GENERALIZED CLUSTER TREES AND SINGULAR MEASURES

BY YEN-CHI CHEN

Department of Statistics
University of Washington
February 16, 2018

In this paper, we study the $\alpha$-cluster tree ($\alpha$-tree) under both singular and nonsingular measures. The $\alpha$-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 $\alpha$-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 $\alpha$-tree fails to converge in the $L_\infty$ 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 $\alpha$-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 $f_0$ (Stuetzle, 2003; Klemelä, 2004). 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 $\lambda$-tree of a distribution (Balakrishnan et al., 2012; Chaudhuri and Dasgupta, 2010; Balakrishnan and Pati, 2012; Pati and Patr Four steps...
Chaudhuri et al., 2014; Chen et al., 2016b; 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).

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 \( \hat{\alpha}_n(x) = \hat{P}_n(\{y : \hat{p}_n(y) \leq \hat{p}_n(x)\}) \) where \( \hat{P}_n \) is the empirical measure and \( \hat{p}_n \) 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), \( 5 \times 10^5 \) (middle), and \( 5 \times 10^7 \) (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).
  2. We derived the convergence rate of \( \hat{\alpha}_n \) (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 \( \hat{\alpha}_n \) is inconsistent under the \( L_\infty \) 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 \( T_{\hat{\alpha}_n} \) 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 Hartigan (1981); Chaudhuri and Dasgupta (2010); Chaudhuri et al. (2014); Eldridge et al. (2015). The convergence rate and the minimax theory are studied in Chaudhuri and Dasgupta (2010); Balakrishman et al. (2012); Chaudhuri et al. (2014). Chen et al. (2016b) 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; Wasserman, 2016). In particular, a cluster tree contains information about the zeroth order homology groups (Cohen-Steiner et al., 2007; Fasy et al., 2014; Bobrowski et al., 2017; Bubenik, 2015). In our analysis, we generalize the Morse theory to a non-smooth and even discontinuous function. Baryshnikov et al. (2013) also generalizes the Morse
Fig 2. Example of the estimated $\lambda$-tree and $\alpha$-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 $\lambda$-tree. The bottom panel displays the estimated $\alpha$-tree. From left to right, we increase the sample size from $5 \times 10^3$, $5 \times 10^5$, to $5 \times 10^7$. Because the distribution is singular, there is no population $\lambda$-tree so when the smoothing bandwidth decreases (when the sample size increases), the estimated $\lambda$-tree is getting degenerated. On the other hand, the estimated $\alpha$-trees remain stable regardless of the smoothing bandwidth. Note that every level set in a $\lambda$-tree corresponds to a level set in the $\alpha$-tree but the value of the corresponding level (Y-axis) will be different. The function used in constructing a $\lambda$-tree may have an unbounded range when the sample size goes to infinity, whereas the function for building an $\alpha$-tree always has a range of $[0, 1]$. 
theory to a non-smooth 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 et al. (2013) 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: 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 of cluster trees and the geometric concepts used in this paper in Section 2. In Section 3, we derive the convergence rate for the \(\alpha\)-tree estimator under nonsingular measures. In Section 4, we study the behavior of the KDE and the stability of the estimated \(\alpha\)-tree under singular measures. In Section 5, we investigate critical points of singular measures and derive the topological consistency of the estimated \(\alpha\)-tree. We summarize this paper and discuss possible future directions in Section 6. We leave all proofs in the supplementary materials (Chen, 2016).

2. Backgrounds.

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

**Definition 1 (Definition 1 in Chen et al. (2016b)).** For any \(f : K \mapsto [0, \infty)\) we define \(T_f : \mathbb{R} \mapsto 2^{2^K}\), where \(2^K\) denotes the set of all subsets of \(K\), \(2^{2^K}\) denotes the collection of all sets of subsets of \(K\), and \(T_f(\lambda)\) is the set of connected components of the upper level set \(\{ x \in K : f(x) \geq \lambda \} \). 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 \(K\). 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.
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 $E(T_f) = \{C_{\text{red}}, C_{\text{blue}}, C_{\text{black}}\}$ and we have the ordering $C_{\text{red}} \leq C_{\text{black}}$ and $C_{\text{blue}} \leq C_{\text{black}}$. 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 $C_{\text{red}}$.

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.

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

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

$$\forall C \in \{T_f\} \text{ such that } C \notin [A, B] \cup [B, A],$$

$$C \subset A \iff C \subset B, \quad A \subset C \iff B \subset C.$$  

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 $C \in E(T_f)$ is called an edge, which contains many nested connected components of the cluster tree $\{T_f\}$ (i.e. if $C_1, C_2 \in C$, 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 $C_1, C_2 \in E(T_f)$ be two edges, we write
\( C_1 \leq C_2 \) if and only if \( A \subset B \) for all \( A \in C_1 \) and \( B \in C_2 \). 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.

**Definition 2.** For two functions \( f : \mathbb{K} \rightarrow [0, \infty) \) and \( g : \mathbb{K} \rightarrow [0, \infty) \), we say \( T_f \) and \( T_g \) are topological equivalent, denoted as \( T_f \approx T_g \), if there exists a bijective mapping \( S : E(T_f) \rightarrow E(T_g) \) such that for any \( C_1, C_2 \in E(T_f) \),

\[
C_1 \leq C_2 \iff S(C_1) \leq S(C_2).
\]

For each \( C \in E(T_f) \), we define

\[
U(C) = \sup \{ \lambda : \exists C \in T_f(\lambda), C \in C \}
\]

to be the maximal level of an edge \( C \). We define the critical tree-levels of function \( f \) as

\[
\mathcal{A}_f = \{ U(C) : C \in E(T_f) \}.
\]

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., 2016b; Eldridge et al., 2015), the cluster tree is referred to as the \( \lambda \)-tree, which uses the probability density function \( p \) to build a cluster tree. Namely, the \( \lambda \)-tree is \( T_p \).

In this paper, we focus on the \( \alpha \)-tree (Kent, 2013) that uses the function

\[
\alpha(x) = P(\{ y : p(y) \leq p(x) \}) = 1 - P(\{ y : p(y) > p(x) \}) = 1 - P(L_p(x))
\]

to build the cluster tree \( T_\alpha \) (\( T_\alpha \) is called the \( \alpha \)-tree). The function \( \alpha(x) \) is also called density ranking in Chen and Dobra (2017). 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). A feature of the \( \alpha \)-tree is that the function \( \alpha(x) \) depends only on the ordering of points
within $K$. Namely, any function that assigns the same ordering of points within $K$ as the density function $p$ can be used to construct the function $\alpha(x)$. Specifically, we write

\[ x_1 \succ_p x_2 \iff p(x_1) > p(x_2), \]
\[ x_1 \prec_p x_2 \iff p(x_1) < p(x_2), \]
\[ x_1 \sim_p x_2 \iff p(x_1) = p(x_2). \]

Then

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

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

One feature of the $\alpha$-tree is that it is topological equivalent to the $\lambda$-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 $\lambda$-tree and $\alpha$-tree are topological equivalent. Namely,

\[ T_p \cong T_\alpha. \]

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

When we use the $\alpha$-tree, the induced upper level set

\[ \mathcal{A}_\omega = \{x : \alpha(x) \geq \omega\} \]

is called an $\alpha$-level set.

**Remark 1 ($\kappa$-tree).** Kent (2013) also proposed another cluster tree—the $\kappa$-tree—which uses the probability content within each edge set defined by an $\alpha$-tree (or a $\lambda$-tree) to compute the function $\kappa(x)$. Because it is just a rescaling from the $\alpha$-tree, the theory of $\alpha$-tree also works for the $\kappa$-tree. For simplicity, we only study the theory of the $\alpha$-tree in this paper.
2.2. Singular Measure. When the probability measure is singular, the $\lambda$-tree is no longer well-defined because there is no density function. However, the $\alpha$-tree can still be defined.

A key feature for constructing the $\alpha$-tree is the ordering. Here we will use a generalized density function, the Hausdorff density (Preiss, 1987; Mattila, 1999), to define the $\alpha$-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 $B(x, r) = \{y : \|y - x\| \leq r\}$ and $C_s$ is the $s$-dimensional volume of an $s$-dimensional unit ball for $s \geq 1$ and $C_0 = 1$.

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

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

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

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

That is, for any pair of points, we first compare their dimensions $\tau(x)$. The point with the lower dimensional value $\tau$ 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 $\alpha$-tree, we use equation (3):

$$\alpha(x) = P\left(\{y : y \preceq_{\tau, \rho} x\}\right).$$

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 $\alpha$-level set $\mathcal{A}_\varpi = \{x : \alpha(x) \geq \varpi\}$ is well-defined in a singular measure.
Fig 4. The supports $K_2$, $K_1$, and $K_0$ for the bivariate random variable in Example 1. The yellow region is $K_2 = [-1,1]^2 \setminus R_{0.5}$, where $R_{0.5} = \{(x,y) : x^2 + y^2 = 0.5^2\}$. The red ring is $K_1 = R_{0.5} \setminus \{(0.5,0)\}$. The blue dot is the location of $(0.5,0)$.

2.3. Geometric Concepts. We first define some notations for sets. For a set $A$, define $\overline{A}$ to be the closure of $A$, $\mathring{A}$ to be the relative interior of $A$, $\partial A$ to be the boundary of set $A$, and $A^C = \mathbb{K} \setminus A$ to be the complement of set $A$ restricted to the support $\mathbb{K}$. We define $A \triangle B = (A \setminus B) \cup (B \setminus A)$ to be the symmetric difference for sets $A$ and $B$

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

$$\mathbb{K} = K_d \bigcup K_{d-1} \bigcup \cdots \bigcup K_0,$$

where $K_s = \{x : \tau(x) = s\}$. Thus, $\{K_0, \cdots, K_d\}$ forms a partition of the entire support $\mathbb{K}$. We call each $K_s$ an $s$-dimensional support. When we analyze the support $K_s$, any $K_{s'}$ with $s' > s$ is called a higher dimensional support (with respect to $K_s$) 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, $K_2 = [-1,1]^2 \setminus R_{0.5}$, $K_1 = R_{0.5} \setminus \{(0.5,0)\}$, and $K_0 = \{(0.5,0)\}$. Figure 4 shows these supports under different colors. The yellow rectangular region is $K_2$, the red ring area is $K_1$, and the blue dot denotes the location of $K_0$.

To regularize the behavior of $\rho(x)$ on each support $K_s$, we assume that the
generalized cluster trees, is an $s$-dimensional smooth manifold (properties of a smooth manifold can be found in Lee 2012; Tu 2010).

For an $s$-dimensional smooth manifold $M$, the tangent space on each point of $M$ changes smoothly (Tu, 2010; Lee, 2012). Namely, for $x \in M$, 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 $\{v_1(x), \cdots, v_s(x)\}$ spans the tangent space of $M$ at $x$ and each $v_\ell(x)$ is a smooth (multivalued) function on $M$. For simplicity, for $x \in K_s$, we denote $T_x(x)$ as the tangent space of $K_s$ at $x$, $N_x(x)$ as the normal space of $K_s$ at $x$, and $\nabla T_x(x)$ to be taking the derivative in the tangent space.

For a function $f : M \to \mathbb{R}$ defined on a smooth manifold $M$, the function $f$ is a Morse function (Milnor, 1963; Morse, 1925, 1930) if all its critical points are non-degenerate. Namely, the eigenvalues of the Hessian matrix of $f$ at each critical point are non-zero. When the function $f$ is a Morse function, its $\lambda$-tree is stable (Chazal et al., 2014; Chen et al., 2016b) in the $L_\infty$ 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

$$\rho^\dagger_s : K_s \mapsto [0, \infty)$$

such that $\rho^\dagger_s(x) = \lim_{x_n \in K_s : x_n \to x} \rho(x_n)$ provided the limit exists and does not depend on the choice of the sequence $x_n$. It is easy to see that $\rho^\dagger_s(x) = \rho(x)$ when $x \in K_s$ but now it is defined on a smooth manifold $K_s$. We say $\rho(x)$ is a generalized Morse function if the corresponding $\rho^\dagger_s(x)$ is a Morse function for $s = 1, \cdots, d$. Later we will show that this generalization leads to a stable $\alpha$-tree for a singular measure.

For $K_s$, let $C_s = \{x \in K_s : \nabla T_x(x) \rho^\dagger_s(x) = 0\}$ be the collection of its critical points. Then the fact that $\rho^\dagger_s(x)$ is a Morse function implies that the eigenvalues of the Hessian matrix $\nabla T_x(c) \nabla T_x(c) \rho^\dagger_s(c)$ are non-zero for every $c \in C_s$. We call $g_s(x) = \nabla T_x(x) \rho^\dagger_s(x)$ the generalized gradient and $H_s(x) = \nabla T_x(x) \nabla T_x(x) \rho^\dagger_s(x)$ the generalized Hessian. For the case $s = 0$ (point mass), we define $C_0 = K_0$. The collection $C = \bigcup_{s=1,\cdots,d} C_s$ is called the collection of generalized critical points of $\rho(x)$. Each element $c \in C$ is called a generalized critical point.

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

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

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 $M$. 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.

Remark 2. If we further assume that the supports satisfy

$$K \supset K_d \supset K_{d-1} \supset \cdots \supset K_0,$$

then the support $K$ forms a stratified space (Goresky and MacPherson, 1980, 2012). Roughly speaking, a stratified space is a topological space $W$ such that there exists a decomposition (called stratification) $W_0, \cdots, W_d$ of $W$ with the properties that (i) each $W_k$ is a $k$-dimensional smooth manifold, (ii) $W = \bigcup_{\omega=0}^{d} W_\omega$, and (iii) for any $k \leq \ell$,

$$W_k \cap \overline{W_\ell} \neq \emptyset \leftrightarrow W_k \subset \overline{W_\ell}.$$

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 ($W_k$), 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 (Goresky and MacPherson, 1980; Edelsbrunner and Harer, 2008; Friedman, 2014). 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 $\alpha$ function and the $\alpha$-tree. In this paper, we focus on estimating $\alpha$-trees via the KDE:

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

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

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

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

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

$$\|\hat{\alpha}_n - \alpha\|_\infty = \sup_x |\hat{\alpha}_n(x) - \alpha(x)|.$$
The $L_\infty$ error has been used in several cluster tree literature; see, e.g., Eldridge et al. (2015); Chen et al. (2016b). An appealing feature of the $L_\infty$ 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). 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, because of the equivalence between the $L_\infty$ error and the merge distortion metric, convergence in $L_\infty$ 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

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

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

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

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 $L_\infty$ error—a large difference in a small region will not affect on these errors much.

To quantify the uncertainty in the topology of $\alpha$-tree, we introduce the notion of topological error, which is defined as

$$P\left( T_{\hat{\alpha}_n} \not\cong T_\alpha \right) = 1 - P\left( T_{\hat{\alpha}_n} \cong T_\alpha \right).$$

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

Finally, we define the following notations. For a smooth function $p$, we define $\|p\|_{\ell,\infty}$ as the supremum maximal norm of $\ell$-th derivative of $p$. For instance,

$$\|p\|_{0,\infty} = \sup_{x \in \mathbb{R}} p(x), \quad \|p\|_{1,\infty} = \sup_{x \in \mathbb{R}} \|g(x)\|_{\max}, \quad \|p\|_{2,\infty} = \sup_{x \in \mathbb{R}} \|H(x)\|_{\max},$$

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 non-negative integers
is called a multi-index with $|\beta| = \beta_1 + \beta_2 + \cdots + \beta_d$ and the corresponding derivative operator is

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

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 $K$ and is a Morse function and is four times differentiable with $\|p\|_{\ell, \infty} < \infty$ for $\ell = 0, \cdots, 4$.

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

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

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

(K2) Let

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

where $K^{(\beta)}$ is defined in equation (8) and $K_r^* = \bigcup_{r=0}^{l} K_r$ and $\bar{h}$ is some positive number. We assume that $K_2^*$ is a VC-type class, i.e. there exists constants $A, v$ and a constant envelope $b_0$ such that

$$(9) \quad \sup_Q N(K_2^*, \mathcal{L}^2(Q), b_0 \epsilon) \leq \left( \frac{A}{\epsilon} \right)^v,$$

where $N(T, d_T, \epsilon)$ is the $\epsilon$-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., 2014; Chen et al., 2016b). 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.
GENERALIZED CLUSTER TREES

(Wasserman, 2006; Scott, 2015). 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 Genovese et al. (2009, 2014); Chen et al. (2015b).

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

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

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

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

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

\[
P \left( T_{\hat{p}_n} \approx T_p \right) \geq 1 - c_0 \cdot e^{-c_1 \cdot nh^{d+4}},
\]

for some \( c_0, c_1 > 0 \).

The rate of consistency under the \( L_\infty \) 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 the topological error bound follows Lemma 2 in Chen et al. (2016b) and the concentration of \( L_\infty \) 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 et al., 2015b; Genovese et al., 2014), critical points (Chazal et al., 2014; Chen et al., 2016a), and persistent diagrams (Cohen-Steiner et al., 2007; Fasy et al., 2014; Chen, 2017).

Now we turn to the consistency for \( \alpha \)-tree. To derive the rate for the \( \alpha \)-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 (K1–2) and $d \geq 2$. Let $\lambda$ be a density level corresponding to the density of a critical point. When $h \to 0$, $\frac{\log n}{nh^{d+4}} \to 0$,

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

The rate in Theorem 3 is slower than the usual density estimation rate. This is because the boundary of $L_\lambda$ hits a critical point when $\lambda$ 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 (Giné and Guillou, 2002; Einmahl and Mason, 2005; Genovese et al., 2014) that under assumption (P) and (K1–2),

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

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 $\| \hat{p}_n - p \|_\infty$. 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 $\lambda$ 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 (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 level $\lambda$ (Singh et al., 2009; Rinaldo et al., 2012), (iii) an $(\epsilon, \sigma)$ condition for density (Chaudhuri and Dasgupta, 2010; Chaudhuri et al., 2014). When $\lambda$ equals a critical level, none of these assumptions hold.

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

**Theorem 4.** Assume (P1) and (K1–2) and $d \geq 2$ and the smoothing bandwidth satisfies $h \to 0$, $\frac{\log n}{nh^{d+4}} \to 0$. Let $C = \{ x : \nabla p(x) = 0 \}$ be the
collection of critical points and let \( a_n \) be a sequence of \( n \) such that \( \| \hat{p}_n - p \|_\infty = o(a_n) \). Then uniformly for all \( x \),

\[
\hat{\alpha}_n(x) - \alpha(x) = \begin{cases} 
O_P \left( \| \hat{p}_n - p \|_\infty \right) & \text{if } |p(x) - p(c)| > a_n \text{ for all } c \in C, \\
O_P \left( \| \hat{p}_n - p \|_{d+1} \right) & \text{otherwise}.
\end{cases}
\]

Theorem 4 shows uniform error rates for \( \hat{\alpha}_n \). 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 \( \hat{\alpha}_n \), we can choose \( a_n = \frac{1}{\log n} \left( O(h^2) + O_P \left( \sqrt{\frac{\log n}{nh^d}} \right) \right) \) which leads to the following result.

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

\[
\| \hat{\alpha}_n - \alpha \|_\infty = O \left( h^{\frac{2d}{d+1}} \right) + O_P \left( \frac{\log n}{nh^d} \right) \frac{4}{2(d+1)},
\]

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

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

\[
P \left( T_{\hat{\alpha}_n} \approx T_\alpha \right) \geq 1 - c_0 \cdot e^{-c_1 \cdot nh^{d+4}},
\]

for some \( c_0, c_1 > 0 \).

Comparing Corollary 5 to Theorem 2, we see that only the \( L_\infty \) 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 \( L_\infty \) error will be slowed down by these points. However, the collection of points \( \{ x : p(x) = p(c) \text{ for some } c \in C \} \) 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: \( T(\hat{p}_n) \approx T_{\hat{\alpha}_n}, T(p) \approx T_\alpha \).
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 $\lambda$-tree. However, as we discussed in Section 4, we are still able to define the $\alpha$-tree. Thus, in this section, we will focus on error rates for the $\alpha$-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 $p_h(x) = \mathbb{E}(\hat{\rho}_n)$, which is also known as the smoothed density.

Assumption.

(S) For all $s < d$, $\mathbb{K}_s$ 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 $0 < \rho_{\min} \leq \rho(x) \leq \rho_{\max} < \infty$ for all $x$. Moreover, for any $s > 0$, $\rho^s$ 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 $x \in \mathbb{K}_\ell$ and define $m(x) = \min\{\ell > s : x \in \mathbb{K}_\ell\} - s$. Let $C^\dagger_\ell = (\int_{B_\ell} K(||y||)dy)^{-1}$, where $B_\ell = \{y : ||y|| \leq 1, y_{\ell+1} = y_{\ell+2} = \cdots = y_d = 0\}$ for $\ell = 1, \cdots, d$ and $dy$ is integrating with respect to $\ell$-dimensional area and $C^\dagger_0 = 1/K(0)$. Then for a fixed $x$, when $h \to 0$ and $m(x) > 0$,

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

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

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

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 $K_s$ 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 $K(x) = \frac{1}{2}I(0 \leq x \leq 1)$ and consider the smoothing bandwidth $h \to 0$. This choice of kernel yields $C_1^\dagger = 1$. Now consider a sequence of points $x_h = 2 + \frac{h}{2}$. Then the smoothed density at each $x_h$ is

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

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, $|E(p_h(x_h) - \rho(x_h))|$ does not converge.

**Remark 4.** The scaling factor in Lemma 6 $C_{\tau(x)}^\dagger h^{d-\tau(x)}$ depends on the support $K_s$ where $x$ resides. In practice, we do not know $\tau(x)$ so we cannot properly rescale $\hat{p}_n(x)$ 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 $x_1 \lesssim_{\tau, \rho} x_2$). When $x_1$ and $x_2$ are both in $K_s$ 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. $x_1 \in K_{s_1}, x_2 \in K_{s_2}$, where $s_1 \neq s_2$), $p_h(x_1)$ 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 $\hat{\alpha}_n$ under other error measurements.

Due to the failure of uniform convergence in the bias, the $L_\infty$ error of $\hat{\alpha}_n$ does not converge under singular measures.
**Fig 5.** Good regions $K_s(h)$ for Example 1. Left: the original $K_2$, $K_1$, and $K_0$. Right: the corresponding $K_2(h)$, $K_1(h)$, and $K_0(h)$. The yellow area is the set $K_2(h)$, which are regions in $K_2$ away from lower-dimensional supports $K_1$ and $K_0$. The red area is the set $K_1(h)$, which are regions in $K_1$ where regions close to $K_0$ have been removed. The blue dot is $K_0(h)$, which is the same as $K_0$ because there is no lower dimensional support. Note that $K(h)$ is the union of the three color regions in the right panel.

**Corollary 7 (L∞ error for singular measures).** Assume $(S, P2)$. When $K_d \cap K_s \neq \phi$, for some $s < d$, $\|\hat{\alpha}_n - \alpha\|_\infty$ does not converge to 0. Namely, there exists $\epsilon > 0$ such that

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

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 $L_\infty$ error of the estimator $\hat{\alpha}_n$ does not converge in general. Thus, there is no guarantee for the Hartigan consistency of the estimated $\alpha$-tree.

**4.2. Error Measurements.** Although Corollary 7 presents a negative result on estimating the $\alpha$-tree, in this section, we show that the estimator $\hat{\alpha}_n$ is still consistent under other error measurements. A key observation is that there is a good region where the scaled KDE converges uniformly.

Let $K_s(h) = K_s \setminus (\bigcup_{\ell \leq s} K_\ell \oplus h)$ be the set that is in the interior of $K_s$ and is away from lower dimensional supports for $s > 0$; in the case of $s = 0$, we define $K_0(h) = K_0$. We define $K(h) = \bigcup_{s \leq d} K_s(h)$, which is the union of each $K_s(h)$. Figure 5 shows the good region of support $K_s(h)$ and the original support $K_s$ in Example 1. Later we will show that the set $K(h)$ is the good region.

In Lemma 6, the quantity

$$m(x) = \min\{\ell \geq \tau(x) : x \in K_\ell\} - \tau(x)$$
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, we introduce the quantity
\[
m_{\text{min}} = \inf_{x \in \mathbb{K}, m(x) > 0} m(x).
\]

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

**Theorem 8 (Consistency of the KDE under singular measures).** Assume \((S, P2, K1–2)\). Let \( C_\ell^\dagger \) be the constant in Lemma 6. Define

\[
\delta_{n,h,s} = \sup_{x \in \mathbb{K}_s(h)} \| C_\ell^\dagger h^{d-s} \hat{p}_n(x) - \rho(x) \|
\]

\[
\delta_{n,h,s}^{(1)} = \sup_{x \in \mathbb{K}_s(h)} \| C_\ell^\dagger h^{d-s} \nabla T_s(x) \hat{p}_n(x) - \nabla T_s(x) \rho(x) \|_{\text{max}}
\]

\[
\delta_{n,h,s}^{(2)} = \sup_{x \in \mathbb{K}_s(h)} \| C_\ell^\dagger h^{d-s} \nabla T_s(x) \nabla T_s(x) \hat{p}_n(x) - \nabla T_s(x) \nabla T_s(x) \rho(x) \|_{\text{max}},
\]

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

\[
\delta_{n,h,s} = O(h^2 \wedge m_{\text{min}}) + O_P \left( \sqrt{\frac{\log n}{nh^s}} \right),
\]

\[
\delta_{n,h,s}^{(1)} = O(h^2 \wedge m_{\text{min}}) + O_P \left( \sqrt{\frac{\log n}{nh^{s+2}}} \right),
\]

\[
\delta_{n,h,s}^{(2)} = O(h^2 \wedge m_{\text{min}}) + O_P \left( \sqrt{\frac{\log n}{nh^{s+4}}} \right),
\]

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 \( L_\infty \) error rate in the \( s \)-dimensional case with a modified bias term (bias is affected by a higher dimensional support).

**Remark 5. (Non-convergence of the integrated distance of the KDE)** One may wonder if the scaled KDE \( (C^\dagger_\tau(x) h^{d-\tau(x)} \cdot \hat{p}_n(x)) \) converges to the
generalized density $\rho(x)$ under the integrated distance. There is no guarantee for such a convergence because

$$\int \|C^\dagger_{\tau(x)} h^{d-\tau(x) \cdot \hat{p}_n(x) - \rho(x)} dx = O_P(1).$$

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

$$C^\dagger_{\tau(x)} h^{d-\ell} \cdot \hat{p}_n(y) - \rho(y) = O_P(h^{s-\ell}).$$

Such $y$ has $\ell$-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 $\hat{\alpha}_n$.

**Theorem 9** (Nearly uniform consistency of $\alpha$-trees). Assume $(S, P^2, K1-2)$. Let $C_s$ be the collection of generalized critical points of $K_s$. Let $\delta_{n,h,s}$ be defined in equation (11) and $r_{n,h,s}$ be a quantity such that $\delta_{n,h,s} = O_P(1)$. Then when $h \to 0, n h^{d+2} / \log n \to \infty$, uniformly for every $x \in K_s(h)$,

$$\hat{\alpha}_n(x) - \alpha(x) = \begin{cases} O(\delta_{n,h,s}) & \text{if } \inf_{c \in C_s} |\rho(x) - \rho(c)| > r_{n,h,s}, \\ O(\delta_{n,h,s}^{\frac{1}{d+2}}) & \text{otherwise} \end{cases}$$

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 $\alpha$ value of a generalized critical point (a critical $\alpha$ level), we have the usual convergence rate. When $\alpha(x)$ is close to a critical $\alpha$ level, the convergence rate is slower. The quantity $r_{n,h,s}$ behaves like the quantity $\varpi_n$ in Theorem 4, which was introduced to guarantee the uniform convergence. To derive the consistency of $\hat{\alpha}_n$ 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 $\alpha$-trees). Assume $(S, P^2, K1-2)$. Then

$$\|\hat{\alpha}_n - \alpha\|_P = O(\delta_{n,h,d}),$$
$$\|\hat{\alpha}_n - \alpha\|_\mu = O(\delta_{n,h,d}).$$
Theorem 9 shows that the quantity $\alpha(x)$ can be consistently estimated by $\hat{\alpha}_n(x)$ for the majority points. This implies that the ordering of points using $\hat{p}_n$ is consistent with the ordering from $\tau, \rho$ in most areas of $\mathbb{K}$.

5. Singular Measures: Critical Points and Topology. Recall from Section 2.1 that the topology of an $\alpha$-tree $T_\alpha$ is determined by its edge set $E(T_\alpha)$ and the relation among edges $C \in E(T_\alpha)$. The set $\mathcal{A}_\alpha$ (critical tree-levels) contains the levels where the upper level set $A_\varpi = \{x : \alpha(x) \geq \varpi\}$ changes its shape. For nonsingular measures, $\mathcal{A}_\alpha$ corresponds to the density value of some critical points. 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

$$\xi(\varpi) = \begin{cases} \max \{w \geq 0 : K_w \subset A_\varpi\}, & \text{if } K_0 \subset A_\varpi \\ -1, & \text{if } K_0 \not\subset A_\varpi \end{cases}.$$ 

Namely, $\xi(\varpi)$ is the highest dimensional structure that is covered by the level set $A_\varpi$.

**Definition 3.** For $\varpi \in \mathcal{A}$, we say $x$ is a DCP if the following holds...
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 $K_1$, 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 $A_\varpi$ where the level $\varpi$ 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.

(1) $x \in K_\ell$ for some $\ell \leq \xi(\varpi)$.

(2) There exists an edge $C \in E(T_\alpha)$ such that

(i) $x \notin C$ for all $C \in C$,

(ii) $d(x, C_\epsilon) \to 0$ when $\epsilon \to 0$, where $C_\epsilon = C \cap T_\alpha(\varpi + \epsilon)$.

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

The first requirement is to ensure that $x$ is on a lower dimensional support ($K_\ell : \ell < \xi(\varpi)$). The second requirement is to show that the DCP $x$ is not in the same edge $C$, 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 $A_{\varpi+\epsilon}$ for a sufficiently small $\epsilon$. Therefore, we can find $C' \in E(T_\alpha)$ such that every element $C \in C'$ contains $x$. Because $x$ is (i) in the elements of edge $C'$, and (ii) not in any element of edge $C$, and because (iii) the distance from $x$ to the element of $C$ converges to 0 when the level decreases to the level $\varpi$, $x$ is a merging point of edges $C'$ and $C$ and the level $\varpi$ 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, 2012).

Let $C^D$ be the collection of DCPs. For a point $c \in C^D$, we denote $\alpha^\dagger(c)$ as the level of $\alpha$ 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 $\alpha$ 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. $x \in K_{\xi(\varpi)+1}$.
2. There exists two different edges $C_1, C_2 \in E(T_\alpha)$ such that
   (i) $x \notin C_1, x \notin C_2$ for all $C_1 \in C_1$ and $C_2 \in C_2$,
   (ii) $d(x, C_{1,\epsilon}) \to 0$ when $\epsilon \to 0$, where $C_{1,\epsilon} = C_1 \cap T_\alpha(\varpi + \epsilon)$,
   (iii) $d(x, C_{2,\epsilon}) \to 0$ when $\epsilon \to 0$, where $C_{2,\epsilon} = C_2 \cap T_\alpha(\varpi + \epsilon)$.

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 $K_\ell$, $\ell \leq \xi(\varpi)$, then it is a DCP, and if $x$ lives in the support $K_{\xi(\varpi)+1}$, 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, $A_\alpha = \{1, 0.7, 0.0319\}$; 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 $x \in \partial K_s$ ($s > 0$) 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

$$\inf_{c \in C_s} d(c, K_\ell) \geq \eta_0,$$

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

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 $\mathcal{A}_\sigma$ 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 $K(h)$.

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

- If a new connected component of $\mathcal{A}_\sigma$ is created at $\sigma \in \mathcal{A}$, then there is a local mode $c$ of $\rho$ or an element in $K_0$ such that $\sigma = \alpha(c)$. Namely, DCPs only merge connected components.
- For any value $\sigma \in \mathcal{A}$, either $\sigma = \alpha(c)$ for some $c \in C$ or there is a DCP associated with $\sigma$.

Lemma 11 provides two basic properties of DCPs. First, DCPs only merge connected components. Moreover, when the topology of connected components of $\alpha$-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 $\alpha$-tree. Thus, in what follows, we will study the stability of generalized critical points and DCPs.
Lemma 12 (Stability of generalized critical points). Assume \((S, P_2, K_{1–2}, C)\). Let \(c \in K_s\) be a generalized critical point with \(n(c)\) negative eigenvalues of its generalized Hessian matrix. Let \(\hat{C}\) be the collection of critical points of \(\hat{p}_n\). Then when \(h \to 0\), \(\frac{nh^{d+4}}{\log n} \to \infty\), there exists a point \(\hat{c} \in \hat{C}\) such that

\[
\|\hat{c} - c\| = O(h) + O_P \left( \frac{1}{\sqrt{nh^{s+2}}} \right)
\]

\[
\|\hat{\alpha}_n(\hat{c}) - \alpha(c)\| = O_P \left( \frac{\delta_{n,h,s}(c_0(c))}{\log n}^{s+1} \right)
\]

and the estimated Hessian matrix at \(\hat{c}\) 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. (2014). 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 \((S, P_2, K_{1–2}, A, B)\). Assume there are \(k\) DCPs. Let \(\hat{C}\) be the critical points of \(\hat{p}_n\). Define \(\hat{G} \subset \hat{C}\) as the collection of estimated critical points corresponding to the generalized critical points. Let \(\hat{D} = \hat{C} \setminus \hat{G}\) be the remaining estimated critical points. Then when \(h \to 0\), \(\frac{nh^{d+4}}{\log n} \to \infty\),

- \(\hat{D} \subset K_C(h)\),
- \(|\hat{D}| \geq k\), where \(|A|\) for a set \(A\) is the cardinality,
- \(\hat{D}\) contains no local mode of \(\hat{p}_n\).

Lemma 13 provides several useful properties of the estimated critical points (critical points of \(\hat{p}_n\)). First, the estimated critical points are all in the bad region \(K_C(h)\), 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 \((S, P_2, K_{1–2}, A, B)\). Let \(c\) be a DCP and \(\alpha^\dagger(c) \in A\) be the associated level. Let \(\hat{\mathcal{D}}\) be defined as Lemma 13. Then when \(h \to 0\), \(\frac{nh^{d+2}}{\log n} \to \infty\), there exists a point \(\hat{c} \in \hat{\mathcal{D}}\) such that

\[
\|\hat{\alpha}_n(\hat{c}) - \alpha^\dagger(c)\| = O \left( \frac{\delta_{n,h,s}(c_0(c))}{\log n}^{s+1} \right),
\]
where $\delta_{n,h,s}$ is defined in (11). Moreover, the $\hat{K}_{n,\hat{c}}$ and $\hat{K}_{n,c}$ 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 $\hat{c} \in \hat{D}$, as described in Lemma 14, whose estimated $\alpha$-level is approaching the $\alpha$-level of the DCP at $x = 2$.

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

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

for some $c_0, c_1 > 0$.

Theorem 15 quantifies the topological error of the estimated $\alpha$-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 $\hat{p}_n$ (estimated $\lambda$-tree) converges topologically to a population cluster tree defined by the function $\alpha(x)$. To see this, recall that by Lemma 1, $T_{\hat{p}_n} \approx T_{\hat{\alpha}_n}$. This, together with Theorem 15, implies

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

under a suitable choice of $h$. This shows that even when the population distribution is singular, the estimated $\lambda$-tree still converges topologically to the population $\alpha$-tree.

**6. Discussion.** In this paper, we study how the $\alpha$-tree behaves under singular and nonsingular measures. In the nonsingular case, the error rate under the $L_{\infty}$ 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 $\lambda$-tree.

When a distribution is singular, we obtain many fruitful results for both the KDE and the estimated $\alpha$-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 $\lambda$-tree) converges topologically to a population cluster tree defined by $\alpha$.

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

1. $\hat{\alpha}_n$ is a pointwise consistent estimator of $\alpha$;
2. $\hat{\alpha}_n$ is a uniformly consistent estimator for the majority of the support;
3. $\hat{\alpha}_n$ is a consistent estimator of $\alpha$ under the integrated distance and probability distance; and
4. the estimated $\alpha$-tree converges topologically to the population $\alpha$-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 (Fasy et al., 2014; Bobrowski et al., 2017). 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 $L_\infty$ metric (Cohen-Steiner et al., 2007; Edelsbrunner and Morozov, 2012), many bounds on the bottleneck distance are derived via bounding the $L_\infty$ metric (Fasy et al., 2014; Bobrowski et al., 2017). However, for $\alpha$-trees under singular measures, the $L_\infty$ 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 $L_\infty$ 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 (Cohen-Steiner et al., 2007; Fasy et al., 2014; Bubenik, 2015) 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 (Friedman, 2014; Goresky and MacPherson, 1980; Edelsbrunner and Harer, 2008). 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 $\lambda$-tree under nonsingular measures and proved that the the k-nearest neighbor estimator is minimax. When the distribution is nonsingular, the $\alpha$-tree and $\lambda$-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.
Acknowledgements. 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: COMPLETED BY THE TYPESETTER; .pdf). This document contains all proofs to the theorems and lemmas in this paper.

References.


