The Annals of Statistics
2019, Vol. 47, No. 4, 2174–2203
https://doi.org/10.1214/18-AOS1744

GENERALIZED CLUSTER TREES AND SINGULAR MEASURES

BY YEN-CHI CHEN

University of Washington

In this paper we study the α-cluster tree (α-tree) under both singular and nonsingular measures. The α-tree uses probability contents within a set created by the ordering of points to construct a cluster tree so that it is well defined even for singular measures. We first derive the convergence rate for a density level set around critical points, which leads to the convergence rate for estimating an α-tree under nonsingular measures. For singular measures, we study how the kernel density estimator (KDE) behaves and prove that the KDE is not uniformly consistent but pointwise 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. DCPs are points that contribute to cluster tree topology but cannot be defined using density gradient. Building on the analysis of the KDE and DCPs, we prove the topological consistency of an 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 a level set $\{x : f(x) \geq f_0\}$ when we move down the level f_0 [Klemelä (2004), Stuetzle (2003)]. Because cluster trees keep track of the connected components of level sets, the shape of a cluster tree contains topological information about the underlying function f. Moreover, a cluster tree can be displayed on a two-dimensional plane regardless of the dimension of $\mathcal{M}$, which makes it an attractive approach for visualizing f. Figure 1 provides an example illustrating the construction of a cluster tree in a $1D$ Euclidean space. In this paper, we focus on the case where $f \equiv f_P$ is some function of the underlying distribution P. In this context, the cluster tree of f reveals information about P.

Most of cluster trees being studied in the literature are the λ-tree of a distribution [Balakrishnan et al. (2012), Chaudhuri and Dasgupta (2010), Chaudhuri et al. (2014), Chen et al. (2016), Kpotufe and Luxburg (2011), Stuetzle (2003)]. The λ-tree of a distribution is the cluster tree of the density function p of that distribution. In this case, the tree structure contains the topological information of p and we can use the λ-tree to visualize a multivariate density function. When we use an λ-tree for visualization purposes, it is also called a density tree [Klemelä (2004, 2006, 2009)].

Received November 2016; revised February 2018.

MSC2010 subject classifications. Primary 62G20; secondary 62G05, 62G07.

Key words and phrases. Cluster tree, kernel density estimator, level set, singular measure, critical points, topological data analysis.

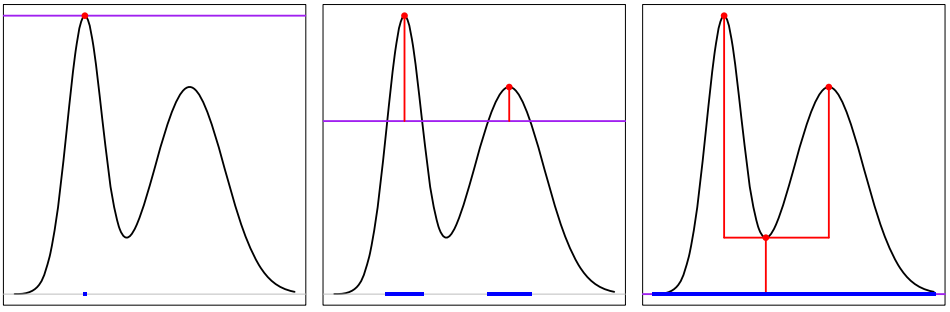

FIG. 1. *An example of constructing a cluster tree in $d = 1$ case. The purple horizontal lines indicates the level f_0 we are using in each panel. The blue region at the bottom indicates the corresponding level set $\{x : f(x) \geq f_0\}$. From left to right, we gradually decrease the level f_0 and use red lines to indicate how the connected components evolve. The resulting red tree in the right panel is a cluster tree.*

Kent (2013) proposed a new type of cluster tree of a distribution—the α-tree. The α-tree uses the function $\alpha(x) = P(\{y : p(y) \leq p(x)\})$ to construct a cluster tree. Note that such a function is also called *density ranking* in our following paper [Chen and Dobra (2017)] and it shows great potential in analyzing GPS datasets. When the distribution is nonsingular and smooth, the α-tree and the λ-tree are topologically equivalent (Lemma 1), so they both provide similar topological information for the underlying distribution. To estimate an 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 $\widehat{p}_n$ is the kernel density estimator (KDE). Namely, we first use the KDE to estimate the density of each data point and count the number of data points with a density below the density of that given point.

When a distribution is singular, the λ-tree is ill-defined because of the lack of a probability density function, but the α-tree is still well defined under a mild modification. For an illustrating example, see Figure 2. These are random samples from a distribution mixed with a point mass at $x = 2$ with a probability of 0.3 and a standard normal distribution with a probability of 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 with the smoothing bandwidth selected by the default rule in R. The estimated density and λ-trees (red trees) are displayed in the top row. It can be seen that when the sample size increases, 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 2).

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

- When the distribution is nonsingular:

1. We derive the convergence rate for the estimated level set when the level equals the density value of a critical point (Theorem 3).

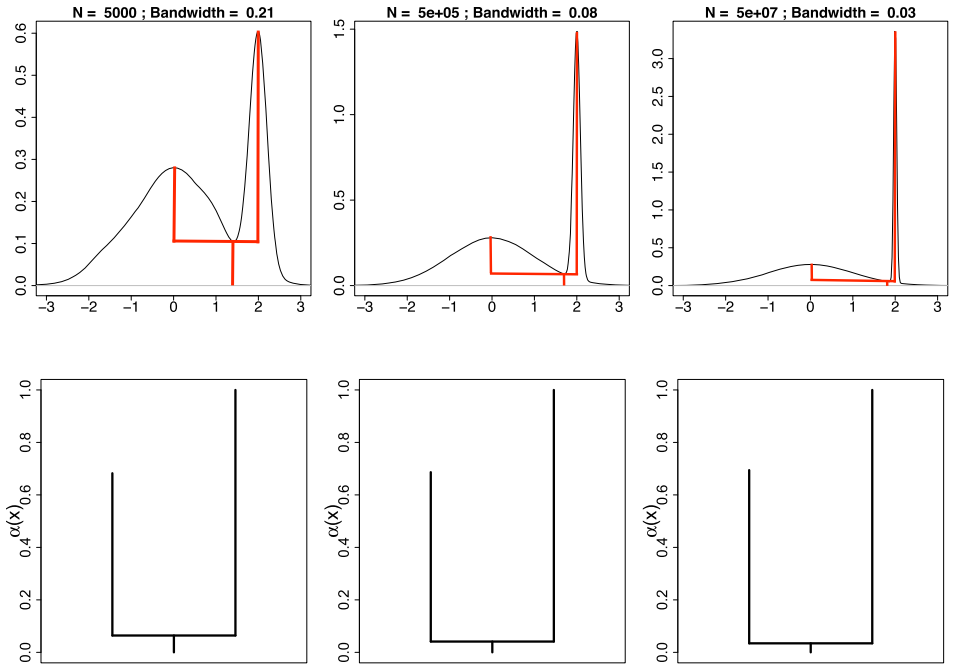

FIG. 2. *Example of the estimated λ-tree and α-tree of a singular distribution. This is a random sample from a singular distribution where with a probability of* 0.3, *it puts a point mass at* $x = 2$, *and with a probability of* 0.7, *we sample from a standard normal. The top panel shows the density estimated by the 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 estimated λ-tree is getting degenerated. On the other hand, the estimated α-trees remain stable regardless of the smoothing bandwidth. Note that every level set in a λ-tree corresponds to a level set in the α-tree but the value of the corresponding level* (*Y-axis*) *will be different. The function used in constructing a λ-tree may have an unbounded range when the sample size goes to infinity, whereas the function for building an α-tree always has a range of* $[0, 1]$.

2. We derived the convergence rate 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 pointwise, but not uniformly, consistent (Theorem 8).

5. We prove that $\widehat{\alpha}_n$ is inconsistent under the L_∞ metric (Corollary 7) but consistent under the L_1 and $L_1(P)$ distance (Theorem 10).

6. We identify a new type of critical points, the dimensional critical points (DCPs), which also contribute to the change in cluster tree topology. We analyze their properties in Lemma 11, 13 and 14.

7. We demonstrate that the estimated α-tree $T_{\widehat{\alpha}_n}$ is topologically equivalent to the population α-tree with probability exponentially converging to 1 (Theorem 15).

Related work. There is extensive literature on theoretical aspects of the λ-tree. Notions of consistency are analyzed in Chaudhuri and Dasgupta (2010), Chaudhuri et al. (2014), Eldridge, Belkin and Wang (2015), Hartigan (1981). The convergence rate and the minimax theory are studied in Balakrishnan et al. (2012), Chaudhuri and Dasgupta (2010), Chaudhuri et al. (2014). Chen et al. (2016) 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 (2013), Wasserman (2018)]. In particular, a cluster tree contains information about the zeroth-order homology groups [Bobrowski, Mukherjee and Taylor (2017), Bubenik (2015), Cohen-Steiner, Edelsbrunner and Harer (2007), Fasy et al. (2014)]. In our analysis, we generalize the Morse theory to a nonsmooth and even discontinuous function. Baryshnikov, Bubenik and Kahle (2014) also generalizes the Morse theory to a nonsmooth function using the concept of configuration space (the collection of n points in a bounded area in R^d). Note that their setting is different from us because we are working on a probability density function whereas the function in Baryshnikov, Bubenik and Kahle (2014) is related to pairwise distance between points and distance to the boundary of certain area. The theory of estimating a cluster tree is closely related to the theory of estimating a level set; an incomplete list of literature is as follows: Mason and Polonik (2009), Polonik (1995), Rinaldo and Wasserman (2010), Singh, Scott and Nowak (2009), Steinwart (2011), Tsybakov (1997), Walther (1997).

Outline. We begin with an introduction of cluster trees and the geometric concepts used in this paper in Section 2. In Section 3, we derive the convergence rate for the α-tree estimator under nonsingular measures. In Section 4, we study the behavior of the KDE and the stability of the estimated α-tree under singular measures. In Section 5, we investigate critical points of singular measures and derive the topological consistency of the estimated α-tree. We summarize this paper and discuss possible future directions in Section 6. We leave all proofs in the supplementary materials [Chen (2019)].

2. Backgrounds.

2.1. *Cluster trees.* Here we recall the definition of cluster trees in Chen et al. (2016). 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 function f is defined as follows.

DEFINITION 1 [Definition 1 in Chen et al. (2016)]. For any $f : \mathbb{K} \mapsto [0, \infty)$, we define $T_f : \mathbb{R} \mapsto 2^{2^{\mathbb{K}}}$, where $2^{\mathbb{K}}$ denotes the set of all subsets of $\mathbb{K}$, $2^{2^{\mathbb{K}}}$ denotes the collection of all sets of subsets of $\mathbb{K}$, and $T_f(\lambda)$ is the set of 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 $\mathbb{K}$. We called $\{T_f\}$ the cluster tree of f.

Clearly, the cluster tree $\{T_f\}$ has a tree structure, because for each 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 a 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 relation with respect to clusters outside an edge is equivalent. So when the collection of connected components is divided into edges, we observe the same inclusion relation 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, that is,

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

We define the equivalence relation $\sim$ such that $A \sim B$ if and only if

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

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, which 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 function 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 $A \subset B$ for all $A \in \mathbb{C}_1$ and $B \in \mathbb{C}_2$. Then the topology of the cluster tree (the shape of the cluster tree) is completely determined by the edge set $E(T_f)$ and the partial order among the edge set. Figure 3 provides an example of the connected components, edges and 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.

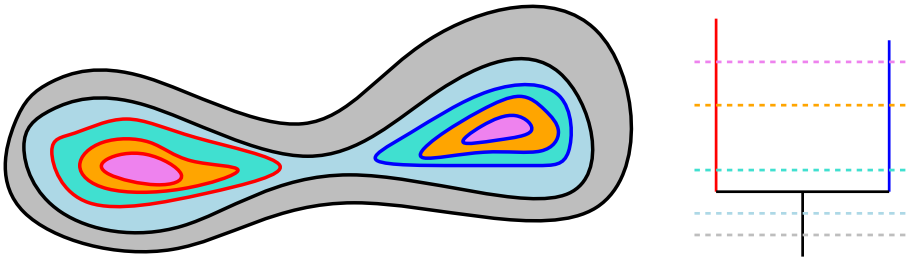

FIG. 3. *Connected components, edges, and edge set of a cluster tree. Left: We display connected components of level sets under five different levels (indicated by the colors: magenta, yellow, sea green, sky blue and gray). The color of boundaries of each connected component denotes the edge they correspond to. Right: The cluster tree. We color the three edges (vertical lines) red, blue and black. The edge set* $E(T_f) = \{\mathbb{C}_{\text{red}}, \mathbb{C}_{\text{blue}}, \mathbb{C}_{\text{black}}\}$ *and we have the ordering* $\mathbb{C}_{\text{red}} \le \mathbb{C}_{\text{black}}$ *and* $\mathbb{C}_{\text{blue}} \le \mathbb{C}_{\text{black}}$. *Note that the solid black horizontal line is not an edge set; it is a visual representation of connecting the blue and red edges to the black edge. The horizontal dashed lines indicate the five levels corresponding to the left panel. In the left panel, the three connected components with red boundaries are elements of the edge* $\mathbb{C}_{\text{red}}$.

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

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

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 function f as

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

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

In most of the cluster tree literature [Balakrishnan et al. (2012), Chaudhuri and Dasgupta (2010), Chaudhuri et al. (2014), Chen et al. (2016), Eldridge, Belkin and Wang (2015)], the cluster tree is referred to as the λ-tree, which uses 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)] that uses the function

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

to build the cluster tree T_α (T_α is called the *α-tree*). The function $\alpha(x)$ is also called *density ranking* in Chen and Dobra (2017). The set $L_\lambda = \{x : p(x) \ge \lambda\}$ 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]. A feature of the 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 of points within $\mathbb{K}$ as the density function p can be used to construct the function $\alpha(x)$. Specifically, we write

$$\begin{aligned} x_1 \succ_p x_2 \quad &\Leftrightarrow \quad p(x_1) > p(x_2), \\ x_1 \prec_p x_2 \quad &\Leftrightarrow \quad p(x_1) < p(x_2), \\ x_1 \simeq_p x_2 \quad &\Leftrightarrow \quad p(x_1) = p(x_2). \end{aligned}$$

Then

$$\alpha(x) = P\big(\{y : p(y) \leq p(x)\}\big) = P\big(\{y : y \preceq_p x\}\big), \tag{3}$$

where $y \preceq_p x$ means either $y \prec_p x$ or $y \simeq_p x$. Note that if we replace the function $p(x)$ by $2p(x)$ or $\log p(x)$, the ordering remains the same (i.e., $x \succ_p y \Leftrightarrow x \succ_{2p} y \Leftrightarrow x \succ_{\log p} y$). We will use this feature later to generalize equation (2) to singular measures.

One feature of the α-tree is that it is topological equivalent to the λ-tree.

LEMMA 1. *Assume the distribution P has a bounded density function p and p is a Morse function with a compact support. Then the λ-tree and α-tree are topological equivalent. Namely,*

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

The proof of this lemma follows from the argument at the beginning of Section 4.1 of Cadre, Pelletier and Pudlo (2009) so we ignore the proof. The main idea is that by equation (2) and the fact that p is Morse, α is a strictly monotonic transformation of the density p so the topology is preserved.

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

$$\mathbb{A}_\varpi = \{x : \alpha(x) \geq \varpi\}$$

is called an *α-level set*.

REMARK 1 (κ-tree). Kent (2013) also proposed another cluster tree—the κ-tree—which uses 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 the κ-tree. For simplicity, we only study the theory of the α-tree in this paper.

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

A key feature for constructing the α-tree is the ordering. Here we will use a generalized density function, the Hausdorff density [Mattila (1995), Preiss (1987)], to define the α-tree under singular measures. Given a probability measure P, the s-density (s dimensional Hausdorff density) is

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

provided the limit exists. Note that $B(x,r) = \{y : \|y - x\| \le r\}$ and C_s is the s-dimensional volume of an s-dimensional unit ball for $s \ge 1$ and $C_0 = 1$.

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

$$\tau(x) = \max\{s \le d : \mathcal{H}_s(x) < \infty\},$$
$$\rho(x) = \mathcal{H}_{\tau(x)}(x).$$

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 such that $x_1 \succ_{\tau,\rho} x_2$ if

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

That is, for any pair of points, we first compare their dimensions $\tau(x)$. The point with the lower dimensional value τ will be ranked higher than the other point. If two points have the same dimension, then we compare their corresponding Hausdorff density. When the distribution is nonsingular, $\tau(x) = d$ for every $x \in \mathbb{K}$ and $\rho(x) = p(x)$ is the usual density function. Thus, the ordering is determined by the density function $p(x)$.

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

$$\alpha(x) = P\big(\{y : y \preceq_{\tau,\rho} x\}\big). \tag{4}$$

Namely, $\alpha(x)$ is the probability content of regions of points whose ordering is lower than or equal to x. As shown in equation (3), equation (4) is the same as equation (2) when P is nonsingular. Note that by equation (4), the α-level set $\mathbb{A}_\varpi = \{x : \alpha(x) \ge \varpi\}$ is well defined in a singular measure.

2.3. *Geometric concepts*. We first define some notation for sets. For a set A, define $\overline{A}$ to be the closure of A, $\mathring{A}$ to be the interior of A, ∂A to be the boundary of set A, and $A^C = \mathbb{K} \backslash A$ to be the complement of set A restricted to the support $\mathbb{K}$. When A is a manifold, ∂A and $\mathring{A}$ will be the boundary and interior of the manifold,

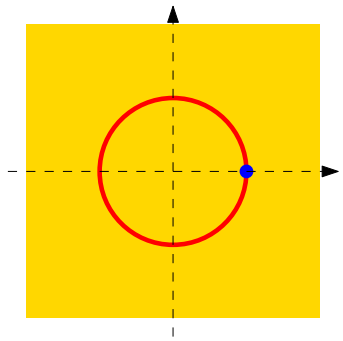

FIG. 4. *The supports* $\mathbb{K}_2$, $\mathbb{K}_1$, *and* $\mathbb{K}_0$ *for the bivariate random variable in Example* 1. *The yellow region is* $\mathbb{K}_2 = [-1,1]^2 \backslash R_{0.5}$, *where* $R_{0.5} = \{(x,y) : x^2 + y^2 = 0.5^2\}$. *The red ring is* $\mathbb{K}_1 = R_{0.5} \backslash \{(0.5,0)\}$. *The blue dot is the location of* $(0.5,0)$.

respectively. We define $A \triangle B = (A \backslash B) \cup (B \backslash A)$ to be the symmetric difference for sets A and B

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

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

where $\mathbb{K}_s = \{x : \tau(x) = s\}$. Thus, $\{\mathbb{K}_0, \ldots, \mathbb{K}_d\}$ forms a partition of the entire support $\mathbb{K}$. We call each $\mathbb{K}_s$ an s-dimensional support. 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.

EXAMPLE 1. Let X be a bivariate random variable such that (i) with a probability of 0.7, it is from a uniform distribution on $[-1,1]^2$, (ii) with a probability of 0.25, it is from a uniform distribution on the ring $R_{0.5} = \{(x,y) : x^2 + y^2 = 0.5^2\}$, (iii) with a probability of 0.05, it is equal to $(0.5,0)$. Apparently, the distribution of X is singular and the support $\mathbb{K} = [-1,1]^2$. In this case, $\mathbb{K}_2 = [-1,1]^2 \backslash R_{0.5}$, $\mathbb{K}_1 = R_{0.5} \backslash \{(0.5,0)\}$, and $\mathbb{K}_0 = \{(0.5,0)\}$. Figure 4 shows these supports under different colors. The yellow rectangular region is $\mathbb{K}_2$, the red ring area is $\mathbb{K}_1$ and the blue dot denotes the location of $\mathbb{K}_0$.

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 of a smooth manifold can be found in Lee (2013), Tu (2008)].

For an s-dimensional smooth manifold $\mathcal{M}$, the tangent space on each point of $\mathcal{M}$ changes smoothly [Tu (2008), Lee (2013)]. Namely, for $x \in \mathcal{M}$, we can find an orthonormal basis $\{v_1(x), \ldots, v_s(x) : v_\ell(x) \in \mathbb{R}^d, \ell = 1, \ldots, s\}$ such that $\{v_1(x), \ldots, v_s(x)\}$ spans the tangent space of $\mathcal{M}$ at 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, $N_s(x)$ as the normal space of $\mathbb{K}_s$ at x, and $\nabla_{T_s(x)}$ to be taking the 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), Morsel (1925, 1930)] if all its critical points are nondegenerate. Namely, the eigenvalues of the Hessian matrix of f at each critical point are nonzero. When the function f is a Morse function, its λ-tree is stable [Chazal et al. (2017), Chen et al. (2016)] in the L_∞ metric under a small perturbation of f.

To link the concept of the 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 $\rho_s^\dagger(x) = \lim_{x_n \in \mathbb{K}_s : x_n \to x} \rho(x_n)$ provided the limit exists and does not depend on the choice of the sequence 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 $s = 1, \ldots, d$. Later we will show that this generalization leads to a stable α-tree for a singular measure.

For $\overline{\mathbb{K}_s}$, let $\mathcal{C}_s = \{x \in \overline{\mathbb{K}_s} : \nabla_{T_s(x)} \rho_s^\dagger(x) = 0\}$ be the collection of its critical points. Then the fact that $\rho_s^\dagger(x)$ is a Morse function implies that the eigenvalues of the Hessian matrix $\nabla_{T_s(c)} \nabla_{T_s(c)} \rho_s^\dagger(c)$ are nonzero for every $c \in \mathcal{C}_s$. We call $g_s(x) = \nabla_{T_s(x)} \rho_s^\dagger(x)$ 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 collection $\mathcal{C} = \bigcup_{s=1,\ldots,d} \mathcal{C}_s$ is called the collection of *generalized critical points* of $\rho(x)$. Each element $c \in \mathcal{C}$ is called a generalized critical point.

Finally, we introduce the concept of *reach* [Chen, Genovese and Wasserman (2017), Federer (1959)] for a smooth manifold $\mathcal{M}$. The reach of $\mathcal{M}$ is defined as

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

where $A \oplus r = \{x : d(x, A) \leq r\}$. One can view reach as the radius of the largest ball that can roll freely outside $\mathcal{M}$. More details about reach can be found in Chen, Genovese and Wasserman (2017), Federer (1959). Reach plays a key role in the stability of a level set; see Chen, Genovese and Wasserman (2017) for more details.

REMARK 2. If we further assume that the supports satisfy

$$\mathbb{K} \supset \overline{\mathbb{K}}_d \supset \overline{\mathbb{K}}_{d-1} \supset \cdots \supset \overline{\mathbb{K}}_0, \tag{6}$$

then the support $\mathbb{K}$ forms a stratified space [Goresky and MacPherson (1980, 1988)]. Roughly speaking, a stratified space is a topological space $\mathbb{W}$ such that there exists a decomposition (called stratification) $\mathbb{W}_0, \ldots, \mathbb{W}_d$ of $\mathbb{W}$ with the properties that (i) each $\mathbb{W}_k$ is a k-dimensional smooth manifold, (ii) $\mathbb{W} = \bigcup_{\omega=0}^d \mathbb{W}_\omega$, and (iii) for any $k \leq \ell$,

$$\mathbb{W}_k \cap \overline{\mathbb{W}}_\ell \neq \varnothing \quad \Leftrightarrow \quad \mathbb{W}_k \subset \overline{\mathbb{W}}_\ell.$$

From the properties of a stratified space, one can see how equation (6) is related to a stratified space. Note that for a stratified space, if we consider a probability measure that is a mixture of probability measures defined on each stratum ($\mathbb{W}_k$), this defines a singular measure as the one being considered in this paper. The topology of a stratified space can be defined using the intersection homology [Edelsbrunner and Harer (2008), Friedman (2014), Goresky and MacPherson (1980)]. The notion of intersection homology and stratified space will be particularly useful if we want to work on higher-order homology groups.

2.4. *Estimating the α function and the α-tree.* In this paper, we focus on estimating α-trees via the 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 $\widehat{\alpha}_n$:

$$\widehat{\alpha}_n(x) = \widehat{P}_n\big(\{y : \widehat{p}_n(y) \le \widehat{p}_n(x)\}\big), \tag{7}$$

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

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

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

The L_∞ error has been used in several cluster tree literature; see, for example, Chen et al. (2016), Eldridge, Belkin and Wang (2015). An appealing feature of the 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, Belkin and Wang (2015)]. Convergence in the merge distortion metric implies the Hartigan consistency [Eldridge, Belkin and Wang (2015)], a notion of consistency of a cluster tree estimator described in Chaudhuri and Dasgupta (2010), Chaudhuri et al. (2014), Hartigan (1981). Thus, because of the equivalence between the 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 [Scott (2015), Wasserman (2006)]. The *integrated error* is

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

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

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

which 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 affect on these errors much.

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

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

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 notation. For a smooth function p, we define $\|p\|_{\ell,\infty}$ as the supremum maximal norm of ℓth derivative of p. For instance,

$$\begin{aligned}
\|p\|_{0,\infty} &= \sup_{x \in \mathbb{K}} p(x), \\
\|p\|_{1,\infty} &= \sup_{x \in \mathbb{K}} \|g(x)\|_{\max}, \\
\|p\|_{2,\infty} &= \sup_{x \in \mathbb{K}} \|H(x)\|_{\max},
\end{aligned}$$

where $g(x) = \nabla p(x)$ and $H(x) = \nabla\nabla p(x)$ are the gradient and Hessian matrix of $p(x)$, respectively. A vector $\beta = (\beta_1, \ldots, \beta_d)$ of nonnegative integers is called a multiindex with $|\beta| = \beta_1 + \beta_2 + \cdots + \beta_d$ and the corresponding derivative operator is

$$D^\beta = \frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}} \cdots \frac{\partial^{\beta_d}}{\partial x_d^{\beta_d}}, \tag{8}$$

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

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 and is four times differentiable with $\|p\|_{\ell,\infty} < \infty$ for $\ell = 0, \ldots, 4$.

(K1) $K(x)$ has compact support and is nonincreasing on $[0, 1]$, and has at least fourth-order bounded derivative and

$$\int \|x\|^2 K^{(\beta)}(\|x\|) \, dx < \infty, \qquad \int \left(K^{(\beta)}(\|x\|)\right)^2 dx < \infty$$

for $|\beta| \le 2$ and $K^{(2)}(0) < 0$.

(K2) Let

$$\mathcal{K}_r = \left\{ y \mapsto K^{(\beta)}\left(\frac{\|x-y\|}{h}\right) : x \in \mathbb{R}^d, |\beta| = r, \bar{h} > h > 0 \right\},$$

where $K^{(\beta)}$ is defined in equation (8) and $\mathcal{K}_l^* = \bigcup_{r=0}^{l} \mathcal{K}_r$ and $\bar{h}$ is some positive number. We assume that $\mathcal{K}_2^*$ is a VC-type class, that is, there exist constants A, v and a constant envelope b_0 such that

$$\sup_Q N\big(\mathcal{K}_2^*, \mathcal{L}^2(Q), b_0\varepsilon\big) \le \left(\frac{A}{\varepsilon}\right)^v, \tag{9}$$

where $N(T, d_T, \varepsilon)$ is the ε-covering number for a 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 that critical points are well separated and will not move too far away under a small perturbation on the gradient and Hessian of the density function [Chazal et al. (2017), Chen et al. (2016)]. We need the fourth-order derivative to ensure the estimated density Hessian matrix converges to the population density Hessian matrix (the bias in estimating the Hessian matrix depends on fourth-order derivatives). Assumption (K1) is a standard condition on kernel function [Scott (2015), Wasserman (2006)]. Assumption (K2) regularizes the complexity of kernel functions so we have uniform bounds on density, gradient, and Hessian estimation. It was first proposed by Giné and Guillou (2002) and Einmahl and Mason (2005) and was later used in various studies such as Chen, Genovese and Wasserman (2015), Genovese et al. (2009, 2014).

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

THEOREM 2. *Assume* (P1), (K1)–(K2). *Then, when* $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$,

$$\|\widehat{p}_n - p\|_\infty = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right),$$

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

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

$$P(T_{\widehat{p}_n} \overset{T}{\approx} T_p) \ge 1 - c_0 \cdot e^{-c_1 \cdot nh^{d+4}},$$

for some $c_0, c_1 > 0$.

The rate of consistency under the L_∞ error can be found in Chen, Genovese and Wasserman (2017), Einmahl and Mason (2005), Giné and Guillou (2002). The integrated error and probability error can be seen in Scott (2015). And the topological error bound follows Lemma 2 in Chen et al. (2016) and the concentration of L_∞ metric for the estimated Hessian matrix.

The requirement of h in Theorem 2 enforces the uniform convergence of the KDE as well as its first and second derivative. Uniform convergence of derivatives of the KDE implies the convergence of some geometric structures of the density function, such as the ridges [Chen, Genovese and Wasserman (2015), Genovese et al. (2014)], critical points [Chazal et al. (2017), Chen, Genovese and Wasserman (2016)] and persistent diagrams [Chen (2017), Cohen-Steiner, Edelsbrunner and Harer (2007), Fasy et al. (2014)].

Now we turn to the consistency for α-tree. To derive the rate for the α-tree, we need to study the convergence rate of an estimated 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)–(K2) *and* $d \geq 2$. *Let* λ *be a density level corresponding to the density of a critical point. When* $h \to 0$, $\frac{\log n}{nh^{d+4}} \to 0$,

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

The rate in Theorem 3 is slower than the usual density estimation rate. This is because the boundary of L_λ hits a critical point when λ equals the density of a critical point. The regions around a critical point have a very low gradient, which leads to a slower convergence rate. It is well known [Einmahl and Mason (2005), Genovese et al. (2014), Giné and Guillou (2002)] that under assumption (P) and (K1)–(K2),

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

In Theorem 3, we see that when d is large, the quantity $\frac{d}{d+1} \to 1$ so the error rate is similar to $\|\widehat{p}_n - p\|_\infty$. This is because the regions that slow down the error rate are areas around the critical points. These areas occupy a small volume when d is large, which decreases the difference in the rate.

REMARK 3. Theorem 3 complements many existing level set estimation theories. To our knowledge, no literature has worked on the situation where λ equals

the density of a critical point. Level set theories mostly focus on one of the following three cases: (i) the gradient on the level set boundary $\partial L_\lambda = \{x : p(x) = \lambda\}$ is bounded away from 0 [Cadre (2006), Chen, Genovese and Wasserman (2017), Laloe and Servien (2013), Mammen and Polonik (2013), Molchanov (1990), Tsybakov (1997), Walther (1997)], (ii) a lower bound on the density changing rate around level λ [Singh, Scott and Nowak (2009), Rinaldo et al. (2012)], (iii) an (ε, σ) condition for density [Chaudhuri and Dasgupta (2010), Chaudhuri et al. (2014)]. When λ equals a critical level, none of these assumptions hold.

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

THEOREM 4. *Assume* (P1) *and* (K1)–(K2) *and* $d \geq 2$ *and the smoothing bandwidth satisfies* $h \to 0$, $\frac{\log n}{nh^{d+4}} \to 0$. *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,

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

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

COROLLARY 5. *Assume* (P1) *and* (K1)–(K2) *and* $d \geq 2$. *Then, when* $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$,

$$\|\widehat{\alpha}_n - \alpha\|_\infty = O(h^{\frac{2d}{d+1}}) + O_P\left(\left(\frac{\log n}{nh^d}\right)^{\frac{d}{2(d+1)}}\right),$$

$$\|\widehat{\alpha}_n - \alpha\|_\mu = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right),$$

$$\|\widehat{\alpha}_n - \alpha\|_P = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right),$$

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

for some $c_0, c_1 > 0$.

Comparing Corollary 5 to Theorem 2, we see that only the L_∞ error rate has a major difference and the other two errors differ by a $\sqrt{\log n}$ factor. This is because Theorem 4 proves that, only at the level of a critical point, we will have a slower convergence rate. Thus, the L_∞ error will be slowed down by these points. However, the collection of points $\{x : p(x) = p(c) \text{ for some } c \in \mathcal{C}\}$ has Lebesgue measure zero so the slow convergence rate does not translate to the integrated error and the 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 error rates under singular measures. When a measure is singular, the usual (Radon–Nikodym) density cannot be defined. Thus, we cannot define the λ-tree. However, as we discussed in Section 4, we are still able to define the α-tree. Thus, in this section, we will focus on error rates for the α-tree.

4.1. *Analysis of the KDE under singular measures.* To study the convergence rate, 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.

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

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

Assumption (S) ensures that $\mathbb{K}_s$ is smooth and every connected component of $\mathbb{K}_s$ is separated for each s. Assumption (P2) is a generalization of (P1) to singular distributions.

LEMMA 6 (Bias of the smoothed density). *Assume* (S), (P2). *Let* $x \in \mathring{\mathbb{K}}_s$ *and define* $m(x) = \min\{\ell > s : x \in \overline{\mathbb{K}_\ell}\} - s$. *Let* $C_\ell^\dagger = (\int_{B_\ell} K(\|y\|)\,dy)^{-1}$, *where* $B_\ell = \{y : \|y\| \le 1, y_{\ell+1} = y_{\ell+2} = \cdots = y_d = 0\}$ *for* $\ell = 1, \ldots, d$ *and* dy *is integrating with respect to* ℓ*-dimensional area and* $C_0^\dagger = 1/K(0)$. *Then for a fixed* x, *when* $h \to 0$ *and* $m(x) > 0$,

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

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

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

Lemma 6 describes the bias of the KDE. The scaling factor $C^{\dagger}_{\tau(x)}h^{d-\tau(x)}$ rescales 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: $O(h^2)$, the usual smoothing bias, and $O(h^{m(x)})$, the bias from a higher dimensional support. This is because the KDE is isotropic, so the probability content outside $\mathbb{K}_s$ will also be included, which causes additional bias. The second assertion states that the smoothed density does not uniformly converge to the generalized density $\rho(x)$, so together with the first assertion, we conclude that the smoothing bias converges pointwisely but not uniformly. Next, we provide an example showing the failure of uniform convergence of a singular measure.

EXAMPLE 2 (Failure of uniform convergence). We consider X from the same distribution as in Figure 2: with a probability of 0.3, $X = 2$, and with a probability of 0.7, X follows a standard normal. For simplicity, we assume that the kernel function is the spherical kernel $K(x) = \frac{1}{2}I(0 \le x \le 1)$ and consider the smoothing bandwidth $h \to 0$. This choice of kernel yields $C^{\dagger}_1 = 1$. Now consider a sequence of points $x_h = 2 + \frac{h}{2}$. Then the smoothed density at each x_h is

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

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 Lemma 6 $C^{\dagger}_{\tau(x)}h^{d-\tau(x)}$ depends on the support $\mathbb{K}_s$ where x resides. 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 Lemma 6. To see this, let x_1 and x_2 be the two points that we want to compare their orderings (i.e., we want to know $x_1 \prec_{\tau,\rho} x_2$ or $x_1 \succ_{\tau,\rho} x_2$ or $x_1 \simeq_{\tau,\rho} x_2$). 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 that 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 $s_1 \ne s_2$), $p_h(x_1)$ and $p_h(x_2)$ diverge at different rates, meaning that we can eventually distinguish them. Thus, ordering for most points can still be recovered under singular measures. 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 measures). *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 converge to* 0. *Namely, there exists* $\varepsilon > 0$ *such that*

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

The proof of Corollary 7 is a direct application of the failure of uniform convergence in smoothing bias shown in Lemma 6. This corollary shows that for a singular measure, the L_∞ error of the estimator $\widehat{\alpha}_n$ does not converge 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 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 the scaled KDE converges uniformly.

Let $\mathbb{K}_s(h) = \mathring{\mathbb{K}}_s \backslash (\bigcup_{\ell<s} \mathbb{K}_\ell \oplus h)$ be the set that is in the interior of $\mathbb{K}_s$ and is away from lower dimensional supports for $s > 0$; in the case of $s = 0$, we define $\mathbb{K}_0(h) = \mathbb{K}_0$. We define $\mathbb{K}(h) = \bigcup_{s\le d} \mathbb{K}_s(h)$, which is the union of each $\mathbb{K}_s(h)$. Figure 5 shows the good region of support $\mathbb{K}_s(h)$ and the original support $\mathbb{K}_s$ in Example 1. Later we will show that the set $\mathbb{K}(h)$ is the good region.

In Lemma 6, the quantity

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

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

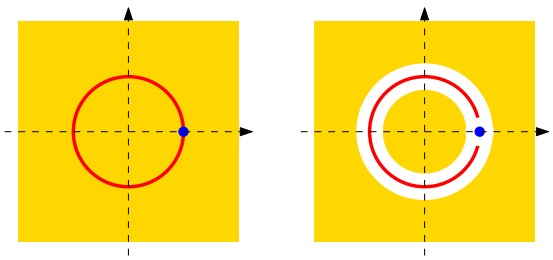

FIG. 5. *Good regions* $\mathbb{K}_s(h)$ *for Example 1. Left: the original* $\mathbb{K}_2, \mathbb{K}_1$, *and* $\mathbb{K}_0$. *Right: the corresponding* $\mathbb{K}_2(h), \mathbb{K}_1(h)$, *and* $\mathbb{K}_0(h)$. *The yellow area is the set* $\mathbb{K}_2(h)$, *which are regions in* $\mathbb{K}_2$ *away from lower-dimensional supports* $\mathbb{K}_1$ *and* $\mathbb{K}_0$. *The red area is the set* $\mathbb{K}_1(h)$, *which are regions in* $\mathbb{K}_1$ *where regions close to* $\mathbb{K}_0$ *have been removed. The blue dot is* $\mathbb{K}_0(h)$, *which is the same as* $\mathbb{K}_0$ *because there is no lower dimensional support. Note that* $\mathbb{K}(h)$ *is the union of the three color regions in the right panel.*

we introduce the quantity

$$m_{\min} = \inf_{x\in\mathbb{K}, m(x)>0} m(x). \tag{10}$$

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

THEOREM 8 (Consistency of the KDE under singular measures). *Assume* (S), (P2), (K1)–(K2). *Let* $C_\ell^\dagger$ *be the constant in Lemma* 6. *Define*

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

where $\nabla_{T_s(x)}$ *is taking gradient with respect to the tangent space of* $\mathbb{K}_s$ *at* x. *Then, when* $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$,

$$\begin{aligned}
\delta_{n,h,s} &= O\big(h^{2\wedge m_{\min}}\big) + O_P\left(\sqrt{\frac{\log n}{nh^s}}\right), \\
\delta_{n,h,s}^{(1)} &= O\big(h^{2\wedge m_{\min}}\big) + O_P\left(\sqrt{\frac{\log n}{nh^{s+2}}}\right), \\
\delta_{n,h,s}^{(2)} &= O\big(h^{2\wedge m_{\min}}\big) + O_P\left(\sqrt{\frac{\log n}{nh^{s+4}}}\right),
\end{aligned} \tag{11}$$

where $a \wedge b = \min\{a, b\}$.

Theorem 8 shows that after rescaling, the KDE is uniformly consistent within the good region $\mathbb{K}_s(h)$ for 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 a higher dimensional support).

REMARK 5 (Nonconvergence of the integrated distance of the KDE). One may wonder if the scaled KDE $[C_{\tau(x)}^\dagger h^{d-\tau(x)} \cdot \widehat{p}_n(x)]$ converges to the generalized density $\rho(x)$ under the integrated distance. There is no guarantee for such a convergence because

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

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 $y \in B(x,h) \cap \mathbb{K}_\ell$, $\tau(y) = \ell$ but the KDE $\widehat{p}_n(y)$ is at rate $O_P(h^{s-d})$. Thus, the difference between the scaled KDE and the generalized density is

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

Such y has ℓ-dimensional volume at rate $O(h^{\ell-s})$, so the integrated error is at rate $O_P(h^{s-\ell}) \times O(h^{\ell-s}) = O_P(1)$.

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

THEOREM 9 (Nearly uniform consistency of α-trees). *Assume* (S), (P2), (K1)–(K2). *Let $\mathcal{C}_s$ be the collection of generalized critical points of $\mathbb{K}_s$. Let $\delta_{n,h,s}$ be defined in equation* (11) *and $r_{n,h,s}$ be a quantity such that $\frac{\delta_{n,h,s}}{r_{n,h,s}} = o_P(1)$. Then, when $h \to 0$, $\frac{nh^{d+2}}{\log n} \to \infty$, uniformly for every $x \in \mathbb{K}_s(h)$,*

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

The convergence rate in Theorem 9 is similar to that in Theorem 8: for a given point x when $\alpha(x)$ is away from the α value of a generalized critical point (a critical α level), we have the usual convergence rate. When $\alpha(x)$ is close to a critical α level, the convergence rate is slower. The quantity $r_{n,h,s}$ behaves like the quantity ϖ_n in Theorem 4, which was introduced to guarantee the uniform convergence. To derive the consistency of $\widehat{\alpha}_n$ under the integrated error and the probability error, we 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)–(K2). *Then*

$$\|\widehat{\alpha}_n - \alpha\|_P = O(\delta_{n,h,d}),$$
$$\|\widehat{\alpha}_n - \alpha\|_\mu = O(\delta_{n,h,d}).$$

Theorem 9 shows that the quantity $\alpha(x)$ can be consistently estimated by $\widehat{\alpha}_n(x)$ for the majority points. This implies that the ordering of points using $\widehat{p}_n$ is consistent with the ordering from τ, ρ in most areas of $\mathbb{K}$.

5. Singular measures: Critical points and topology. Recall 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 $\mathbb{C} \in E(T_\alpha)$. The set $\mathcal{A}_\alpha$ (critical tree-levels) contains the levels where the upper level set $\mathbb{A}_\varpi = \{x : \alpha(x) \geq \varpi\}$ changes its shape. For nonsingular measures, $\mathcal{A}_\alpha$ corresponds to the density value of some critical points.

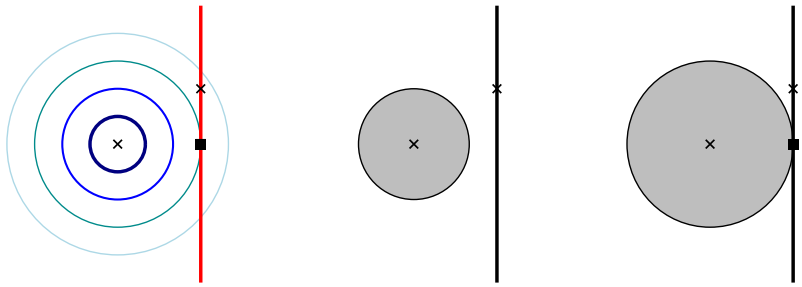

FIG. 6. *Example of DCPs. This is a $d = 2$ case; there is a $2D$ spherical distribution mixed with a $1D$ distribution on a line segment. Left: the blue contours are density contours of the $2D$ spherical distribution and the red line segment is $\mathbb{K}_1$, the support of the $1D$ singular distribution. 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 DCPs merge two connected components, we consider the middle and the right panel, which are level sets of $\alpha(x)$ at two different levels. Middle: the level set $\mathbb{A}_{\varpi}$ where the level ϖ 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 merge so there is only one connected component. The merging point is the square point, which is defined as a DCP.*

For singular measures, this is not true even when $\rho(x)$ is a generalized Morse function.

Consider the example in Figure 6. The solid box in the left panel indicates a new type of critical points, where merges between elements in different edge sets occur (change in the topology of level sets occurs). By the definition of $\mathcal{A}_\alpha$, this corresponds to an element in $\mathcal{A}_\alpha$, but it is clearly not a generalized critical point. We call this type of critical points the *dimensional critical points (DCPs)*. In Figure 6, the dimension $d = 2$ and we have a $2D$ spherical distribution mixed with a $1D$ singular measure that is distributed on the red curves $\mathbb{K}_1$. 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 merge at the location of the solid box. Although the location of the solid box does not belong to the collection of generalized critical points $\mathcal{C}$, this point does correspond to the merging of connected components in the level sets. Thus, this point corresponds to an element in $\mathcal{A}_\alpha$.

Here is the formal definition of the DCP. Recall that $\mathcal{C}$ is the collection of generalized critical points of $\rho(x)$ and from equation (1), $\mathcal{A}_\alpha$ is the collection of levels of function $\alpha(x)$ such that the creation of a new connected component or merging of connected components occurs. For simplicity, we denote $\mathcal{A} = \mathcal{A}_\alpha$. For any level

$\varpi \in \mathcal{A}$, we define

$$\xi(\varpi) = \begin{cases} \max\{w \geq 0 : \mathbb{K}_w \subset \mathbb{A}_\varpi\} & \text{if } \mathbb{K}_0 \subset \mathbb{A}_\varpi, \\ -1 & \text{if } \mathbb{K}_0 \not\subset \mathbb{A}_\varpi. \end{cases} \tag{12}$$

Namely, $\xi(\varpi)$ is the highest dimensional structure that is covered by the level set $\mathbb{A}_\varpi$.

DEFINITION 3. For $\varpi \in \mathcal{A}$, we say x is a DCP if the following hold:

(1) $x \in \mathbb{K}_\ell$ for some $\ell \leq \xi(\varpi)$.
(2) There exists an edge $\mathbb{C} \in E(T_\alpha)$ such that:

(i) $x \notin C$ for all $C \in \mathbb{C}$,
(ii) $d(x, C_\varepsilon) \to 0$ when $\varepsilon \to 0$, where $C_\varepsilon = \mathbb{C} \cap T_\alpha(\varpi + \varepsilon)$.

Note that x may not exist. In such a case, there is no DCPs for level ϖ.

The first requirement is to ensure that x is on a lower dimensional support $[\mathbb{K}_\ell : \ell < \xi(\varpi)]$. The second requirement is to show that the DCP x is not in the same edge $\mathbb{C}$, but its 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 $\mathbb{A}_{\varpi+\varepsilon}$ for a 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 $\mathbb{C}'$, and (ii) not in any element of edge $\mathbb{C}$, and because (iii) the distance from x to the element of $\mathbb{C}$ converges to 0 when the level decreases to the level ϖ, x is a merging point of edges $\mathbb{C}'$ and $\mathbb{C}$ and the level ϖ is their merging level.

Note that since DCPs occur when different dimensional regions intersect each other, the topological structure of such an intersection may be related to the intersection homology and stratified space [Edelsbrunner and Harer (2008), Goresky and MacPherson (1980, 1988)].

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 corresponding to the DCP at c. Note that $\alpha^\dagger(c) \neq \alpha(c)$ when c is a DCP. Since a DCP is generally in a lower dimensional support than the support that the merging happens, $\alpha(c) > \alpha^\dagger(c)$. Remark 7 provides an example of how α and $\alpha^\dagger$ differs.

REMARK 6 (Relation to the usual critical points). The definition of DCPs is similar to that of saddle points or local minima, who contribute to the merging of level sets. Saddle points (or local minima) that contribute to a merging of level sets can be defined as a point x with the following properties:

(1) $x \in \mathbb{K}_{\xi(\varpi)+1}$.
(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) $d(x, C_{1,\varepsilon}) \to 0$ when $\varepsilon \to 0$, where $C_{1,\varepsilon} = \mathbb{C}_1 \cap T_\alpha(\varpi + \varepsilon)$,
(iii) $d(x, C_{2,\varepsilon}) \to 0$ when $\varepsilon \to 0$, where $C_{2,\varepsilon} = \mathbb{C}_2 \cap T_\alpha(\varpi + \varepsilon)$.

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

REMARK 7. Note that a DCP might be at the same position as a critical point. Consider the example in Figure 2 and Example 2. In this case,

$$\alpha(x) = \begin{cases} 0.7 \cdot 2 \cdot \Phi_0(-|x|) & \text{if } x \neq 2, \\ 1 & \text{if } x = 2, \end{cases}$$

where $\Phi_0(x)$ is the cumulative distribution of a standard normal. Moreover, $\mathcal{A}_\alpha = \{1, 0.7, 0.0319\}$; the first element $\{1\}$ is the level of the point mass located at $x = 2$, the second element $\{0.7\}$ is the level of the mode of the standard normal distribution, the last element is the level where the connected components created at levels 1 and 0.7 merged so it comes from a DCP. This DCP also located at $x = 2$, which coincides with a local mode. Note that the number $0.0319 = 0.7 \cdot 2 \cdot \Phi_0(-2)$, which is the level where the merging occurred. At the critical point $x = 2$, $\alpha(2) = 1$ and $\alpha^\dagger(2) = 0.0319$.

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 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 $y \in B(x, r) \cap \mathbb{K}_s$ 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 $s = 1, 2, \ldots, d$.

Assumption (A) is to ensure that no multiple topological changes will occur at the same level so each level corresponds to only a merging or a creation. Assumption (B) is to guarantee that no new connected component at the boundary of a lower dimensional manifold will be created. Thus, any creation of a new connected component of the level set $\mathbb{A}_\varpi$ occurs only at a (generalized) local mode. 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* $\mathbb{A}_{\varpi}$ *is created at* $\varpi \in \mathcal{A}$, *then there is a local mode* c *of* ρ *or an element in* $\mathbb{K}_0$ *such that* $\varpi = \alpha(c)$. *Namely, DCPs only merge connected components*.
- *For any value* $\varpi \in \mathcal{A}$, *either* $\varpi = \alpha(c)$ *for some* $c \in \mathcal{C}$ *or there is a DCP associated with* ϖ.

Lemma 11 provides two basic properties of DCPs. First, DCPs only merge connected components. Moreover, when the topology of connected components of α-level sets changes (when we decrease the level), either a critical point or a DCP must be responsible for this. Therefore, 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 (Stability of generalized critical points). *Assume* (S), (P2), (K1)–(K2), (C). *Let* $c \in \mathbb{K}_s$ *be a generalized critical point with* $n(c)$ *negative eigenvalues of its generalized Hessian matrix. Let* $\widehat{\mathcal{C}}$ *be the collection of critical points of* $\widehat{p}_n$. *Then, when* $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$, *there exists a point* $\widehat{c} \in \widehat{\mathcal{C}}$ *such that*

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

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

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* (11).

Lemma 12 is a generalization of the stability theorem of critical points given in Lemma 16 of Chazal et al. (2017). Note that the bias is now of the order $O(h)$; this is due to the smoothing effect from a higher dimensional support.

LEMMA 13 (Properties of estimated critical points). *Assume* (S), (P2), (K1)–(K2), (A), (B). *Assume there are* k *DCPs. Let* $\widehat{\mathcal{C}}$ *be the critical points of* $\widehat{p}_n$. *Define* $\widehat{\mathcal{G}} \subset \widehat{\mathcal{C}}$ *as 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* $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$:

- $\widehat{\mathcal{D}} \subset \mathbb{K}^C(h)$,
- $|\widehat{\mathcal{D}}| \geq k$, *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 of the estimated critical points (critical points of $\widehat{p}_n$). First, the estimated critical points are all in the bad region $\mathbb{K}^C(h)$, except for 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)–(K2), (A), (B). *Let c be a DCP and $\alpha^\dagger(c) \in \mathcal{A}$ be the associated level. Let $\widehat{\mathcal{D}}$ be defined as 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*

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

where $\delta_{n,h,s}$ is defined in (11). Moreover, the $\widehat{\mathbb{A}}_{\widehat{\alpha}_n(\widehat{c})+\varepsilon}$ and $\widehat{\mathbb{A}}_{\widehat{\alpha}_n(\widehat{c})}$ are not topological equivalent.

Lemma 14 illustrates the stability of critical tree-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 the estimated level sets.

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

EXAMPLE 3 (A DCP and its estimator). Consider again the example in Figure 2 and Example 2. We have a singular distribution mixed with a point mass at $x = 2$ with a probability of 0.3 and a standard normal with a probability of 0.7. In this case, as indicated, a DCP is located at $x = 2$ with level 0.0319 (see Remark 7). In every panel of the top row of Figure 2, there is a local minimum located in the region $x \in [1.5, 2]$. Moreover, when we increase the sample size (from left to right), this local minimum is moving toward $x = 2$. This local minimum is an estimated critical point $\widehat{c} \in \widehat{\mathcal{D}}$, as 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)–(K2), (A), (B), (C). *Then, when $h \to 0$, $\frac{nh^{d+4}}{\log n} \to \infty$,*

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

for some $c_0, c_1 > 0$.

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

REMARK 8. 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, recall that by Lemma 1, $T_{\widehat{p}_n} \overset{T}{\approx} T_{\widehat{\alpha}_n}$. This, together with Theorem 15, implies

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

under a 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 singular and nonsingular measures. In the nonsingular case, the error rate under the L_∞ metric is slower than other metrics because of the slow rate of level set estimation around saddle points. However, other error rates are the same as estimating the λ-tree.

When a 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 pointwise consistent estimator after rescaling;
2. the KDE is a uniformly consistent estimator after rescaling for the majority of the support; and
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 pointwise consistent estimator of α;
2. $\widehat{\alpha}_n$ is a uniformly consistent estimator for the majority of the support;
3. $\widehat{\alpha}_n$ is a consistent estimator of α under the integrated distance and probability distance; and
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 the estimated critical points from the KDE approximate these DCPs.

Finally, we point out some possible future directions.

- *Persistence homology.* The cluster tree is closely related to the persistent homology of level sets [Bobrowski, Mukherjee and Taylor (2017), Fasy et al. (2014)]. In the persistent homology, a common metric for evaluating the quality of an estimator is the bottleneck distance [Cohen-Steiner, Edelsbrunner and Harer (2007), Edelsbrunner and Morozov (2013)]. Because the bottleneck distance is bounded by the L_∞ metric [Cohen-Steiner, Edelsbrunner and Harer (2007), Edelsbrunner and Morozov (2013)], many bounds on the bottleneck distance are derived via bounding the L_∞ metric [Bobrowski, Mukherjee and Taylor (2017), Fasy et al. (2014)]. However, for α-trees under singular measures, the L_∞ metric does not converge (Corollary 7) but we do have topological consistency (Theorem 15), which implies convergence in the bottleneck distance. This provides an example where we have consistency under the bottleneck distance and inconsistency in the L_∞ metric. How this phenomenon affects the persistence homology is unclear and we leave that line of study for future work.
- *Higher-order homology groups and stratified space.* Our definition of DCPs is for connected components, which are zeroth-order homology groups [Bubenik (2015), Cohen-Steiner, Edelsbrunner and Harer (2007), Fasy et al. (2014)] and sufficient for analyzing cluster trees. 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 whether the KDE is also consistent in recovering higher-order homology groups under singular measures. Moreover, as is mentioned in Remark 2, the supports we are analyzing are related to the stratified space [Friedman (2014), Goresky and MacPherson (1980)] and the DCPs might be related to the intersection homology [Edelsbrunner and Harer (2008), Friedman (2014), Goresky and MacPherson (1980)]. The intersection homology extends the definition of homology group to a stratified space so it provides a tool to analyze higher-order homology groups in our setting. Thus, finding the connection between theories of stratified space and the higher-order homology groups in our settings will be another future research direction.
- *Minimax theory.* Chaudhuri and Dasgupta (2010), Chaudhuri et al. (2014) derived the minimax theory for estimating the λ-tree under nonsingular measures and proved 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. However, for singular measures, it is unclear how to derive the minimax theory so we plan to investigate this in the future.

Acknowledgments. We thank reviewers for their very insightful comments. We also thank members in the CMU Topstat group for useful discussion.

SUPPLEMENTARY MATERIAL

Supplementary proofs: Generalized cluster trees and singular measures (DOI: 10.1214/18-AOS1744SUPP; .pdf). This document contains all proofs to the theorems and lemmas in this paper.

REFERENCES

BALAKRISHNAN, S., NARAYANAN, S., RINALDO, A., SINGH, A. and WASSERMAN, L. (2012). Cluster trees on manifolds. In *Advances in Neural Information Processing Systems* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) **26** 2679–2687. Curran Associates, Red Hook, NY.

BARYSHNIKOV, Y., BUBENIK, P. and KAHLE, M. (2014). Min-type Morse theory for configuration spaces of hard spheres. *Int. Math. Res. Not. IMRN* **9** 2577–2592. MR3207377

BOBROWSKI, O., MUKHERJEE, S. and TAYLOR, J. E. (2017). Topological consistency via kernel estimation. *Bernoulli* **23** 288–328. MR3556774

BUBENIK, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16** 77–102. MR3317230

CADRE, B. (2006). Kernel estimation of density level sets. *J. Multivariate Anal.* **97** 999–1023. MR2256570

CADRE, B., PELLETIER, B. and PUDLO, P. (2009). Clustering by estimation of density level sets at a fixed probability. Available at https://hal.archives-ouvertes.fr/file/index/docid/397437/filename/tlevel.pdf.

CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 255–308. MR2476414

CHAUDHURI, K. and DASGUPTA, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems* 343–351.

CHAUDHURI, K., DASGUPTA, S., KPOTUFE, S. and VON LUXBURG, U. (2014). Consistent procedures for cluster tree estimation and pruning. *IEEE Trans. Inform. Theory* **60** 7900–7912. MR3285753

CHAZAL, F., FASY, B., LECCI, F., MICHEL, B., RINALDO, A. and WASSERMAN, L. (2017). Robust topological inference: Distance to a measure and kernel distance. *J. Mach. Learn. Res.* **18** Paper No. 159, 40. MR3813808

CHEN, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. ArXiv preprint. Available at arXiv:1704.03924.

CHEN, Y.-C. (2019). Supplement to "Generalized cluster trees and singular measures". DOI:10.1214/18-AOS1744SUPP.

CHEN, Y.-C. and DOBRA, A. (2017). Measuring human activity spaces with density ranking based on GPS data. ArXiv preprint. Available at arXiv:1708.05017.

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2015). Asymptotic theory for density ridges. *Ann. Statist.* **43** 1896–1928. MR3375871

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2016). A comprehensive approach to mode clustering. *Electron. J. Stat.* **10** 210–241. MR3466181

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2017). Density level sets: Asymptotics, inference, and visualization. *J. Amer. Statist. Assoc.* **112** 1684–1696. MR3750891

CHEN, Y.-C., KIM, J., BALAKRISHNAN, S., RINALDO, A. and WASSERMAN, L. (2016). Statistical inference for cluster trees. ArXiv preprint. Available at arXiv:1605.06416.

COHEN-STEINER, D., EDELSBRUNNER, H. and HARER, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120. MR2279866

EDELSBRUNNER, H. and HARER, J. (2008). Persistent homology—A survey. In *Surveys on Discrete and Computational Geometry. Contemp. Math.* **453** 257–282. Amer. Math. Soc., Providence, RI. MR2405684

EDELSBRUNNER, H. and MOROZOV, D. (2013). Persistent homology: Theory and practice. In *European Congress of Mathematics* 31–50. Eur. Math. Soc., Zürich. MR3469114

EINMAHL, U. and MASON, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380–1403. MR2195639

ELDRIDGE, J., BELKIN, M. and WANG, Y. (2015). Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. In *Proceedings of the* 28*th Conference on Learning Theory* 588–606.

FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S. and SINGH, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42** 2301–2339. MR3269981

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

FRIEDMAN, G. (2014). Singular intersection homology. Preprint.

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2009). On the path density of a gradient field. *Ann. Statist.* **37** 3236–3271. MR2549559

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2014). Nonparametric ridge estimation. *Ann. Statist.* **42** 1511–1545. MR3262459

GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré B, Probab. Stat.* **38** 907–921. MR1955344

GORESKY, M. and MACPHERSON, R. (1980). Intersection homology theory. *Topology* **19** 135–162. MR0572580

GORESKY, M. and MACPHERSON, R. (1988). *Stratified Morse Theory. Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **14**. Springer, Berlin. MR0932724

HARTIGAN, J. A. (1981). Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.* **76** 388–394. MR0624340

KENT, B. P. (2013). Level set trees for applied statistics. Ph.D. thesis, Carnegie Mellon Univ., Pittsburgh, PA.

KLEMELÃ, J. (2004). Visualization of multivariate density estimates with level set trees. *J. Comput. Graph. Statist.* **13** 599–620. MR2087717

KLEMELÃ, J. (2006). Visualization of multivariate density estimates with shape trees. *J. Comput. Graph. Statist.* **15** 372–397. MR2256150

KLEMELÃ, J. (2009). *Smoothing of Multivariate Data*: *Density Estimation and Visualization*. Wiley, Hoboken, NJ. MR2640738

KPOTUFE, S. and LUXBURG, U. V. (2011). Pruning nearest neighbor cluster trees. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*-11) (L. Getoor and T. Scheffer, eds.) 225–232. International Machine Learning Society, Madison, WI.

LALOE, T. and SERVIEN, R. (2013). Nonparametric estimation of regression level sets. *J. Korean Statist. Soc.*

LEE, J. M. (2013). *Introduction to Smooth Manifolds*, 2nd ed. *Graduate Texts in Mathematics* **218**. Springer, New York. MR2954043

MAMMEN, E. and POLONIK, W. (2013). Confidence regions for level sets. *J. Multivariate Anal.* **122** 202–214. MR3189318

MASON, D. M. and POLONIK, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.* **19** 1108–1142. MR2537201

MATTILA, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces*: *Fractals and Rectifiability. Cambridge Studies in Advanced Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR1333890

MILNOR, J. (1963). *Morse Theory. Based on Lecture Notes by M. Spivak and R. Wells. Annals of Mathematics Studies*, *No.* 51. Princeton Univ. Press, Princeton, NJ. MR0163331

MOLCHANOV, I. S. (1990). Empirical estimation of quantiles of distributions of random closed sets. *Teor. Veroyatn. Primen.* **35** 586–592. MR1091220

MORSE, M. (1925). Relations between the critical points of a real function of n independent variables. *Trans. Amer. Math. Soc.* **27** 345–396. MR1501318

MORSE, M. (1930). The foundations of a theory of the calculus of variations in the large in m-space. II. *Trans. Amer. Math. Soc.* **32** 599–631. MR1501555

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—An excess mass approach. *Ann. Statist.* **23** 855–881. MR1345204

PREISS, D. (1987). Geometry of measures in $\mathbf{R}^n$: Distribution, rectifiability, and densities. *Ann. of Math.* (2) **125** 537–643. MR0890162

RINALDO, A. and WASSERMAN, L. (2010). Generalized density clustering. *Ann. Statist.* **38** 2678–2722. MR2722453

RINALDO, A., SINGH, A., NUGENT, R. and WASSERMAN, L. (2012). Stability of density-based clustering. *J. Mach. Learn. Res.* **13** 905–948. MR2930628

SCOTT, D. W. (2015). *Multivariate Density Estimation*: *Theory*, *Practice*, *and Visualization*, 2nd ed. Wiley, Hoboken, NJ. MR3329609

SINGH, A., SCOTT, C. and NOWAK, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.* **37** 2760–2782. MR2541446

STEINWART, I. (2011). Adaptive density level set clustering. In *COLT* 703–738.

STUETZLE, W. (2003). Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification* **20** 25–47. MR1983120

TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. MR1447735

TU, L. W. (2008). *An Introduction to Manifolds*. Springer, New York. MR2356311

WALTHER, G. (1997). Granulometric smoothing. *Ann. Statist.* **25** 2273–2299. MR1604445

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. *Springer Texts in Statistics*. Springer, New York. MR2172729

WASSERMAN, L. (2018). Topological data analysis. *Ann. Rev. Stat. Appl.* **5** 501–535. MR3774757

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195
USA
E-MAIL: yenchic@uw.edu