{(2)} \xrightarrow{P} 0$.

Second assertion: number of estimated critical points. This follows directly from Lemma 14 and Assumption (A) and the fact that $\delta_{n,h,d}^{(2)} \xrightarrow{P} 0$.

Third assertion: no local modes. By assumption (B), the only case where a creation of a connected component occurs is a (generalized) local mode. These population local modes will correspond to elements in $\widehat{\mathcal{C}}$. Thus, any estimated local mode in $\widehat{\mathcal{D}}$ does not have a population target so they are away from the population local modes and by first assertion, it has to be in the bad region $\mathbb{K}^C(h)$. It is easy to see that we cannot have any estimated local mode under such a constraint when we have the gradient and Hessian consistency of the scaled KDE ($\delta_{n,h,d}^{(2)} \xrightarrow{P} 0$).

□

Proof of Lemma 14. Let c be a DCP and $a_0(c)$ be the corresponding α-level of merging. By assumption (A), c is a unique point for $\alpha_0(c) \in \mathcal{D}$ and for any ϵ that is sufficiently small, there is an unique connected component $C_\epsilon \in \mathbb{A}_{a_0(c)+\epsilon}$ and a support $\mathbb{K}_\ell$ with $\ell \leq \xi(a_0(c))$ such that $x \in \mathbb{K}_\ell$, $c \notin C_\epsilon$ and $d(c, C_\epsilon) \to 0$ when $\epsilon \to 0$.

The idea of the proof is to find $\widehat{a}^+$ and $\widehat{a}^-$ such that in the set $\mathbb{A}_{\widehat{a}^+}$, the merging has not yer happened and in the set $\mathbb{A}_{\widehat{a}^-}$, the merging has happened. Then we know the actual value $\widehat{a}_n(\widehat{c})$ lies within the interval $[\mathbb{A}_{\widehat{a}^-}, \mathbb{A}_{\widehat{a}^+}]$.

Case: lower bound. To derive the upper bound, recalled that $\delta_{n,h,s} = O(h^{2 \wedge m(x)}) + O_P\left(\sqrt{\frac{\log n}{nh^s}}\right)$. By Theorem 9, any point $y \in K(h)$ satisfies $|\widehat{\alpha}_n(y) - \alpha(y)| = \delta_{n,h,\tau(y)}$. Thus, for any C_ϵ defined as the connected component within $\mathbb{A}_{a_0(x)+\epsilon}$ that is about to merge with $\mathbb{K}_\ell$,

$$\inf_{y \in C_\epsilon} \widehat{\alpha}_n(y) > a_0(x) - \delta_{n,h,\xi(a_0(c))+1}.$$

This is because $\inf_{y \in C_\epsilon} \alpha(y) \geq a_0(c)$ by definition and for $C_\epsilon \cap \mathbb{K}(h)$, we have the uniform bound from Theorem 9. For $y \in C_\epsilon \backslash \mathbb{K}(h)$, the estimated $\widehat{\alpha}_n(y)$ will be influenced by $\mathbb{K}_\ell$, which is a lower dimensional support. So the estimated α value will be more than $a_0(c) - \delta_{n,h,\xi(a_0(c))+1}$. Thus, we can pick the lower bound $\mathbb{A}_{\widehat{a}^-} = a_0(c) - \delta_{n,h,\xi(a_0(c))+1}$.

Case: upper bound. To prove the upper bound, the idea is very simple. We will show that when the α-level is high enough, $\widehat{\mathbb{A}}_{a'}$ will not contain the set $\widetilde{C}_\ell \oplus h$ where $c \in \widetilde{C}_\ell$ and $\widetilde{C}_\ell$ is a connected component of $\mathbb{A}_{a_0}$.

Let $\widetilde{C}_\ell$ be defined as the above. Because $\widetilde{C}_\ell$ is a subset of $\bigcup_{s\le\ell}\mathbb{K}_s$, the set $C_\epsilon \cap (\widetilde{C}_\ell \oplus h)$ is always within the bad region $\mathbb{K}^C(h)$. Thus, $P(C_\epsilon \cap (\widetilde{C}_\ell \oplus h)) \le P(\mathbb{K}^C(h)) = O(m^{2\wedge m_{\min}})$ by Lemma 17.

Now we consider the boundary $\partial(\widetilde{C}_\ell \oplus h) = \{x : d(x, \widetilde{C}_\ell) = h\}$. Because $P(C_\epsilon \cap (\widetilde{C}_\ell \oplus h)) \le O(m^{2\wedge m_{\min}})$,

$$\sup_{x\in\partial(\widetilde{C}_\ell\oplus h)} \alpha(x) \le a_0(c) + O(m^{2\wedge m_{\min}}).$$

Moreover, outside the boundary $\partial(\widetilde{C}_\ell \oplus h)$ we can apply Theorem 9 to bound $\widehat{\alpha}_n(x) - \alpha(x)$. Thus,

$$\sup_{x\in\partial(\widetilde{C}_\ell\oplus h)} \widehat{\alpha}_n(x) \le a_0(c) + O(m^{2\wedge m_{\min}}) + \delta_{n,h,\xi(a_0(c))+1}.$$

This suggests that $\mathbb{A}_{\widehat{a}^+} = a_0 + O(m^{2\wedge m_{\min}}) + \delta_{n,h,\xi(a_0(c))+1} = a_0 + \delta_{n,h,\xi(a_0(c))+1}$ because $O(m^{2\wedge m_{\min}})$ is part of the term $\delta_{n,h,\xi(a_0(c))+1}$.

Thus, the quantity $\widehat{\alpha}_n(\widehat{c})$ must lie within $a_0 \pm \delta_{n,h,\xi(a_0(c))+1}$, which is the desired bound. Because there is a topological change of upper level set for $\widehat{p}_n$ at such $\widehat{\alpha}_n(\widehat{c})$, $\widehat{c}$ must be a critical point of $\widehat{p}_n$. This completes the proof.

□

Proof of Theorem 15. The main idea is to show that there exists a constant $a_0 > 0$ such that

$$\max_{j=0,1,2}\max_{s=0,\cdots,d} |\delta^{(j)}_{n,h,s}| < a_0 \Longrightarrow T_{\widehat{\alpha}_n} \overset{T}{\approx} T_\alpha. \tag{37}$$

Note that $\delta^{(0)_{n,h,s}} = \delta_{n,h,s}$.

By assumption (A), there exists a constant $a_1 = \min\{|\alpha_1 - \alpha_2| : \alpha_1, \alpha_2 \in \mathcal{A}, \alpha_1 \ne \alpha_2\} > 0$. Without loss of generality, let the elements in $\mathcal{A}$ be $\alpha_1 > \alpha_2 > \cdots > \alpha_m$ (assume $\mathcal{A}$ has m elements) and $c(\alpha_j)$ be the corresponding critical point or the DCP for level α_j. By Lemma 12 and 14, we can find a sequence of points $\widehat{c}_1, \cdots, \widehat{c}_m$ such that each $\widehat{c}_j$ is the estimator to $c(\alpha_j)$. And again by Lemma 12 and 14, when

$$\max_{s=0,\cdots,d} |\delta^{(0)}_{n,h,s}| < \frac{a_1}{2}, \tag{38}$$

we have

$$\widehat{\alpha}_n(\widehat{c}_1) > \cdots > \widehat{\alpha}_n(\widehat{c}_m).$$

Namely, the ordering will be the same when $\max_{s=0,\cdots,d} |\delta^{(0)}_{n,h,s}|$ is sufficiently small. Thus, we need to prove that (1) no other connected component of $\widehat{\alpha}_n$ and (2) there is no other merging point to get the topological equivalent.

By Lemma 13, there will be no estimated local mode in $\widehat{\mathcal{D}}$ so there will be no extra connected component. In the proof of Lemma 14, we have shown that each estimator of a DCP corresponds to a merging in the estimated level sets and similarly if connected components are merged at a saddle point or a local minimum, the corresponding estimator will be a merging point. To use the conclusion of Lemma 12, we need the uniform consistency in both gradient and Hessian estimation. Namely, we need $\delta^{(1)}_{n,h,s}, \delta^{(2)}_{n,h,s}$ to be sufficiently small. The gradient consistency is to regularize the positions of estimators of generalized

critical points and the bound on Hessian matrix is to guarantee that the eigenvalues remain the same sign. Thus, to apply Lemma 12, there is some $a_2 > 0$ such that the conclusion of Lemma 12 holds whenever

$$\max_{j=1,2} \max_{s=0,\cdots,d} |\delta_{n,h,s}^{(j)}| < a_2. \tag{39}$$

To use Lemma 14, we need $\delta_{n,h,s}^{(1)}$ to be sufficiently small; this comes from Theorem 9.

Combining equation (38) and (39), we obtain equation (37). For the quantity $\delta_{n,h,s}^{j}$, the slowest rate occurs at $j = 2$ and $s = d$. Thus, there are some constants $c_1, c_2 > 0$ such that

$$\max_{j=0,1,2} \max_{s=0,\cdots,d} |\delta_{n,h,s}^{(j)}| < c_1 h^{2 \wedge m(x)} + c_2 Z_d, \tag{40}$$

where $Z_d = O_P\left(\sqrt{\frac{\log n}{nh^{d+4}}}\right)$ is the stochastic variation. Recalled that in the proof of Theorem 8, Z_d is

$$Z_d = \sup_{x \in \mathbb{K}} \|\nabla\nabla \widehat{p}_n(x) - \nabla\nabla p_h(x)\|_{\max}.$$

By assumption (K2) and the Talagrand's inequality (Giné and Guillou, 2002; Einmahl and Mason, 2005), there is constants $c_3, c_4 > 0$ such that when $nh^{d+4} \to \infty$

$$P(Z_d > t) < c_3 \cdot e^{-c_4 \cdot nh^{d+4} \cdot t^2}. \tag{41}$$

Thus, when Using equation (40) and (41), when $h^{2 \wedge m(x)} < \frac{a_0}{c_1}$,

$$\begin{aligned} P(T_{\widehat{\alpha}_n} \overset{T}{\approx} T_\alpha) &\geq P(\max_{s=0,\cdots,d} |\delta_{n,h,s}^{(1)}| < a_0) \\ &\geq 1 - P\left(Z_d > \frac{a_0}{c_2}\right) \\ &\geq 1 - c_3 \cdot e^{-c_4 \cdot nh^{d+4} \cdot (\frac{a_0}{c_2})^2} \\ &= 1 - c_3 \cdot e^{-c_5 \cdot nh^{d+4}}, \end{aligned}$$

where $c_5 = c_4 \cdot (\frac{a_0}{c_2})^2 > 0$. This proves the desired result. □

REFERENCES

S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, 2012.

O. Bobrowski, S. Mukherjee, and J. E. Taylor. Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*, 2014.

P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1): 77–102, 2015.

B. Cadre. Kernel estimation of density level sets. *Journal of multivariate analysis*, 97(4):999–1023, 2006.

G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pages 343–351, 2010.

K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. von Luxburg. Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60(12):7900–7912, 2014.

F. Chazal, A. Lieutier, and J. Rossignac. Normal-map between normal-compatible manifolds. *International Journal of Computational Geometry & Applications*, 17(05):403–421, 2007.

F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Density level sets: Asymptotics, inference, and visualization. *arXiv:1504.05438*, 2015a.

Y.-C. Chen, C. R. Genovese, L. Wasserman, et al. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5): 1896–1928, 2015b.

Y.-C. Chen, C. R. Genovese, L. Wasserman, et al. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016a.

Y.-C. Chen, J. Kim, S. Balakrishnan, A. Rinaldo, and L. Wasserman. Statistical inference for cluster trees. *arXiv preprint arXiv:1605.06416*, 2016b.

D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

H. Edelsbrunner and D. Morozov. Persistent homology: theory and practice. In *Proceedings of the European Congress of Mathematics*, pages 31–50, 2012.

U. Einmahl and D. M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

J. Eldridge, M. Belkin, and Y. Wang. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. In *Proceedings of The 28th Conference on Learning Theory*, pages 588–606, 2015.

B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, et al. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

H. Federer. Curvature measures. *Trans. Am. Math. Soc*, 93, 1959.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman, et al. On the path density of a gradient field. *The Annals of Statistics*, 37(6A):3236–3271, 2009.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

J. A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 1981.

B. P. Kent. *Level Set Trees for Applied Statistics*. PhD thesis, Carnegie Mellon University, 2013.

J. Klemelä. Visualization of multivariate density estimates with level set trees. *Journal of Computational and Graphical Statistics*, 13(3), 2004.

J. Klemelä. Visualization of multivariate density estimates with shape trees. *Journal of Computational and Graphical Statistics*, 15(2):372–397, 2006.

J. Klemelä. *Smoothing of multivariate data: density estimation and visualization*, volume 737. John Wiley & Sons, 2009.

S. Kpotufe and U. V. Luxburg. Pruning nearest neighbor cluster trees. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 225–232, 2011.

T. Laloe and R. Servien. Nonparametric estimation of regression level sets. *Journal of the Korean Statistical Society*, 2013.

J. Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2012.

E. Mammen and W. Polonik. Confidence regions for level sets. *Journal of Multivariate Analysis*, 122:202–214, 2013.

D. M. Mason and W. Polonik. Asymptotic normality of plug-in level set estimates. *The Annals of Applied Probability*, 19 (3):1108–1142, 2009.

P. Mattila. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*, volume 44. Cambridge university press, 1999.

J. W. Milnor. *Morse theory*. Number 51. Princeton university press, 1963.

I. Molchanov. Empirical estimation of distribution quantiles of random closed sets. *Theory of Probability & Its Applications*, 35(3):594–600, 1991.

M. Morse. Relations between the critical points of a real function of n independent variables. *Transactions of the American Mathematical Society*, 27(3):345–396, 1925.

M. Morse. The foundations of a theory of the calculus of variations in the large in m-space (second paper). *Transactions of the American Mathematical Society*, 32(4):599–631, 1930.

W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, pages 855–881, 1995.

D. Preiss. Geometry of measures in r n: distribution, rectifiability, and densities. *Annals of Mathematics*, 125(3):537–643, 1987.

A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, Oct. 2010. ISSN

0090-5364. . URL http://arxiv.org/abs/0907.3454.

A. Rinaldo, A. Singh, R. Nugent, and L. Wasserman. Stability of density-based clustering. *The Journal of Machine Learning Research*, 13(1):905–948, 2012.

D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

A. Singh, C. Scott, and R. Nowak. Adaptive hausdorff estimation of density level sets. *The Annals of Statistics*, 37(5B): 2760–2782, 2009.

I. Steinwart. Adaptive density level set clustering. In *COLT*, pages 703–738, 2011.

W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, 20(1):025–047, 2003.

A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

L. W. Tu. *An introduction to manifolds*. Springer Science & Business Media, 2010.

G. Walther. Granulometric smoothing. *The Annals of Statistics*, 25(6):2273–2299, 1997.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195
E-mail: yenchic@uw.edu