Journal of Machine Learning Research 20 (2019) 1-50 Submitted 7/18; Revised 5/19; Published 9/19

DBSCAN: Optimal Rates For Density-Based Cluster Estimation

Daren Wang DARENW@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Xinyang Lu XLU8@LAKEHEADU.CA
Mathematical Sciences Department
Lakehead University
Thunder Bay, ON P7B 5E1, Canada

Alessandro Rinaldo ARINALDO@CMU.EDU
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Ingo Steinwart

Abstract

We study the problem of optimal estimation of the density cluster tree under various smoothness assumptions on the underlying density. Inspired by the seminal work of Chaudhuri et al. (2014), we formulate a new notion of clustering consistency which is better suited to smooth densities, and derive minimax rates for cluster tree estimation under Hölder smooth densities of arbitrary degree. We present a computationally efficient, rate optimal cluster tree estimator based on simple extensions of the popular DBSCAN algorithm of Ester et al. (1996). Our procedure relies on kernel density estimators and returns a sequence of nested random geometric graphs whose connected components form a hierarchy of clusters. The resulting optimal rates for cluster tree estimation depend on the degree of smoothness of the underlying density and, interestingly, match the minimax rates for density estimation under the sup-norm loss. Our results complement and extend the analysis of the DBSCAN algorithm in Sriperumbudur and Steinwart (2012). Finally, we consider level set estimation and cluster consistency for densities with jump discontinuities. We demonstrate that the DBSCAN algorithm attains the minimax rate in terms of the jump size and sample size in this setting as well.

Keywords: DBSCAN; density-based clustering; cluster tree; minimax optimality; Hölder smooth density.

1. Introduction

Clustering is one of the most basic and fundamental tasks in statistics and machine learning, used ubiquitously and extensively in the exploration and analysis of data. The literature on this topic is vast, and practitioners have at their disposal a multitude of algorithms and heuristics to perform clustering on data of virtually all types. However, despite its importance and popularity, rigorous statistical theories for clustering, leading to inferential procedures with provable theoretical guarantees, have been traditionally lacking in the literature. As a result, the practice of clustering, a central tasks in the analysis and manipulation of data, still relies in many cases on methods and heuristics of unknown or even dubious scientific validity. One of the most striking instances of such a disconnect is the DBSCAN algorithm of Ester et al. (1996), an extremely popular and relatively efficient (see Gan and Tao, 2015; Wang et al., 2015) clustering methodology whose statistical properties have been properly analyzed only very recently: see Sriperumbudur and Steinwart (2012), Jiang (2017a) and Steinwart et al. (2017).

In this paper, we provide a complementary and thorough study of DBSCAN, and show that this simple algorithm can deliver optimal statistical performance in *density-based* clustering. Density-based clustering (see, e.g., Hartigan, 1981) provides a general and rigorous probabilistic framework in which the clustering task is well-defined and amenable to statistical analysis. Given a probability distribution P on $\mathbb{R}^d$ with a corresponding continuous density p and a fixed threshold $\lambda \geq 0$, the λ*-clusters* of p are the connected components of the upper λ-level set of p, the set $\{x \in \mathbb{R}^d \colon p(x) \geq \lambda\}$ of all points whose density values exceed the level λ. With this definition, clusters are the high-density regions, subsets of the support of P with the largest probability content among all sets of the same volume.

As noted in Hartigan (1981), the hierarchy of inclusions of all clusters of p is a tree structure indexed by $\lambda > 0$, called the *cluster tree* of p. The chief goal of density clustering is to estimate the cluster tree of p, given an i.i.d. sequence $\{X_i\}_{i=1}^n$ of points with common distribution P. A cluster tree estimator is also a tree structure, consisting of a hierarchy of nested subsets of the sample points, and typically relies on non-parametric estimators of p in order to determine which sample points belong to high-density regions of p. A cluster tree estimator is deemed accurate if, with high probability, the hierarchy of clusters it encodes is close to the hierarchy that would have been obtained should p be known.

Density-based clustering, an instance of hierarchical clustering, enjoys several advantages: (1) it imposes virtually no restrictions on the shape, size and number of clusters, at any level of the tree; (2) unlike *flat* (i.e. non-hierarchical) clustering, it does not require a pre-specified number of clusters as an input and in fact the number of clusters itself is a quantity that may change depending on the level of the tree; (3) it provides a multi-resolution representation of all the clustering features of p across all levels λ at the same time; (4) it allows for an efficient representation and storage of the entire tree of clusters with a compact data structure that can be easily accessed and queried, and (5) the main object of interest for inference, namely the cluster tree of p, is a well-defined quantity.

Despite the appealing properties of the density-based clustering framework, a rigorous quantification of the statistical performance of this type of algorithms has proved difficult. Previous results by Hartigan (1981) and then Penrose (1995) have demonstrated a weaker notion of consistency achieved by the popular single-linkage algorithm. More recently Chaudhuri et al. (2014) have developed a general framework for defining consistency of cluster tree estimators based on a separation criterion among clusters. The authors further demonstrated that two graph-based algorithms, both based on k-nearest neighbors graphs over the sample points, achieve such consistency and provided minimax optimal consistency rates with respect to the parameters specifying the amount of cluster separation. Such results hold with virtually no assumptions on the underlying density. However, because of this generality, these consistency rates do not directly reflect any degree of regularity or smoothness of the underlying density. In particular, it remains unclear whether cluster tree estimation would be easier with smoother densities.

In this paper we provide further contributions to the theory of density based clustering by deriving novel, nearly minimax-optimal rates for cluster tree estimation that depend explicitly on the smoothness of the underlying density function. Our results further confirm that the smoother the density the faster the rate of consistency for the cluster tree estimation problem, a finding that is consistent with analogous results about non-parametric density estimation. Interestingly, our rates match those for estimating smooth densities in the L_∞ norm. To the best of our knowledge, this finding and the implication that density based clustering is no easier – at least in our setting – than density estimation, has not been rigorously shown before. In order to account explicitly for the smoothness of the density, we have developed a new criterion for cluster consistency that is better suited for smooth densities. In terms of procedures, we consider cluster tree estimators that arise from applying a very simple generalization of the well-known DBSCAN algorithm and are computationally efficient. Furthermore, our DBSCAN-based estimator is minimax optimal over arbitrary smooth densities according to our notion of consistency under appropriate conditions.

Related work

The idea of using the probability density function in order to study clustering structure dates back to Hartigan (1981), who formalized the notion of clusters as the connected components of high density regions and of cluster tree. Much of the subsequent theoretical work focused on consistency for "flat" clustering at a fixed level, which effectively reduces to level set estimation. The literature on this topic is vast and offers a multitude of results covering different settings and metric for consistency. See, e.g., Penrose (1995) Polonik (1995), Tsybakov et al. (1997), Cuevas and Fraiman (1997), BaÍllo et al. (2000), Klemelä (2004), Willett and Nowak (2007), Singh et al. (2009), Rigollet and Vert (2009), Rinaldo and Wasserman (2010). In contrast, there have been fewer contributions to the theory of practice of cluster tree estimation: see, e.g., Stuetzle (2003); Stuetzle and Nugent (2010), Klemelä (2009) and Rinaldo et al. (2012). The work of Chaudhuri et al. (2014) (see also Kpotufe and Luxburg (2011)) represented a significant advance in the theory of density-based clustering, as it derived a new framework and consistency rates for cluster tree estimation. Balakrishnan et al. (2012) generalized these results to the probability distributions supported over well-behaved manifolds, with consistency rates depending on the reach of the manifold and its intrinsic dimension. Corresponding guarantees in Hausdorff distance have been recently obtained by Jiang (2017a). Eldridge et al. (2015) developed a unified theory for consistency in cluster tree estimation that encompasses the original framework of Hartigan while Kim et al. (2016) investigated the challenging problems of defining adequate metrics over the space of cluster tree and of constructing confidence sets for cluster tree structures. Chen et al. (2017) provides bootstrap-based methods for constructing confidence sets for density level sets and for visualization of high-density clusters. Recently, Jang and Jiang (2018) proposed a variant of the DBSCAN algorithm with both minimax clustering rate and sub-quadratic computational complexity while Jiang et al. (2019) studied DBSCAN under possibly adversarial contamination of the input data.

In a parallel and important line of work, Steinwart (2011, 2015) developed a rigorous, measure-theoretic approach to density-based clustering whereby the cluster tree is recovered by estimating the lowest split level of the density and then proceeding recursively. The corresponding results demonstrate a direct link between density based clustering and optimal level set estimation. This approach was applied in Sriperumbudur and Steinwart (2012) to show that the DBSCAN algorithm yield consistent estimator of density trees, a result that was then extended in Steinwart et al. (2017) to allow for more general, KDE-based procedures. Our work built directly upon the contributions of Chaudhuri et al. (2014) and Steinwart (2015).

Organization of the paper

The rest of the paper is organized as follows. In Section 3, we describe the DBSCAN algorithm and establish its connections with non-parametric density estimation. In Section 4 we introduce a new notion of cluster consistency, called δ-consistency that is tailored to Hölder-continuous densities. We describe a DBSCAN-based algorithm for clustered tree estimation that is computational efficient and delivers nearly optimal minimax rates that depend explicitly on the degree of smoothness of the underlying density, whereby cluster tree of smoother densities can be estimated at faster rates. Interestingly and, perhaps surprisingly, for the class of DBSCAN-based algorithms we consider, we observe a trade-off between statistical optimality and computational cost for smoother Hölder densities of degree $\alpha > 1$. In these situations, minimax rates can still be achieved by our computationally efficient algorithm provided that the underlying density satisfies additional geometric regularity conditions around the split levels. Such conditions are relatively mild and have been exploited before; see in particular Steinwart (2015). Finally, in Section 5 we consider a different scenario in which the underlying density exhibits jump discontinuities. We are particularly interested in level set and cluster estimation at the jump, with the assumption that the size of the discontinuity is vanishing when $n \to \infty$ so that clustering becomes increasingly difficult. We show that, with suitable inputs, the DBSCAN algorithm returns a Devroye-Wise type of estimator which is minimax optimal for cluster recovery and level set estimation. In addition, we derive the minimax scaling for the size of the jump discontinuity.

Notation

We denote with p a density for the distribution P of the i.i.d. sample $\{X_i\}_{i=1}^n \subset \mathbb{R}^d$. For a constant $\lambda > 0$, we set $L(\lambda) = \{p \geq \lambda\}$ to be the λ-upper upper level set of the density p. We use T_p to denote the cluster tree generated by the density p and $\widehat{T}$ to density any estimator of T_p. We use subscript n to emphasize any global variable which may change with respect to n. $\mathcal{L}$ represents the Lebesgue measure in $\mathbb{R}^d$ and $B(x,r)$ the closed d dimensional Euclidean ball centered at x with radius r and $V_d = \mathcal{L}(B(0,1))$ the volume of the unit ball $B(0,1)$. For a vector x we denote with $\|x\|$ and $\|x\|_\infty$ its Euclidean and L_∞ norms, respectively. With a slight abuse of notation, if f is a real valued function defined over a subset S of $\mathbb{R}^d$, we let $\|f\|_\infty = \sup_{x\in S} |f(x)|$ its L_∞ norm. For any $h > 0$ and a measurable set $A \subset \mathbb{R}^d$ we set

$$A_h = \bigcup_{x \in A} B(x,h) \quad \text{and} \quad A_{-h} = \{x \in A : B(x,h) \subset A\}. \tag{1}$$

For any two real sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ we write $a_n = O(b_n)$ if there exists $C > 0$ such that $\limsup_{n\to\infty} |a_n/b_n| < C$ and write $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. For any two closed subsets A and B of $\mathbb{R}^n$, we use $d(A,B) = \inf_{x\in A, y\in B} \|x-y\|$ to represent the ordinary distance between them.

2. Cluster Trees Estimation

Let P be a probability distribution with a continuous[1] Lebesgue density p and with support $\Omega \subset \mathbb{R}^d$. For any $\lambda \geq 0$, let $\{x \in \Omega \colon p(x) \geq \lambda\}$ be the λ-upper level set of p and the λ-cluster of p are the connected components of $L(\lambda)$. See Appendix A for definition of connectedness. Notice that the set of all clusters is an indexed collection of subsets of Ω, whereby each cluster of p is assigned the index λ associated to the corresponding super-level set $L(\lambda)$, and that many clusters may be indexed by the same level λ. The cluster tree of p is the collection T_p of all clusters of p, that is

$$T_p = \{L(\lambda)\}_{\lambda \geq 0}.$$

We can think of the cluster tree of p as the function defined on $[0,\infty)$ and for each $\lambda \geq 0$, it returns the set of λ-clusters of p. Thus, $T_p(\lambda)$ consists of disjoint connected subsets of Ω. We remark that, since the density p is unique only up to sets of Lebesgue measure zero, the cluster tree T_p is also not unique. In fact, Steinwart (2015) shows that there exists a well-defined notion of cluster tree for the distribution P that is independent of the choice of the density. Furthermore, if P admits an upper semi-continuous density p, then the cluster tree is in fact composed of the hierarchy of the (closures of the) upper level sets of such density. As P is assumed, throughout most of the article, to have a density that is continuous everywhere on its support, when we speak of "the" density of P, we will refer to this canonical choice.

The concept of cluster tree owes its name to the easily verifiable property (see Hartigan (1981)) that if A and B are elements of T_p, i.e. distinct clusters of p, then $A \cap B = \emptyset$ or $A \subseteq B$ or $B \subseteq A$. This induces a partial order on the set of clusters. In particular, for any $\lambda_1 \geq \lambda_2 \geq 0$, if $A \in T_p(\lambda_1)$ and $B \in T_p(\lambda_2)$ then either $A \cap B = \emptyset$ or $B \subseteq A$. As a result, T_p can be represented as a dendrogram with height indexed by $\lambda \geq 0$. We refer to Kim et al. (2016) for a formal definition of the dendrogram encoding a cluster tree.

Let $\{X_i\}_{i=1}^n$ be i.i.d. samples from P. In order to estimate the cluster tree of p we will consider tree-valued estimators, defined below.

Definition 1. *A cluster tree estimator of T_p is a collection $\widehat{T}_n$ of subsets of $\{X_i\}_{i=1}^n$ indexed by $[0,\infty)$ such that*
- *for each $\lambda \geq 0$, $\widehat{T}_n(\lambda)$ consists of disjoint subsets of $\{X_i\}_{i=1}^n$ (including, possibly the empty set), called clusters, and*
- *$\widehat{T}_n$ satisfies the tree property: for any $\lambda_1 \geq \lambda_2 \geq 0$, if $A \in \widehat{T}_n(\lambda_1)$ and $B \in \widehat{T}_n(\lambda_2)$ then either $A \cap B = \emptyset$ or $A \subseteq B$.*

[1] Density based clustering does not require in general continuous densities.

It is important to realize that, while the cluster tree T_p is a collection of connected subsets of the support of p, the cluster tree estimators considered in this paper are collections of subsets of the sample points.

In order to quantify how well a cluster-tree estimator approximates the true cluster tree, we will make use of the notion of cluster tree consistency put forward by Chaudhuri et al. (2014).

In detail, let $\mathcal{A}_n$ denote a collection of connected subsets of the support of p, which may depend on n. A cluster tree estimator $\widehat{T}_n$ is consistent with respect to $\mathcal{A}_n$ if, with probability tending to 1 as $n \to \infty$, the following holds simultaneously over all A and A' in $\mathcal{A}_n$: the smallest clusters in $\widehat{T}_n$ containing $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ are disjoint. The requirement for consistency outlined above is rather natural: if a cluster tree is deemed consistent with respect to the sequence $\mathcal{A}_n$, then it should, with probability tending to 1, cluster the sample points perfectly well, as if we had the ability of verifying, for each pair of sample points X_i and X_j and each connected set $A \in \mathcal{A}_n$, whether both X_i and X_j are in A.

We allow $\mathcal{A}_n$ to grow larger and more complex with n, so that the cluster tree estimator will be able to discriminate among clusters of p that are barely distinguishable given the size of the sample. An example of a sequence $\{\mathcal{A}_n\}_{n=1}^\infty$ is the set of δ_n-separated clusters according to Definition 2, where the parameter δ_n is taken to be vanishing as $n \to \infty$. The sequence of target subsets $\{\mathcal{A}_n\}_{n=1}^\infty$ may not be chosen to be too large: for example if $\mathcal{A}_n$ equals to the set of all clusters of p, then, depending on the complexity of p, no cluster tree estimator need to be consistent. A natural way to define $\{\mathcal{A}_n\}_{n=1}^\infty$ is by specifying a *separation criterion* for sets, which may become less strict as n grows, and then populate each $\mathcal{A}_n$ using only the connected subsets of the support of p fulfilling such a criterion. In particular, Chaudhuri et al. (2014) develop a criterion known as the (ϵ, σ)-separation, which requires two connected subsets A and A' to be far apart from each other in terms of their "horizontal" distance $d(A, B)$ and their "vertical" distance, in the sense that the smallest cluster containing both A and B should belong to a level set of p indexed by a value of λ significantly smaller to the values indexing the level sets of A and B. See Definition 10 below for details. One of the main contributions of this paper is to replace this rather general notion of separation by a simpler one, the δ-separation criterion in Definition 2, which is better suited deal with smooth densities. This allows us to derive new rates of consistency that depend explicitly on the smoothness of the density.

As explained in Eldridge et al. (2015), the cluster tree consistency guarantees based on separation criteria can be fairy coarse, as they only require $\widehat{T}_n$ to preserve the connectivity of all the sets in $\mathcal{A}_n$. In particular, a tree estimator that is consistent with respect to such definition needs not yield a good clustering of the sample points. Concretely, $\widehat{T}_n$ might have additional unwanted clusters, referred to as *false* in Chaudhuri et al. (2014), that do not correspond to any disjoints sets in $\mathcal{A}_n$, a phenomenon referred to as *over-segmentation* by Eldridge et al. (2015). Similarly, $\widehat{T}_n$ might not conform to the partial order of inclusions among the clusters of p, an issue called *improper nesting*. In fact, the estimators developed in this paper do not suffer from such shortcomings and are consistent in the merge distortion metric of Eldridge et al. (2015), a more refined stronger notion of consistency for cluster trees. See Sections 4.6 and 6 below.

3. The DBSCAN Algorithm

The DBSCAN algorithm, first introduced in Ester et al. (1996), is an extremely popular methodology for "flat" clustering. In this section we introduce a simple generalization of DBSCAN, shown below in Algorithm 1, that yields cluster tree estimators and establish its connections with kernel density estimation.

Algorithm 1 The DBSCAN algorithm.

```
INPUT: i.i.d sample $\{X_i\}_{i=1}^n$, and $h > 0$.
  1. For each $k \in \mathbb{N}$, construct a graph $\mathbb{G}_{h,k}$ with nodes $\{X_i : |B(X_i, h) \cap \{X_j\}_{j=1}^n| \geq k\}$
     and edges $(X_i, X_j)$ if $\|X_i - X_j\| < 2h$.
  2. Compute $\mathbb{C}(h, k)$, the graphical connected components of $\mathbb{G}_{h,k}$.
OUTPUT: $\{\mathbb{C}(h, k), k \in \mathbb{N}\}$.
```

For a fixed value of k, Algorithm 1 is in fact a simplified version of the original DBSCAN procedure of Ester et al. (1996), where the parameters h and k are called instead Eps and MinPts, respectively. Notice that, unlike in the original formulation of DBSCAN, we do not distinguish between core and border points and, furthermore, we evaluate connectivity among the sample points using balls of radius $2h$ instead of h. Such modifications have no impact on the rates of consistency we obtain but simplify the derivations.

Assuming $h > 0$ fixed, by sweeping through all the possible values of k, Algorithm 1 produces a sequence of nested geometric graphs $\widehat{T}_n = \{\mathbb{C}(h, k)\}_{k \in \mathbb{N}}$. It is immediate to see that $\widehat{T}_n$ forms a cluster tree estimator over the sample points $\{X_i\}_{i=1}^n$; see Definition 1. This is because, for each $k_1 \leq k_2$,

$$\bigcup_{\{X_i \colon \, |B(X_i,h) \cap \{X_j\}_{j=1}^n| \geq k_2\}} B(X_i, h) \subseteq \bigcup_{\{X_i \colon \, |B(X_i,h) \cap \{X_j\}_{j=1}^n| \geq k_1\}} B(X_i, h).$$

In practice, Algorithm 1 can be efficiently implemented using a union-find structure in such a way that the set $\mathbb{C}(h, k)$ of the maximal connected components of $\mathbb{G}_{h,k}$ can be computed without using the potentially expensive breadth-first search or depth-first search algorithms. The resulting cluster tree algorithm is simpler than the estimator based on Wishart's algorithm proposed in Chaudhuri et al. (2014). Indeed, the DBSCAN-based estimator is obtained from a sequence of node-induced sub-graphs of the $2h$-neighborhood graph over the sample points. In contrast, Wishart's algorithm entails taking node and edge-induced sub-graphs of the k-nearest neighborhood graph over $\{X_1, \ldots, X_n\}$, which has higher computational complexity.

As explained in Sriperumbudur and Steinwart (2012), DBSCAN is implicitly using a kernel density estimator with kernel corresponding to the indicator function of the unit d-dimensional Euclidean ball to cluster the points. In detail, consider the density estimator $\widehat{p}_h$ given by

$$x \in \mathbb{R}^d \mapsto \widehat{p}_h(x) = \frac{|B(x, h) \cap \{X_i\}_{i=1}^n|}{nh^d V_d} = \frac{1}{nh^d V_d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \tag{2}$$

where

$$K(x) = \begin{cases} 1 & \text{if } x \in B_d(0, 1), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

It is easy to see that $\widehat{p}_h$ is a Lebesgue density, i.e. $\widehat{p}_h(x)$ is a measurable, non-negative function and $\int_{\mathbb{R}^d} \widehat{p}_h(x)dx = 1$. Furthermore, $\mathbb{E}[\widehat{p}_h(x)] = p_h(x)$ for all $x \in \mathbb{R}^d$, where

$$p_h(x) = \frac{1}{h^d V_d} \int_{\mathbb{R}^d} K\left(\frac{x-z}{h}\right) p(z)dz = \frac{P(B(x,h))}{h^d V_d}. \tag{4}$$

For any $\lambda \geq 0$, set $\widehat{D}(\lambda) = \{x \colon \widehat{p}_h(x) \geq \lambda\} \cap \{X_i\}_{i=1}^n$ and

$$\widehat{L}(\lambda) = \bigcup_{X_j \in \widehat{D}(\lambda)} B(X_j, h). \tag{5}$$

Then, setting, for $k > 0$, $\lambda_k = \frac{k}{nh^d V_d}$, one can see that, for clustering purpose, $\mathbb{C}(h,k)$ and $\widehat{L}(\lambda_k)$ convey the same information. Indeed from the definition of $\widehat{L}(\lambda_k)$, it is straightforward to see that

Lemma 1. *Two data points X_i and X_j are in the same connected component of the d-dimensional set $\widehat{L}(\lambda_k)$ if and only if they are in the same connected component of the graph $\mathbb{C}(h,k)$.*

The union of balls $\widehat{L}(\lambda)$ is a renown estimator in the literature on level set estimation, originally studied in Devroye and Wise (1980) (see also Cuevas and Rodríguez-Casal (2004)). In particular, with a suitable choice of the bandwidth parameter h and as n grows unbounded, $\widehat{L}(\lambda)$ is a rate-optimal estimator of the level set $L(\lambda)$ under various loss functions and appropriate assumptions on the underlying density.

4. Clustering Consistency for Hölder Continuous Densities

In this section we show that the DBSCAN algorithm 1 is consistent under Hölder smooth densities. Towards that end, we introduce a new notion of cluster tree consistency, called δ-consistency (see Section 4.2 below), which is well-suited to study cluster trees generated by smooth densities. We will show that DBSCAN, with suitable inputs, will return cluster tree estimators that nearly attain the corresponding minimax optimal rates and that those rates depend on the degree of smoothness of the density.

4.1. Hölder smooth densities

Below we give a recap of well-known results on non-parametric density estimation. Given vectors $s = (s_1, \ldots, s_d)$ in $\mathbb{N}^d$ and $x = (x_1, \ldots, x_d)$ in $\mathbb{R}^d$, set $|s| = s_1 + \cdots + s_d$ and $x^s = x_1^{s_1} \ldots x_d^{s_d}$, and let

$$D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$$

denote the differential operator. A function $p : \mathbb{R}^d \to \mathbb{R}$ is said to belong the Hölder class $\Sigma(L, \alpha)$ with parameters $\alpha > 0$ and $L > 0$ if p is $\lfloor \alpha \rfloor$-times continuously differentiable and, for all $x, y \in \mathbb{R}^d$ and all $s \in \mathbb{N}^d$ with $|s| = \lfloor \alpha \rfloor$,

$$|D^s p(x) - D^s p(y)| \leq L\|x - y\|^{\alpha - s}.$$

Notice that, when $0 < \alpha \leq 1$, the Hölder condition reduces to the Lipschitz condition

$$|p(x) - p(y)| \leq L\|x - y\|^{\alpha}, \quad \forall x, y \in \mathbb{R}^d.$$

Let $\widehat{p}_h$ denote a kernel density estimator with bandwidth h and kernel K, that is

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

Then, we obtain the standard bias-variance decomposition for the KDEs, namely

$$\|\widehat{p}_h - p\|_\infty \leq \|\widehat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty.$$

In order to control the stochastic component $\|\widehat{p}_h - p_h\|_\infty$ we will invoke well-known concentration bounds from density estimation to conclude that, under appropriate and very mild assumptions on K, there exists a constant $C_1 > 0$, depending on $\|p\|_\infty$, d and K, such that, for any $\gamma > 0$ and all n large enough and assuming $nh^d \geq 1$,

$$\mathbb{P}\left(\|\widehat{p}_h - p_h\|_\infty \leq a_n\right) \geq 1 - e^{-\gamma}, \tag{6}$$

where

$$a_n = C_1 \sqrt{\frac{(\gamma + \log(1/h))}{nh^d}}. \tag{7}$$

The verification of this bound can be found in many places in the literature; see, e.g., Giné and Guillou (2002), Sriperumbudur and Steinwart (2012), Jiang (2017b) and Kim et al. (2019). See Appendix B for details. As for the bias term $\|p_h - p\|_\infty$, if K is chosen to be a α-valid kernel[2] (see, e.g., Rigollet and Vert (2009)), then standard calculations yield that

$$\|p_h - p\|_\infty \leq C_2 h^{\alpha}, \tag{8}$$

for an appropriate constant $C_2 > 0$ depending on L and α. In particular, since this type of kernels are polynomials supported on $[0, 1]^d$, they automatically satisfy the VC condition (see lemma 22 of Nolan and Pollard, 1987, for a justification).

Thus combining the bias in (6) and the variance in (8), we conclude that, with probability at least $1 - e^{-\gamma}$, with γ any positive number,

$$\|\widehat{p}_h - p\|_\infty \leq C_1 \sqrt{\frac{\gamma + \log(1/h)}{nh^d}} + C_2 h^{\alpha}. \tag{9}$$

Setting, e.g., $\gamma = \log n$, the optimal choice of the bandwidth is

$$h \asymp \left(\frac{\log n}{n}\right)^{2\alpha + d}, \tag{10}$$

leading to the final rate of $\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}$, which is in fact minimax optimal.

[2] For a fixed $\alpha > 0$, a function $K : \mathbb{R}^d \to \mathbb{R}$ is an α-valid kernel if $\int_{\mathbb{R}^d} K(x)dx = 1$, has finite L_p norm for all $p \geq 1$, $\int_{\mathbb{R}^d} \|x\|^{\alpha} K(x)dx < \infty$ and $\int_{\mathbb{R}^d} x^s K(x)dx = 0$ for all $s = (s_1, \ldots, s_d) \in \mathbb{Z}^d$ such that $1 \leq \sum_{i=1}^{d} s_i \leq \lfloor \alpha \rfloor$, where for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, $x^s = \prod_{i=1}^{d} x_i^{s_i}$. See Definition 1 in Rigollet and Vert (2009).

4.2. The δ-Separation Criterion

We now formulate a notion of cluster separation that is naturally suited to smooth densities and, for continuous densities is equivalent to separation in the merge distance of Eldridge et al. (2015) (see also Kim et al., 2016).

Definition 2. *Let A and A' be subsets the support of the density p and set $\lambda := \inf_{x \in A \cup A'} p(x)$. For $\delta \in (0, \lambda)$, A and A' are said to be δ-separated if they belong to distinct connected components of the level set $\{x \colon p(x) > \lambda - \delta\}$.*

Unlike the separation criterion of Chaudhuri et al. (2014), which requires the specification of two parameters quantifying the horizontal and vertical displacement between clusters, δ-separation only uses one parameter The intuition behind the notion of δ-separation is simple: due to the smoothness of the density, the degrees of "vertical" and "horizontal" separation between clusters are coupled. This is illustrated in Figure 1 and best explained for the case of a density in $\Sigma(L, \alpha)$ with $\alpha \leq 1$. If A and A' are δ-separated, then their distance is at least $\left(\frac{\delta}{L}\right)^{1/\alpha}$. As a result, the degree of separation between clusters of smooth densities can be described using only one parameter, a feature that we will exploit to derive new notion of consistency for clustering.

Definition 3. *Let $\delta > 0$ and $\gamma \in (0, 1)$. A cluster tree estimator based on an i.i.d. sample $\{X_i\}_{i=1}^n$ is (δ, γ)-accurate if, with probability no smaller than $1-\gamma$, for any pair of connected subsets A and A' of the support that are δ-separated, exactly one of the following conditions holds:*

1. *at least one of $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ is empty;*
2. *the smallest clusters in the cluster tree estimator containing $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ are disjoint.*

Let $\{\delta_n\}_n$ be a vanishing sequence of positive numbers and a $\{\gamma_n\}$ a vanishing sequence in $(0, 1)$. We say that the sequence of cluster tree estimators $\{T_n\}_n$, where T_n is based on an i.i.d. sample $\{X_i\}_{i=1}^n$, is δ-consistent with rate δ_n if, for all n large enough, T_n is (δ_n, γ_n)-accurate where γ_n decays polynomially in n.

It is important to realize that the notion of δ-consistency is a *uniform* notion of consistency that is required to hold simultaneously over all possibly pairs of δ-separated connected subsets of the support.

The δ-separation criterion is closely related to the concept of the *merge height* introduced by Eldridge et al. (2015). In the context of hierarchical clustering, the merge height is used to describe the "height" at which two points or two clusters merge into one cluster; see Definition 9. In particular we show below in Lemma 4 that if two subsets A and A' of the support are δ-separated and $\inf_{x \in A \cup A'} p(x) = \lambda$, then their merge height is no larger than $\lambda - \delta$. To further emphasize how similar the two approaches are, we mention that our results about cluster consistency still hold for a slightly stronger notion of cluster consistency, whereby condition 2 in Definition 3 is replaced by the condition

2. there exists a level $\lambda \in [\inf_{x \in A \cup A'} f(x) - \delta, \inf_{x \in A \cup A'} f(x))$ such that $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ are contained in two different λ-clusters of the cluster tree estimator.

The key difference between δ-consistency and this stronger version of δ-consistency is that in the latter case we further constrain the split level for A and A' in the cluster tree estimator to occur at a value less than $\inf_{x \in A \cup A'} f(x)$ by an amount no larger than δ_n. This is precisely what is required for merge distance consistency; see Eldridge et al. (2015). We provide more detailed comparison in Section 6, where we further elucidate the differences between our notion of δ-separation and the (ϵ, σ)-separation criterion of Chaudhuri et al. (2014).

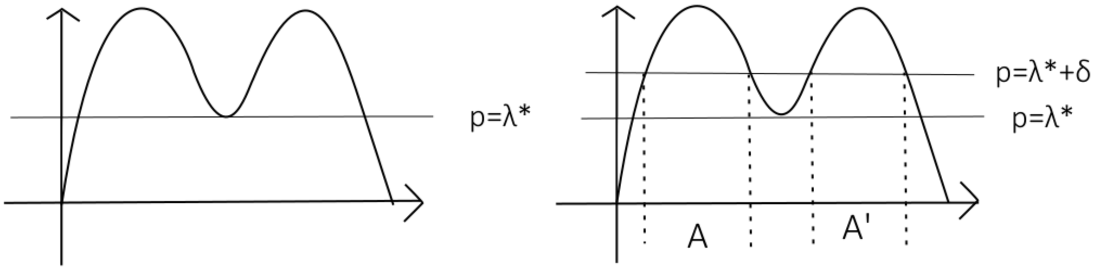

Figure 1: The left figure depicts a split level (defined in λ^* Section 4.3) of the density p. The right figure depicts two sets A and A' being δ separated with respect to λ^*.

4.3. The split levels

One of the most impoertant features of a cluster tree is the collections of levels λ at which the clusters split into two or more disjoint sub-clusters, which we refer to as *split levels.* Such levels belong to the well-known class of "critical levels" in differential topology, which identify critical changes in the topology of the upper level sets of p. See, for example, Hirsch (2012) for more details. In particular, the estimation of split levels is a central theme in the contributions of Sriperumbudur and Steinwart (2012) and in Steinwart (2015). Below, we provide a slightly different characterization of the split levels of continuous densities and relate it to the criterion of δ-separation of clusters. The notion of split levels will be important below in Section 4.4.2 in formalizing conditions under which computationally efficient and statistically optimal cluster tree estimation is feasible for Hölder densities with smoothness degree $\alpha > 1$. It will also be used to demonstrate that we can easily remove false clusters from the cluster tree estimations returned by our algorithms (see Section 4.6).

Definition 4. *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a continuous density function. For a fixed $\lambda^* > 0$, let $\{\mathcal{C}_k\}_{k=1}^K$ be the connected components of $\{p \geq \lambda^*\}$. The value λ^* is said to be a split level of p if there exists a $\mathcal{C}_k$ such that $\mathcal{C}_k \cap \{p > \lambda^*\}$ has two or more connected components.*

The following simple result illustrates the main topological properties of split levels.

Proposition 1. *Suppose that $p : \mathbb{R}^d \to \mathbb{R}$ is compactly supported and that A and A' are subsets of two distinct connected components of $\{p \geq \lambda_1\}$. If A and A' belongs to the same connected components of $\{p \geq \lambda_2\}$, where $\lambda_2 < \lambda_1$, then there is a unique split level $\lambda^* \in [\lambda_2, \lambda_1)$ such that A and A' belong to one connected component of $\{p \geq \lambda^*\}$ and to distinct connected components of $\{p > \lambda^*\}$.*

Proposition 1 suggests that if two connected components merge into one as the density level λ decreases, then there exists one and only one split level at which the corresponding merge takes place. Therefore, the following definition, which characterizes the split level of any two distinct clusters in a cluster tree, seems natural.

Definition 5. *Suppose A and A' are two open subsets of the support of p. Then A and A' are said to split at level λ^* if A and A' belong to one connected component of $\{p \geq \lambda^*\}$ and to two distinct connected components of $\{p > \lambda^*\}$.*

In our next result we illustrate a direct link between the notion of split levels and the criterion of δ-separation introduced above. We will exploit this fact later in Section 4.6 to demonstrate how to prune the cluster tree estimators to yield accurate estimates of the split levels without producing false clusters.

Corollary 1. *Let A and A' be δ-separated. Then there exists a split level λ^* of the density, with*

$$\lambda^* \leq \inf_{x \in A \cup A'} f(x) - \delta,$$

such that A and A' belong to one connected component of $\{p \geq \lambda^*\}$ and to two distinct connected components of $\{p > \lambda^*\}$.

4.4. Rate of consistency for the DBSCAN algorithm

We are now ready to present the main results of the paper, and derive rates of consistency for DBSCAN-based cluster tree estimators with respect to the criterion of δ-separation and for Hölder smooth densities. Specifically, we will show that these estimators are δ-consistent with rate

$$\delta_n \geq C \left(\frac{\log(n)}{n} \right)^{\frac{\alpha}{2\alpha+d}}, \tag{11}$$

for an appropriate constant C that depends on $\|p\|_\infty$, L, K and α. The above rates depend on the smoothness of the underlying density, with smoother densities leading to faster rates, and, as shown in Section 4.5, are in fact minimax optimal. This is one of the main contributions of this article and delivers an extension of the cluster consistency results of Chaudhuri et al. (2014), which are agnostic to the smoothness of p.

4.4.1. Consistency for $\alpha \leq 1$

We first show that, when $\alpha \leq 1$, the DBSCAN algorithm is δ-consistent with rate of order (11). We remark that this type of result can be deduced from several contributions in the literature on density-based clustering, which show that variants of the DBSCAN algorithm lead to some form of cluster consistency when $\alpha \leq 1$. See, e.g., Rinaldo and Wasserman

(2010), Sriperumbudur and Steinwart (2012), Jiang (2017a) and Steinwart et al. (2017). We provide the details for completeness.

In order to demonstrate that DBSCAN is δ-consistent, it will be sufficient to show that the procedure provides an approximation to the upper level sets of $\widehat{p}_h$. This is done in the next result, which relies on general, well-known, finite sample concentration bounds for KDEs along with standard calculations for the bias of a KDE; see Lemma 6 and Proposition 7 in Appendix B.

Lemma 2. *Assume that $p \in \Sigma(L, \alpha)$, where $\alpha \in (0, 1]$, and let K be the spherical kernel. Then, there exist constants C_2, depending on C_1, $\|p\|_\infty$, L and d such that, if $h = C_1 n^{-\frac{1}{2\alpha+d}}$ then uniformly over all $\lambda > 0$, with probability at least $1 - 1/n$,*

$$\left\{p \geq \lambda + C_2 \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)} + Lh^\alpha\right\} \subset \bigcup_{X_j \in \widehat{D}(\lambda)} B(X_j, h) \subset \left\{p \geq \lambda - C_2 \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)} - Lh^\alpha\right\}. \tag{12}$$

As a direct corollary, we see that the DBSCAN algorithm, with an appropriate choice of the bandwidth h, outputs a δ-consistent cluster tree with consistency rates that depend on α.

Corollary 2. *Assume that $p \in \Sigma(L, \alpha)$, where $\alpha \in (0, 1]$, and let K be the spherical kernel. Then, there exist constants C_1 depending on $\|p\|_\infty$, L and d such that, if $h = C_1 \left(\frac{\log(n)}{n}\right)^{\frac{1}{2\alpha+d}}$, the cluster tree returned by the DBSCAN Algorithm 1 is δ-consistent with rate $\delta_n \geq C \left(\frac{\log(n)}{n}\right) n^{\frac{\alpha}{2\alpha+d}}$, where $C = C(\|p\|_\infty, L, d)$.*

4.4.2. Consistency for $\alpha > 1$

When $\alpha > 1$, Algorithm 1 no longer delivers the optimal rate displayed in (11), for two reasons. The first reason stems from standard non-parametric density estimation considerations: when $\alpha > 1$ it becomes necessary to rely on smoother kernels, namely α-valid kernels as indicated before. This will lead to a bias $\|p - p_h\|_\infty$ of the correct order $O(h^\alpha)$. The second reason is more subtle: the straightforward arguments we used to handle the case of $\alpha \leq 1$ do not lead to optimal clustering rates even if the kernel K is chosen to be α-valid. To exemplify, suppose we would like to cluster the sample points $\{X_i\}_{i=1}^n \cap \{x \colon \widehat{p}_h(x) \geq \lambda\}$ for some $\lambda > 0$. The computationally efficient linkage rule implemented by DBSCAN is to cluster the points based on the connected components of the union-of-balls around them, i.e based on the connected components of

$$\widehat{L}(\lambda) = \bigcup_{X_j \in \{\widehat{p}_h \geq \lambda\}} B(X_j, h).$$

Assume now that the gradient of p has norm uniformly bounded by a constant D for all $x \in L(\lambda)$. Then,

$$\max_{X_j \in \{\widehat{p}_h \geq \lambda\}} \sup_{x \in B(X_j, h)} |p(x) - p(X_j)| \leq Dh, \tag{13}$$

and, as a result,

$$\left\{p \geq \lambda + C\left(\sqrt{\frac{\log(n)}{nh^d}} + h^\alpha\right) + Dh\right\} \subset \widehat{L}(\lambda) \subset \left\{p \geq \lambda - C\left(\sqrt{\frac{\log(n)}{nh^d}} - h^\alpha\right) - Dh\right\}, \tag{14}$$

where $C\left(\frac{\log(n)}{\sqrt{nh^d}} + h^\alpha\right)$ comes from the L_∞ error bound of α-valid kernels as in (9) and Dh is due to (13). As $h \to 0$, the term Dh dominates the bias term Ch^α, so that the optimal choice of h is of the order $h \asymp \left(\frac{\log n}{n}\right)^{1/(2+d)}$, which in turn yields a worse rate than (11) when $\alpha > 1$. What is more, $\|\nabla p(X_j)\| > 0$ for any X_j away from the critical points of p. This would mean that

$$\min_{X_j \in \{\widehat{p}_h \geq \lambda\}} \|\nabla p(X_j)\| > 0,$$

Then, as $h \to 0$, $\sup_{x \in B(X_j,h)} |p(x) - p(X_j)| \approx \|\nabla p(X_j)\| h = \Theta(h)$. Thus, the inclusions in (14) are tight, showing that the sub-optimal choice of h cannot be ruled out. We believe that this phenomenon is not specific to DBSCAN only, but applies more broadly to the the class of single-linkage-type clustering algorithms. That is, it seems to us that such algorithms are in general unable, like our DBSCAN-based procedure in Algorithm 1, to take advantage of a higher degree of smoothness of the underlying density.

The issue outlined above can be handled in more than one way. A possible solution, which is nearly trivial but impractical, is to deploy a computationally inefficient algorithm that assumes the ability to evaluate the connected components of the upper level set of $\widehat{p}_h$ exactly: see Algorithm 3 in the appendix. It is immediate to see that this approach produces optimal δ-consistency; see Corollary 3 in Algorithm 3. Unfortunately, this procedure will require evaluating $\widehat{p}_h$ on a fine grid, which is computationally infeasible even in small dimensions. The second, more interesting and novel solution which we describe next, is to further assume that p satisfies additional mild regularity conditions around the split levels. The conditions are of geometric and analytic nature and are reminiscent of low-noise type assumptions in classification. Under these conditions, the modified DBSCAN Algorithm 2 will achieve the optimal rate (11) while remaining computationally efficient. This finding, stated formally in Theorem 1 below, is the main result of this section.

Algorithm 2 The modified DBSCAN

```
INPUT: i.i.d sample $\{X_i\}_{i=1}^n$, a $\alpha$-valid kernel $K$ and $h > 0$.
  1. Compute $\{\widehat{p}_h(X_i), i = 1, \ldots, n\}$.
  2. For each $\lambda \geq 0$, construct a graph $\mathbb{G}_{h,\lambda}$ with node set

         $\widehat{D}(\lambda) = \{X_i : \widehat{p}_h(X_i) \geq \lambda\}$

     and edge set $\{(X_i, X_j) : X_i, X_j \in \widehat{D}(\lambda)$ and $\|X_i - X_j\| < 2h\}$.
  3. Compute $\mathbb{C}(h, \lambda)$, the graphical connected components of $\mathbb{G}_{h,\lambda}$.
OUTPUT: $\widehat{T}_n = \{\mathbb{C}(h, \lambda), \lambda \geq 0\}$.
```

Remark 1. *Despite its seemingly different form, Algorithm 2 is nearly identical to Algorithm 1. The only difference is the use of an α-valid kernel K instead of a spherical kernel. Furthermore, the procedures only require evaluating at most $n+1$ different graphs:*

$$\mathbb{G}_{h,0}, \mathbb{G}_{h,\widehat{p}_h(X_{\sigma_1})}, \ldots, \mathbb{G}_{h,\widehat{p}_h(X_{\sigma_n})},$$

where $(\sigma_1, \ldots, \sigma_n)$ is a permutation of $(1, \ldots, n)$ such that

$$\widehat{p}_h(X_{\sigma_1}) \leq \widehat{p}_h(X_{\sigma_2}) \leq \ldots \widehat{p}_h(X_{\sigma_n})$$

And, again just like with Algorithm 1, the connected components of each $\mathbb{C}(h, \lambda)$ can be easily evaluated by maintaining a union-find structure.

To formulate the the extra regularity conditions on the geometry of the density $p \in \Sigma(\alpha, L)$ around the split levels that guarantee optimality of the clustering Algorithm 2 we first recall some notions commonly used in the literature on level set and support estimation. Below, Ω denotes a generic subset of $\mathbb{R}^d$ of dimension d.

C1. (The Inner Cone Condition) The subset Ω satisfies the inner cone conditions if ihere exist constants $r_I, c_I > 0$ such that, for any $0 \leq r \leq r_I$ and $x \in \Omega$,

$$\mathcal{L}(B(x, r) \cap \Omega) \geq c_I V_d r^d,$$

where $\mathcal{L}$ denotes the Lebesgue measure of $\mathbb{R}^d$.

C2. (The Covering Condition) The subset Ω satisfies the covering condition if there exists a constant C_I such that, for any $0 < r \leq r_I$, there exists a collection of points $\mathcal{N}_r \subset \Omega$ such that $card(\mathcal{N}_r) \leq C_I r^{-d}$ and

$$\bigcup_{y \in \mathcal{N}_r} B(y, r) \supset \Omega.$$

Both assumptions **C1, C2** are rather mild. If Ω is a compact manifold of dimension $b \leq d$ with piecewise Lipschitz boundary, both assumptions are automatically verified with the dimension d replaced by the intrinsic dimension b. (see, e.g. Do Carmo, 1992). The inner cone condition **C1** is used in Korostelev and Tsybakov (1993) and is well-known as the standard condition (Cuevas, 2009) or, more recently, the (a, b) condition of (Chazal et al., 2015). It is essentially equivalent to the level set regularity condition [B] in Singh et al. (2009). The covering condition **C2** holds automatically if Ω is compact. See, e.g., Rinaldo and Wasserman (2010) and Balakrishnan et al. (2012).

Since $p \in \Sigma(L, \alpha)$ with $\alpha > 1$, any level set $\{p \geq \lambda\}$ is a union of connected d dimensional manifolds with C^1 boundary. Therefore it is natural to require both **C1** and **C2** to hold simultaneously for all the upper level-sets of p *right above the split levels.* Specifically, we will assume the following.

C. There exists a $\delta_0 > 0$ such that, for any split level λ^* of p and any $0 < \delta \leq \delta_0$, the set $\{x \colon p(x) \geq \lambda^* + \delta\}$ satisfies conditions **C1** and **C2** with constants r_I,c_I and C_I only depending on p.

We also need the connected components of the upper level sets right above the split levels to satisfy a low-noise condition as follows.

S(α). There exist positive constants δ_S and c_S such that, for each split level λ^* of the density p, the following holds. Let $\{\mathcal{C}_k\}_{k=1}^K$ be the connected components of $\{x\colon p(x) > \lambda^*\}$. Then,

$$\min_{k \neq k'} d(\mathcal{C}_k \cap \{p \geq \lambda^* + \delta\}, \mathcal{C}_{k'} \cap \{p \geq \lambda^* + \delta\}) \geq c_S \delta^{1/\alpha}, \quad \forall \delta \in (0, \delta_S]. \tag{15}$$

Condition **S**(α) constrains the behavior of the density only around the split levels. It is a fairly common assumption in the literature: it coincides with the *separation exponent* condition of Steinwart (2015) (see Definition 4.2 therein), which quantifies the separation of distinct connected components right above the split levels. Furthermore, **S**(α) is implied by *the local density regularity* conditions of Singh et al. (2009), which in turn is used in Jiang (2017a) to define the *β-regularity* condition for cluster separation. Steinwart (2015) provides several specific examples of densities satisfying the **S**(α) condition. In fact, we prove that conditions **C** and **S**(α) are verified in a large non-parametric class of functions. This class consists of Morse density functions, which are widely used in the density based clustering and mode estimation and topological data analysis; see, e.g., Chacón (2015), Arias-Castro et al. (2016) and references therein. We recall that a function p is Morse if all its critical points have a non-degenerate Hessian. An equivalent and more intuitive condition is that p behaves like a quadratic function around its critical points.

Proposition 2. *Suppose $p : \mathbb{R}^d \to \mathbb{R}$ is a Morse function. Then p satisfies* **C** *and* **S(2)**.

Another interesting class of density functions satisfying conditions **C** and **S**(α) can be obtained as follows. Let $\alpha \geq 2$ be any integer and $f_1 : [0,1] \to \mathbb{R}$ be such that $f_1(x) = (x-2)^\alpha$. Then, there exists a polynomial f_2 of degree α such that the function on $\mathbb{R}$ defined point-wise as

$$f(x) = \begin{cases} f_1(x), & x \in [1,2] \\ f_2(x), & x \in [0,1] \\ 0, \text{ otherwise}, \end{cases}$$

has continuous derivatives up to order $\alpha - 1$ and is such that $f(0) = f'(0) =, \ldots, = f^{(\alpha-1)} = 0$. When $\alpha = 3$, f is a natural spline. For any integer $d \geq 1$, let $F : \mathbb{R}^d \to \mathbb{R}$ be such that $F(x) = f(\|x\|_2)$. Then $F \in \Sigma(\alpha, L)$. Denote $x_0 = (2, 0, \ldots, 0)$. Let $G(x) = F(x - x_0) + F(x + x_0)$. It is easy to see that for any $0 < \delta \leq 1$

$$\{G(x) \geq \delta\} = B(x_0, 2 - \delta^{1/\alpha}) \cup B(-x_0, 2 - \delta^{1/\alpha}).$$

As a result, conditions **C** and **S**(α) are trivially satisfied in this simple case.

Our main result of this section is to prove that that the conclusion of Corollary 2 still holds for $\alpha > 1$, provided that the conditions **C** and **S**(α) are met.

Theorem 1. *Let $p \in \Sigma(\alpha > 1, L)$ be any density function with compact and connected support and finitely many split levels. Suppose that conditions* **C** *and* **S**(α) *hold for p. If*

$h \asymp \left(\frac{\log(n)}{n}\right)^{1/(2\alpha+d)}$), then, with probability at least $1 - \frac{1}{n} - O(h^{-d}\exp(-cn^{\alpha/(2\alpha+d)}))$, the cluster tree returned by the modified DBSCAN Algorithm 2 is δ-consistent with rate

$$\delta_n \geq 2a_n + (4h/c_S)^\alpha$$

where c is a constant that depends on p only, $a_n = C_1\sqrt{\frac{(\log n + \log(1/h))}{nh^d}} + C_2 h^\alpha$ is the right hand side of the inequality in (9) and c_S is defined in **C**.

The choice of the parameter h in Theorem 1 yields that Algorithm 2 is δ consistent with rate given by (11).

4.5. Lower bounds

Next, we show that the consistent rates of the DBSCAN algorithm derived in the previous sections are nearly minimax optimal, save for a $\log(n)$ term. We point out that the lower bound results by Chaudhuri et al. (2014) are not directly applicable to our problem, since they rely on discontinuous densities.

Theorem 2. *Suppose $d \geq 1$ and $\alpha > 0$. There exists a finite family $\mathcal{F}$ of d-dimensional probability density functions belonging to the Hölder class $\Sigma(L, \alpha)$ satisfying the conditions* **C**, **S**(α) *and uniformly bounded from above by C_0, and a constant $\mathcal{K}$, depending on L and α, such that when*

$$n \geq \frac{4^d 8\log(32)}{V_d} \quad \textit{and} \quad \delta \leq \min\left\{\left(\frac{\mathcal{K}}{16^\alpha (7C_0)^{\alpha/d}}\right), \|p\|_\infty/(2^{d/2+1})\right\},$$

where V_d denote the volume of a d dimensional ball, the following holds. If cluster tree estimator is $(\delta_n, 1/4)$-accurate when presented with an i.i.d. sample from a density function in $\mathcal{F}$, then it must be the case that

$$n \geq \frac{C_0 \mathcal{K}^{d/\alpha}}{C\delta_n^{2+d/\alpha}}, \tag{16}$$

for some constant C only depends on d.

Therefore, with the constant C_0 in the previous theorem and the dimension d fixed, the bounds obtained in Theorem 1 and Corollary 2 match the minimax bound in (16), up to a $\log(n)$ factor. Thus, together they show that, up to log factors, the optimal rate for δ-consistency of density functions in $\Sigma(L, \alpha)$ is of order $\left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)}$.

Interestingly, the minimax clustering rates we derived match the rates for estimation of a Hölder density p under L_∞ norm; see Section 4.1. While this result may not be entirely surprising in light of the findings of, e.g., Eldridge et al. (2015) and Kim et al. (2016), such a a connection has never been formally established, to the best of our knowledge. In particular, our results seem to settle, at least for the class of Hölder-continuous densities and with respect to the criterion of δ-consistency, a long-standing open problem of how density-based clustering compares to density estimation: both problems exhibit the same degree of statistical difficulty.

4.6. Consistent Estimate of The Split Levels

In this section, we present a simple pruning strategy, leading to consistent estimators of the split levels of the the density. While pruning strategies and consistent estimation of split levels have been considered by several authors, such as Sriperumbudur and Steinwart (2012); Steinwart (2015), Chaudhuri et al. (2014) and Jiang (2017a), the existing results do not yield error bounds that depend on the degree of smoothness α for density $p \in \Sigma(L, \alpha)$ with $\alpha > 1$.

The following definition provides a way to identify significant split levels in the cluster tree estimator returned by Algorithm 2.

Definition 6. *Let $\Delta > 0$. The random variable $\widehat{\lambda^*} \in (0, \infty)$ is said to be a Δ-significant split level of the cluster tree estimator if there exist two data points $X_i, X_j \in D(\widehat{\lambda^*} + \Delta)$ such that*

$$\widehat{\lambda^*} = \sup\{\lambda > 0 : X_i \textit{ and } X_j \textit{ are in the same connected component of } \mathbb{C}(h, \lambda).\} \tag{17}$$

Below, we show that there is a one to one correspondence between Δ-significant split levels of the modified DBSCAN cluster tree estimator from Algorithm 2 and the split levels of the population density under a slightly stronger covering condition than condition **C** given above. Specifically, we assume the following.

C'. There exists a constant $\delta_0 > 0$ such that, for any split level λ^* of p and any $\delta \in \mathbb{R}$ with $|\delta| \leq \delta_0$, $\{p \geq \lambda^* + \delta\}$ satisfies conditions **C1** and **C2** with constants r_I,c_I and C_I only depending on p.

The only difference between **C** and **C'** is that while condition **C** assumes some regularity of p only above split levels, **C'** requires the same type of regularity *around* split levels.

Proposition 3. *Suppose condition* **C'** *and* $\mathbf{S}(\alpha)$ *hold. Let $\Delta = 2a_n + (4h/c_S)^\alpha$ where a_n is defined in* (9) *and $h = C_1 n^{-1/(2\alpha+d)}$. Suppose p has finitely many split levels. Then, with probability at least $1 - 1/n - O(h^{-d}\exp(-cn^{\alpha/(2\alpha+d)}))$, the following additional results hold:*
1. *Let λ^* be a split level of the density p. Suppose $\mathcal{C}$ and $\mathcal{C}'$ are two open sets splitting at λ^* (see Definition 5) and that*

$$\min\{P\left(\mathcal{C} \cap \{p \geq \lambda^* + 2\Delta\}\right), P\left(\mathcal{C}' \cap \{p \geq \lambda^* + 2\Delta\}\right)\} > 0. \tag{18}$$

Then, there exists a Δ-significant split level $\widehat{\lambda^*}$ of the cluster tree estimator returned by the modified DBSCAN such that

$$|\lambda^* - \widehat{\lambda^*}| \leq \Delta. \tag{19}$$

2. *Conversely, suppose that $\widehat{\lambda^*}$ is a Δ-significant split level of the cluster tree estimator. Then there exists a split level λ^* of p such that*

$$|\lambda^* - \widehat{\lambda^*}| \leq \Delta. \tag{20}$$

Proposition 3 says that, with high probability, every Δ-split level corresponds to a density split level and that conversely, any split level of the density p can be found if we have enough data. To prune the cluster tree returned by the modified DBSCAN algorithm, it suffices to remove all the split levels that are not Δ significant.

5. Densities with Gaps

We now consider the particular scenario where the density p exhibiting a jump discontinuity in such a way that, for all levels λ in a given interval of length ϵ, the upper level sets $\{x\colon p(x) \geq \lambda\}$ do not change. The value of ϵ is referred to as the *gap size.* We provide a formal definition next, which we formulate within the general measure-theoretic language of Steinwart (2015) since the underlying densities are not continuous. We recall that $\mathcal{L}$ denotes the Lebesgue measure on $\mathbb{R}^d$.

Definition 7 (Distribution with a gap)**.** *Let P a probability measure on $\mathbb{R}^d$, absolutely continuous with respect to the Lebesgue measure. For any $\lambda^* > 0$, let S_{λ^*} be the support of the sub-probability measure*

$$A \mapsto P(A \cap \{x\colon p'(x) \geq \lambda^*\}), \quad \textit{A Lebesgue measurable,}$$

where p' is any density of P. Then, P is said to have a gap at λ_* of size ϵ, where $0 < \epsilon < \lambda^*$, if $P(S_{\lambda^*}) > 0$ with $\mathcal{L}(\partial S_{\lambda^*}) = 0$ and

$$S_{\lambda_* - \eta} \backslash S_{\lambda_*} = \emptyset, \quad \forall \eta \in (0, \epsilon). \tag{21}$$

We impose the condition that $\mathcal{L}(\partial S_{\lambda^*}) = 0$ in order to avoid pathological cases.

The above definition is independent of the choice of the density of P. At the same time, it also implies that P admits a Lebesgue density p such that

$$S_{\lambda^*} = \mathrm{cl}\left(\{x\colon p(x) \geq \lambda^*\}\right) \quad \text{and} \quad \mathrm{cl}(S^c_{\lambda^*}) = \mathrm{cl}\left(\{x\colon p(x) \leq \lambda^* - \epsilon\}\right). \tag{22}$$

Here, given any set A, $\mathrm{cl}(A)$ denotes the closure of A. Indeed, it is not hard to see that, if p' is any Lebesgue density of P, then the function

$$p(x) = \begin{cases} \max\{p'(x), \lambda^*\} & x \in S_{\lambda^*} \\ \min\{p'(x), \lambda^* - \epsilon\} & x \in S^c_{\lambda^*} \end{cases}$$

is also a density of P, and satisfies (22). Thus, with a slight abuse of notation, we may also speak of "the" density p even in this case, with the understanding that we are referring to any density of P for which (22) holds. See Figure 2 for an illustration.

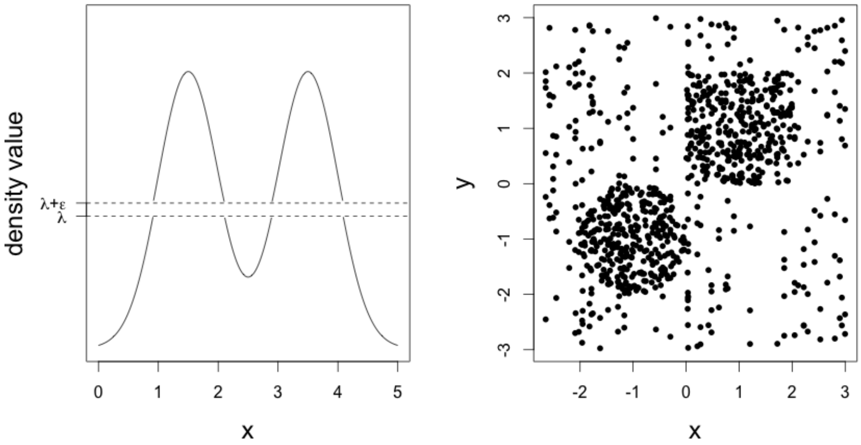

Figure 2: The left plot depicts a one dimensional density with gap of size ϵ at level λ. It is clear that $\{\lambda < p \leq \lambda + \epsilon\}$ is an empty set. The right plot depicts 500 i.i.d sampling from a two dimensional density with a gap. It is clear that the density is low in the background and high in the disk centered at (-1,-1) and in the square centered at (1,1). Finding the samples points with low density values can be thought of as outliers detection in this case.

Though fairly restrictive, the scenario of a distribution with a gap is quite interesting for the purpose of both clustering and level set estimation. Indeed, this situation encompasses the ideal clustering scenario, depicted as examples in Figures 1 and 5 in the original DBSCAN paper Ester et al. (1996), of a piece-wise constant density that is low everywhere on its support with the exception of a few connected, full-dimensional regions, or clusters, where it is higher by a certain amount (in our case the gap ϵ). The size of the gap parameter ϵ and the minimal distance among clusters both affect the difficulty of the clustering task, which becomes harder as the parameters get smaller.

To formally capture the dependence on the distance between clusters we set

$$S_{\lambda^*} = \bigcup_{i=1}^{I} \mathcal{C}_i,$$

where $(\mathcal{C}_1, \ldots, \mathcal{C}_I)$ are disjoint, connected, (necessarily) closed sets of positive P measure and let

$$\sigma = \min_{i \neq j} \operatorname{dist}(\mathcal{C}_i, \mathcal{C}_j) > 0 \tag{23}$$

be the minimal distance between them. The separation parameter σ captures an aspect of the intrinsic difficulty of the clustering task that is complementary to the one quantified by gap parameter ϵ: clusters that are at a small distance σ from each other are hard to separate, for any given value of ϵ. In our analysis, we let both the gap parameter ϵ and the separation parameter σ vary with n (though we do not make this dependence in out notation for ease of readability), thus allowing for harder clustering problems as a function of the sample size.

In the next simple result, we show that a flat version of the vanilla DBSCAN algorithm given in Algorithm 1, with suitable choices for the input parameters, can optimally estimate the clusters at λ^* at a rate that depend explicitly on both ϵ and σ.

Proposition 4. *Let $\{X_1, \dots, X_n\}$ be an i.i.d. sample from a probability distribution P that has a gap of size ϵ at level λ^*. Set $a_n = C_1\sqrt{\frac{\log(n)+\log(1/h)}{nh^d}}$ as in (7) and suppose the input parameters h and k of the DBSCAN algorithm are such that*

$$\sigma/4 \geq h \geq C\left(\frac{\log(n)}{n\epsilon^2}\right)^{1/d} \qquad \text{and} \quad k = \lceil nh^d V_d \lambda \rceil, \tag{24}$$

for any $C > 0$ such that $2a_n < \epsilon$ and any λ in $(\lambda^* - \epsilon + a_n, \lambda^* - a_n]$. Then with probability at least $1 - 1/n$:

i. simultaneously over all connected sets A such that $A_{2h} \subset \mathcal{C}_i$, for some i, all the sample points in A, if any, belong to the same connected component of $\mathbb{G}_{k,h}$;

ii. simultaneously over all connected sets A and A' such that $A_{2h} \subset \mathcal{C}_i$ and $A'_{2h} \subset \mathcal{C}_j$, for some $i \neq j$, the sample points in A and A', if any, belong to distinct connected components of $\mathbb{G}_{k,h}$.

The definition of A_{2h} and A_{-2h} can be found in (1). Notice that the above results hold true for all n large enough such that $a_n < \epsilon/2$. Proposition 4 implies the DBSCAN algorithm will yield clustering consistency, in the sense of Chaudhuri et al. (2014), provided that its input parameters fulfill (24) holds. In particular, this result requires that the sample size relates to the gap parameter ϵ and the separation parameter σ according to the inequality

$$n \geq C\frac{1}{\epsilon^2\sigma^d},$$

for some constant C, depending on d. In fact, such scaling is nearly minimax optimal: no other clustering algorithms can guarantee cluster consistency under the same assumptions and with a better sample complexity as a function of both ϵ and σ. This results follows from the lower bound guarantee given in Theorem VI.1 of Chaudhuri et al. (2014), where we take notice that the parameters σ and ϵ have different, though related, meaning; see Appendix C.4. In fact, the results in Proposition 4 can be further extended to hold over more general settings of arbitrary densities; see Appendix D.

We conclude by noting that the gap size ϵ and the separation parameters σ quantify two very separate notions of intrinsic difficulty of clustering that are unrelated to each other, and the clustering problem becomes impossible whenever either one of them becomes so small to violate the lower bound (53), regardless of the other. In particular, it is easy to give examples in which the clustering task is impossible to solve because ϵ is too small even if σ is large, and the other way around. As a result, the overall hardness of the clustering problem around the gap is a combination of these two parameters. This is in contrast with the settings considered earlier, where, due to the smoothness of the underlying density, only one parameter is sufficient to capture separation among clusters.

5.1. The Devroye-Wise estimator of the Level Sets

The assumption of a density with gap allows us to carry out a further analysis of the DBSCAN algorithm, showing that it is also minimax optimal for estimating the level set itself S_{λ^*}. For this purpose, the DBSCAN algorithm reduces to the renown Devroye-Wise estimator: see Devroye and Wise (1980). Below we provide a novel, sharper analysis of this estimator, where we allow the size of the gap ϵ to decrease with n, and demonstrate that its rate-optimal (again, we will not explicitly express this dependence in our notation for simplicity). To the best of our knowledge, such scaling has not been previously established.

Recall that the DBSCAN algorithm with inputs k and h outputs a set of nodes $\mathbb{G}_{h,k}$. One then may construct the estimator

$$\widehat{S}_h = \bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h) \tag{25}$$

comprised of a union of balls around such points, and use it as an estimator of the corresponding high density region $S = S_{\lambda^*} = \cup_{i=1}^{I} \mathcal{C}_i$ consisting of all the clusters.

We measure the performance of any estimator $\widehat{S}$ with the Lebesgue measure of its symmetric difference with S:

$$\mathcal{L}\left(S\Delta\widehat{S}\right) = \mathcal{L}\left(S \cap \widehat{S}^c\right) + \mathcal{L}\left(S^c \cap \widehat{S}\right).$$

We will in addition impose the following condition:

R. (Level set regularity). There exists constants $h_0 > 0$ and $C_0 > 0$ such that, for all $h \in (0, h_0)$,

$$\mathcal{L}\left(S_h \setminus S_{-h}\right) \leq C_0 h,$$

where S_h and S_{-h} are defined in (1).

Condition **R** is very mild. Indeed, if ∂S is C^2, then the set $N_h := S_h \backslash S_{-h}$ is the tubular neighborhood (see, e.g. Hirsch, 2012) of ∂S in $\mathbb{R}^d$. In particular, every compact domain in $\mathbb{R}^d$ with C^2 boundary satisfies condition **R**. In this case $C_0 \preceq V_{d-1}|\partial S|$, where $|\partial S|$ denotes the surface volume of S. We also note that **R** is equivalent to the "smooth boundary" condition in Steinwart (2015).

Proposition 5. *Let $\{X_1, \ldots, X_n\}$ be an i.i.d. sample from a probability distribution P that has a gap of size ϵ at level λ^*. Let $a_n = C\sqrt{\frac{\log(n)+\log(1/h)}{nh^d}}$ be defined as in (7). Suppose the input parameters (h, k) of the DBSCAN algorithm satisfy*

$$h_0 \geq h \geq C\left(\frac{\log(n)}{n\epsilon^2}\right)^{1/d} \quad \textit{and} \quad k = \lceil nh^d, V_d\lambda \rceil \tag{26}$$

for any $C > 0$ such that $2a_n < \epsilon$ and any λ in $(\lambda^* - \epsilon + a_n, \lambda^* - a_n]$. Then with probability at least $1 - 1/n$,

$$\mathcal{L}(S \triangle \widehat{S}_h) \leq 2C_0 h,$$

*where C_0 is defined in **R** and*

$$\widehat{S}_h = \bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h) \tag{27}$$

Similarly to Proposition 4, the above results hold true for all n large enough such that $a_n < \epsilon/2$. If $P(X \in S) = 1$, the level set estimator $\widehat{S}_h$ is also a support estimator and $\epsilon = \inf_{x \in S} p(x)$. In this case, Proposition 5 says that if the lower bound on the density vanishes no faster than $O(n^{-1/2})$, then support estimation is still possible.

Below we show that the error bound given in Proposition 5 is minimax optimal up to log factors. Consider $\mathcal{P}^n(h_0, \epsilon)$, the class of probability distribution of n i.i.d. random vectors in $\mathbb{R}^d$ whose common density exhibit a gap of size ϵ such that condition (22) holds, and satisfying condition $\mathbf{R}$ with parameter $h_0 > 0$. Then Proposition 5 shows that, for all n large enough,

$$\sup_{P \in \mathcal{P}^n(h_0,\epsilon)} \mathbb{E}_P \left(S \triangle \widehat{S}_h \right) = O\left(\left(\frac{\log(n)}{n\epsilon^2} \right)^{1/d} \right),$$

provided that h is of the order $\left(\frac{\log(n)}{n\epsilon^2}\right)^{1/d}$. Our next result provides a nearly-matching lower bound.

Proposition 6. *There exist constants h_0 and c, depending only on d such that for any $\epsilon \leq 1/4$, for all n large enough here exist probability distributions $\{P_1, \ldots, P_M\}$ in $\mathcal{P}^n(h_0, \epsilon)$ such that*

$$\inf_{\widehat{S}} \sup_{i=1,\ldots,M} \mathbb{E}_{P_i} \left(\mathcal{L}(\widehat{S} \triangle S) \right) \geq c \min \left\{ \left(\frac{1}{n\epsilon^2} \right)^{1/d}, 1 \right\},$$

where the infimum is with respect to all estimators of S.

Thus, if $\frac{\log(n)}{n\epsilon^2} \to 0$ as $n \to \infty$ (so that condition (26) is eventually satisfied), then the bounds given in Proposition 5 and Proposition 6 match, up to a $\log(n)$ factor. That is, with suitable choice of input, DBSCAN can optimally estimate the level set S at the gap.

The performance of the Devroye-Wise estimator is a well-established topic in the literature: see, e.g., Theorem 4 and 5 in Cuevas and Rodríguez-Casal (2004). Our contribution in this regard is two fold: we allow for an explicit dependence on the gap size parameter ϵ and deliver minimax lower bounds. Our rate of convergence confirms the intuition that a smaller gap size leads to a harder estimation problem.

6. Discussion

In this article we propose a new notion of consistency for estimating the clustering structure under various conditions. Our analysis shows that the DBSCAN algorithm is minimax optimal. Interestingly, the rates match, up to log terms, minimax rates for density estimation in the supreme norm for Höloder smooth densities. In particular, our results provide a complete, rigorous justification to the plausible belief, commonly held in density-based clustering, that clustering is as difficult as density estimation. In the rest of the discussion section, we will compare our notion of separation with other existing ones in the literature. For the sake of exposition, we will follow the convention used in much of the literature on density-based clustering of assuming that the cluster tree of the data generating distribution in fact corresponds to the hierarchy of the upper level sets of a canonical density p. As explained in Steinwart (2015), this definition is in general not well-posed, since different densities will yield different trees.

6.1. Hartigan consistency in Hartigan (1981)

We follow Chaudhuri and Dasgupta (2010) and Eldridge et al. (2015) in defining Hartigan consistency in terms of the density cluster tree.

Definition 8 (Hartigan consistency)**.** *Let $\widehat{T}_n$ be a cluster tree estimator constructed from i.i.d. data $\{X_i\}_{i=1}^n$ from a disribution P with Lebesgue density p. For any pair of subsets A and A', let denote A_n and A'_n be the smallest clusters of $\widehat{T}_n$ containing $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$, respectively. The cluster tree estimator $\widehat{T}_n$ is Hartigan consistent if, for any pair of sets A and A' belonging to distinct connected components of $\{x : p(x) \geq \lambda\}$ for some λ, $P(A_n \cap A'_n = \emptyset) \to 1$ as $n \to \infty$.*

It is immediate from Definition 3 that a δ-consistent cluster tree is also Hartigan consistent. While Hartigan consistency is a simple form of *point-wise* cluster tree consistency, which holds for each fixed pairs of disjoint clusters, δ-consistency is a stronger guarantee, as it yields *uniform* consistency over all δ-separated clusters and, furthermore, gives consistency rates depending on the value of the separation parameter δ.

6.2. Comparison with the Merge distortion metric

The notion of δ-separation is closely related to the notion of *merge distance* introduced by Eldridge et al. (2015), which we present next.

Definition 9. *Let p and q be Lebesgue densities in $\mathbb{R}^d$ and let T_p and T_q be the corresponding cluster density trees. The merge distortion distance between T_p and T_q is defined as*

$$d_M(T_p, T_q) = \sup_{x,y \in \mathbb{R}^d} |m_p(x,y) - m_q(x,y)|,$$

where, for a Lebesgue density p,

$$m_p(x,y) = \sup\{\lambda > 0 \in \mathbb{R} : \textit{there exists } C \in T_p(\lambda) \textit{ such that } \{x,y\} \subset C\}.$$

The original definition of merge distortion metric is, in fact, more general but, when specialized to our settings, reduces to the one given above.

The merge distortion distance is closely related to the L_∞ distance between densities. In fact, by Theorem 17 in Eldridge et al. (2015), $d_M(T_p, T_q) \leq \|p - q\|_\infty$, so that, if $\{p_n\}_n$ is a sequence of Lebesgue densities, then $\|p_n - p\| \to 0$ implies that $d_M(T_{p_n}, T_p) \to 0$. In fact, Lemma 1 in Appendix F of Kim et al. (2016) shows that, if p and q are continuous, then $d_M(T_p, T_q) = \|p - q\|_\infty$. As a result, for the class of continuous (and, in particular, Hölder smooth) densities, cluster consistency in the merge distortion distance is equivalent to cluster consistency based on the δ-separation criterion, which in turn is equivalent to estimation consistency of the underlying density in the L_∞ norm.

The above statement immediately applies to the näive cluster tree estimator $T_{\widehat{p}_{h_n}}$ built using the level sets of any density estimator $\widehat{p}_{h_n}$ that is continuous and, as $h_n \to \infty$, consistent in the L_∞ norm. Such estimator is of course computationally unfeasible even in low dimension. In fact, the DBSCAN-based procedures described above in Algorithms 1 and 2, which are applicable in high-dimensional settings, are also consistent in the merge-distortion metric. To see this, and following the arguments of Eldridge et al. (2015), it is sufficient to demonstrate the properties of minimality and separation, as defined in that reference, for the DBSCAN cluster-tree estimators. The separation property, which prevents the emergence of false clusters or over-segmentation, follows directly from the definition of δ-separation; see also the pruning results of Section 4.6. On the other hand, minimality avoids the occurrence of improper nesting and holds for the DBSCAN procedures we consider here in virtue of Lemma 1 and Lemma 6. The fact that our algorithms produce cluster tree estimators that are consistent in the merge distance should not be surprising, since δ-consistency is directly tied to consistency for density estimation in in L_∞. As shown in Eldridge et al. (2015), the robust single-linkage clustering algorithm of Chaudhuri et al. (2014) is also consistent in the merge distance.

6.3. Comparison with the (ϵ, σ)-separation criterion of Chaudhuri et al. (2014)

The criterion of δ-separation we introduce in this paper is most useful when studying smooth densities. Nonetheless, it will be helpful to compare it to the notion of (ϵ, σ)-separation defined in Chaudhuri et al. (2014), which is applicable to arbitrary densities.

Definition 10 ((ϵ, σ)-separation criterion in Chaudhuri et al. (2014))**.**

1. *Let f be a density supported on $X \subset \mathbb{R}^d$. We say that $A, A' \subset X$ are (ϵ, σ)-separated if there exists $S \subset X$ (the separator set) such that (i) any path in X from A to A' intersects S, and (ii) $\sup_{x \in S_\sigma} f(x) < (1-\epsilon) \inf_{x \in A_\sigma \cup A'_\sigma} f(x)$.*
2. *Suppose an i.i.d samples $\{X_i\}_{i=1}^n$ is given. An estimate of the cluster tree is said to be (ϵ, σ) consistent if for any pair A and A' being (ϵ, σ) separated, the smallest cluster containing $A \cap \{X_i\}_{i=1}^n$ is disjoint from the smallest cluster containing $A' \cap \{X_i\}_{i=1}^n$.*

In the following result we make a straightforward connection between the δ-separation and (ϵ, σ)-separation.

Lemma 3. *Assume that $p \in \Sigma(L, \alpha)$ with $\alpha \leq 1$ and that A and A' are δ-separated. Then, A and A' are (ϵ, σ)-separated with*

$$S = \{x \colon p(x) \leq \lambda - \delta\}, \quad \epsilon = \delta/(3\lambda) \quad \textit{and} \quad \sigma^\alpha = \delta/(3L), \tag{28}$$

where $\lambda = \inf_{z \in A \cup A'} p(z)$.

Proof of lemma 3. Denote $\lambda = \inf_{z \in A \cup A'} p(z)$. Suppose for the sake of contradiction that there is a path l connects A and A' and that $l \cap \{p \leq \lambda - \delta\} = \emptyset$. Then by the continuity of p and the compactness of l, there exist $\gamma > 0$ such that $l \subset \{p \geq \lambda - \delta + \gamma\}$. Thus A and A' belongs to the same path connected component of $\{p \geq \lambda - \delta + \gamma\}$. Since $\{p \geq \lambda - \delta + \gamma\} \subset \{p > \lambda - \delta\}$, A and A' belongs to the same path connected component of $\{p > \lambda\}$. Since $\{p > \lambda\}$ is an open set, A and A' be belongs to the same connected component of $\{p > \lambda\}$. This is a contradiction.

Let $\sigma^\alpha = \delta/(3L)$ and $\epsilon = \delta/3$, then for any $x \in S_\sigma$, $p(x) \leq \lambda - \delta + L\sigma^\alpha = \lambda - 2\delta/3$. Similarly if $x \in A_\sigma \cup A'_\sigma$, $p(x) \geq \lambda - L\sigma^\alpha = \lambda - \delta/3$. Thus

$$(1-\epsilon) \inf_{x \in A_\sigma \cup A'_\sigma} f(x) > (1-\epsilon)\lambda - \delta/3 = \lambda - 2\delta/3 > \sup_{x \in S_\sigma} f(x)$$

□

According to the separation criterion in Definition 10, two clusters can be (ϵ, σ)-separated for many the values of ϵ and σ. In particular, by taking the separator set to be larger, it is easy to produce examples of δ-separated clusters that are also (ϵ, σ)-separated such that δ is big but σ is small. This is simply because σ is heavily associated with S. And conversely, by taking an almost flat density function, it is possible to have a very large σ and very small δ.

We remark that when $\alpha > 1$, there is no obvious relationship between the parameter (σ, ϵ) in Definition 10 and δ in Definition 3 as that in Lemma 3. For $\alpha > 1$, while $p \in \Sigma(L, \alpha)$ implies that p is Lipschitz continuous, the Lipschitz constant in this case does not depend on L and α in a simple manner. As a result, the parameter σ, representing the distance between connected components of upper level sets of p, is not straightforwardly related to δ.

References

Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17(1):1487–1514, 2016.

Amparo BaÍllo, Antonio Cuevas, and Ana Justel. Set estimation and nonparametric detection. *Canadian Journal of Statistics*, 28(4):765–782, 2000.

Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, 2012.

José E Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.

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

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

Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16(1):3603–3635, 2015.

Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112 (520):1684–1696, 2017.

Antonio Cuevas. Set estimation: Another bridge between statistics and geometry. *Boletín de Estadística e Investigación Operativa*, 25(2):71–85, 2009.

Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, pages 2300–2312, 1997.

Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36(2):340–354, 2004.

Luc Devroye and Gary L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.

Manfredo Perdigao Do Carmo. *Riemannian geometry*. Birkhauser, 1992.

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

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Kdd, pages 226–231, 1996.

Junhao Gan and Yufei Tao. Dbscan revisited: Mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, 2015.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'IHP Probabilités et statistiques*, 38:907–921, 2002.

John A Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.

Morris W Hirsch. *Differential topology*, volume 33. Springer Science & Business Media, 2012.

Jennifer Jang and Heinrich Jiang. Dbscan++: Towards fast and scalable density clustering. *arXiv preprint arXiv:1810.13105*, 2018.

Heinrich Jiang. Density level set estimation on manifolds with dbscan. *arXiv preprint arXiv:1703.03503*, 2017a.

Heinrich Jiang. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pages 1694–1703, 2017b.

Heinrich Jiang, Jennifer Jang, and Ofir Nachum. Robustness guarantees for density clustering. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3342–3351, 2019.

Jisu Kim, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, and Larry Wasserman. Statistical inference for cluster trees. In *Advances in Neural Information Processing Systems 29*, pages 1839–1847. 2016.

Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3398–3407, 2019.

Jussi Klemelä. Complexity penalized support estimation. *Journal of multivariate analysis*, 88(2):274–297, 2004.

Jussi Klemelä. *Smoothing of Multivariate Data: Density Estimation and Visualization*. Wiley, 2009.

Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 1993.

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

James R Munkres. *Topology*. Prentice Hall, 2000.

Deborah Nolan and David Pollard. U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799, 1987.

Mathew D. Penrose. Single linkage clustering and continuum percolation. *Journal of Multivariate Analysis*, 53:94–109, 1995.

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

Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, pages 1154–1178, 2009.

Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, pages 2678–2722, 2010.

Alessandro Rinaldo, Aarti Singh, Rebecca Nugent, and Larry Wasserman. Stability of density-based clustering. *The Journal of Machine Learning Research*, 13:905–948, 2012.

Arrti Singh, Clayton Scott, Robert Nowak, and Aarti Singh. Adaptive hausdorff estimation of density level sets. *The Annals of Statistics*, 37:2760–2782, 2009.

Bharath K Sriperumbudur and Ingo Steinwart. Consistency and rates for clustering with dbscan. In *AISTATS*, pages 1090–1098, 2012.

Ingo Steinwart. Adaptive density level set clustering. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 703–738, 2011.

Ingo Steinwart. Fully adaptive density-based clustering. *The Annals of Statistics*, 43(5): 2132–2167, 2015.

Ingo Steinwart, Bharath K. Sriperumbudur, and Philipp Thomann. Adaptive clustering using kernel density estimators. arXiv preprint arXiv:1708.05254, 2017.

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

Werner Stuetzle and Rebecca Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2), 2010.

Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics, 2009.

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

Bingchen Wang, Chenglong Zhang, Lei Song, Lianhe Zhao, Yu Dou, and Zihao Yu. Design and optimization of dbscan algorithm based on cuda. *arXiv preprint arXiv:1506.02226*, 2015.

RM Willett and Robert D Nowak. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.

Appendix A. Topological Preliminaries

For completeness, we review the definition of connectedness from the general topology.

Definition 11 (Munkres (2000) Chapter 3). *Let U be any nonempty subset in $\mathbb{R}^d$. Then U is said to be connected, if, for every pair of open subsets A, A' of U such that $A \cup A' = U$, we have either $A = \emptyset$ or $A' = \emptyset$. The maximal connected subsets of U are called the connected components of U.*

We briefly explain why the connected components naturally introduce a hierarchical structure to the level sets of p. Let $\lambda_1 > \lambda_2$, so we have $\{p \geq \lambda_1\} \subset \{p \geq \lambda_2\}$.

- Suppose A is any subset of $\mathbb{R}^d$, and A belongs to the same connected component of $\{p \geq \lambda_1\}$. Then A is contained in the same connected component of $\{p \geq \lambda_2\}$.
- Suppose $A \cup A' \subset \{p \geq \lambda_1\}$ and they belong to distinct connected components of $\{p \geq \lambda_2\}$ Then A and A' are not contained in the same connected component of $\{p \geq \lambda_1\}$.

We also review a closed related concepts, which is call the path connectedness in general topology.

Definition 12. *We say that a subset $U \subset \mathbb{R}^d$ is path connected if for any $x, y \in U$, there exists a path continuous $\mathcal{P} : [0, 1] \to U$ such that $\mathcal{P}(0) = x$ and $\mathcal{P}(1) = y$.*

The main reason we introduce the path connectedness is that if U is an open set in $\mathbb{R}^d$, then U is connected if and only if it is path connected. Therefore a simple but useful consequence is that for any λ, the connected components of $\{p > \lambda\}$ are also the path connected components.

We will repeatedly use these topological properties in our analysis without further mentioning. The proofs of them are omitted and can be found in Munkres (2000) or any other books on general topology.

Appendix B. Proofs from Section 4

We begin by justifying (6). Since this is a well known result, we simply use a result of Sriperumbudur and Steinwart (2012). We will assume the following condition for the kernel K which is fairly standard in the non-parametric literature.

VC. The kernel $K : \mathbb{R}^d \to \mathbb{R}$ has bounded support and integrates to 1. Let $\mathcal{F}$ be the class of functions of the form

$$z \in \mathbb{R}^d \mapsto K\left(x - z\right), \quad z \in \mathbb{R}^d.$$

Then, $\mathcal{F}$ is a uniformly bounded VC class: there exist positive constants A and v such that

$$\sup_P \mathcal{N}(\mathcal{F}, L^2(P), \epsilon\|F\|_{L^2(P)}) \le (A/\epsilon)^v,$$

where $\mathcal{N}(T, d, \epsilon)$ denotes the ϵ-covering number of the metric space (T, d), F is the envelope function of $\mathcal{F}$ and the sup is taken over the set of all probability measures on $\mathbb{R}^d$. The constants A and v are called the VC characteristics of the kernel.

The assumption VC holds for a large class of kernels, including any compact supported polynomial kernel and the Gaussian kernel. See Nolan and Pollard (1987) and Giné and Guillou (2002).

Proposition 7 (Sriperumbudur and Steinwart (2012)). *Let P be the probability measure on $\mathbb{R}^d$ with Lebesgue density bounded by $\|p\|_\infty$ and assume that the kernel K belongs to $L^\infty(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ satisfies the VC assumption. Then for any $\gamma > 0$ and $h > 0$, there exists an absolute constant C depending on the VC characteristic of K such that, with probability no smaller than $1 - e^{-\gamma}$,*

$$\|p_h - \widehat{p}_h\|_\infty \le \frac{C}{nh^d}\left(\gamma + v\log\frac{2A}{\sqrt{h^d\|p\|_\infty\|K\|_2^2}}\right) + C\sqrt{\frac{2\|p\|_\infty}{nh^d}}\left(\gamma\|K\|_\infty^2 + v\|K\|_2^2\log\frac{2A}{\sqrt{h^d\|p\|_\infty\|K\|_2^2}}\right)$$

Lemma 4. *Suppose A' and A' are two clusters of $T_p(\lambda)$ and they are δ-separated. Then their merge height $m_f(A, A')$ satisfies*

$$m_f(A, A') \le \lambda - \delta.$$

Proof. By Definition 2, there A and A' belong distinct connected components of $\{p > \lambda - \delta\}$. For the sake of contradiction, suppose

$$m_f(A, A') > \lambda - \delta.$$

Therefore there exists λ' such that $m_f(A, A') > \lambda' > \lambda - \delta$ and that by Definition 9 A and A' belong the same connected component of $\{p \ge \lambda'\}$. This is a contradiction because $\{p \ge \lambda'\} \subset \{p > \lambda - \delta\}$. □

B.1. Proofs in Section 4.3

Proof of proposition 1. To show proposition 1, we begin by introducing a standard topology lemma.

Lemma 5. *Suppose $p : \mathbb{R}^d \to \mathbb{R}$ are compactly supported. If A and A' are in the same connected components of $\{p \geq \lambda_i\}$ for $i = 1, 2, \dots, \infty$ and that $\lambda_i \leq \lambda_{i+1}$,then A and A' are in the same connected components of $\{p \geq \lambda_0\}$, where $\lambda_0 = \sup_i \lambda_i$.*

Proof of lemma 5. Let $\mathcal{C}_i$ be the connected component of $\{p \geq \lambda_i\}$ that contains A and A'. Thus $\mathcal{C}_i$ are compact and connected. Since $\mathcal{C}_{i+1} \subset \mathcal{C}_i$ for all $i \geq 1$, $\bigcap_{i=1}^{\infty} \mathcal{C}_i$ is connected. Thus $A, A' \subset \bigcap_i \mathcal{C}_i \subset \{p \geq \lambda_0\}$. □

Consider

$$\lambda^* = \sup\{\lambda : A, A' \text{ belongs to the same connected components of } \{p \geq \lambda\}\}$$

Then $\lambda_2 \leq \lambda^* \leq \lambda_1$. By lemma 5 , A and A' are in the same connected components of $\{p \geq \lambda^*\}$. Thus $\lambda_2 \leq \lambda^* < \lambda_1$.
In order to show that λ^* is split level, it suffices to show that A and A' are in the different connected components of $\{p > \lambda^*\}$. Suppose for the sake of contradiction that A and A' are connected in $\{p > \lambda^*\}$. Then A and A' are path connected as $\{p > \lambda^*\}$ is open. Thus there exist $\mathcal{P}$ connects A and A' in $\{p > \lambda^*\}$. Since $\mathcal{P}$ is compact, $p(\mathcal{P}) > \lambda^*$ implies that there exists $a > 0$ such that $\lambda^* + a < \lambda_1$ and $\mathcal{P}, A, A' \subset \{p \geq \lambda^* + a\}$. Thus A and A' belong to the same connected component of $\{p \geq \lambda^* + a\}$. This is a contradiction because by construction of λ^*, A and A' belongs to the different connected components of $\{p \geq \lambda^* + a\}$. □

Proof of corollary 1. Suppose A and A' are δ separated with respect to λ. Then A and A' belongs to distinct connected components of $\{p > \lambda - \delta\}$ where $\lambda = \inf_{x \in A \cup A'} f(x)$. Let $0 < \epsilon \leq \delta$ be given. Then since $\{p \geq \lambda - \delta + \epsilon\} \subset \{p > \lambda - \delta\}$, A and A' belongs to distinct connected components of $\{p \geq \lambda - \delta + \epsilon\}$.
Since $\mathbb{R}^d = \{p \geq 0\}$ is connected, A and A' belongs to the same connected component of $\{p \geq 0\}$. By proposition 1, there exists $0 \leq \lambda^* < \lambda - \delta + \epsilon$ such that A and A' in the same connected component of $\{p \geq \lambda^*\}$ and in different connected components of $\{p > \lambda^*\}$. By taking $\epsilon \to 0$, the claimed result follows.

□

B.2. Proofs in sections 4.4

Proof of Lemma 2. From the proof of Lemma 6, it can be see that

$$\left\{p \geq \lambda + C\frac{\log(n)}{n^{\alpha/(2\alpha+d)}}\right\} \cap \{X_i\}_{i=1}^n \subset \widehat{D}(\lambda) \subset \left\{p \geq \lambda - C\frac{\log(n)}{n^{\alpha/(2\alpha+d)}}\right\}.$$

Thus for any $y \in B(X_j, h)$ for some $X_j \in \hat{D}(\lambda)$,

$$p(y) \geq p(X_j) - Lh^\alpha \geq \lambda - C\frac{\log(n)}{n^{\alpha/(2\alpha+d)}} - Lh^\alpha,$$

where the first inequality follows from $|p(y) - p(X_j)| \leq Lh^{\alpha}$. Therefore the above display implies

$$\bigcup_{X_j \in \hat{D}(\lambda)} B(X_j, h) \subset \left\{ p \geq \lambda - C\frac{\log(n)}{n^{\alpha/(2\alpha+d)}} - Lh^{\alpha} \right\}.$$

For the other inclusion, let $x \in \left\{ p \geq \lambda + C\frac{\log(n)}{n^{\alpha/(2\alpha+d)}} + Lh^{\alpha} \right\}$. Then $\hat{p}_h(x) \geq \lambda + Lh^{\alpha}$. Thus $B(x,h) \cap \{X_i\}_{i=1}^n \neq \emptyset$, or else $\hat{p}_h(x) = 0$. Let $X_j \in B(x,h)$. Therefore

$$p(X_j) \geq p(x) - Lh^{\alpha} \geq \lambda + \frac{\log(n)}{n^{\alpha/(2\alpha+d)}}.$$

Thus $\widehat{p}_h(X_j) \geq \lambda$, which means that $X_j \in \widehat{D}(\lambda)$. So $x \in \bigcup_{X_j \in \hat{D}(\lambda)} B(X_j, h)$ and the first inclusion follows.

□

Proof of Theorem 1. Let $\mathcal{B}$ be the event that

$$\mathcal{B} = \{ \sup_{x \in \mathbb{R}^d} |\widehat{p}_h(x) - p(x)| \leq a_n \}$$

By Proposition 7, we can choose a_n so that $P(\mathcal{B}) \geq 1 - 1/n$ and that $a_n = O\left(\left(\frac{\log(n)}{n} \right)^{\alpha/(d+2\alpha)} \right)$. All the argument will be made on the good event $\mathcal{B}$.

Observe that p has connected support. Therefore $\lambda = 0$ is not a split level. Assume that $\lambda_0 = \min\{\lambda^* : \lambda^* \text{ is a split level of } p\}$. Then $\lambda_0 > 0$. If $h = O(n^{-1/(2\alpha+d)})$, for large n, we have $2a_n + (4h/c_S)^{\alpha} < \min\{\delta_S, \delta_0\}$. Take

$$\delta \geq 2a_n + (4h/c_S)^{\alpha}/c_S.$$

Let A and A' are two sets being δ-separated and let $\lambda = \inf_{x \in A \cup A'} p(x)$. Since A and A' are in distinct connected components of $\{p > \lambda - \delta\}$, by proposition 1 there exists λ^* being a split level of p such that $\lambda^* \leq \lambda - \delta$ and that A and A' belongs to distinct connected components of $\{p > \lambda^*\}$. Thus A and A' belong to distinct connected components of $\{p > \lambda'\}$, where

$$\lambda' = \lambda^* + 2a_n + (4h/c_S)^{\alpha}/c_S.$$

Let $\{\mathcal{C}_k\}_{k=1}^K$ be the collection of connected components of $\{p > \lambda'\}$. Thus we have $A \subset \mathcal{C}_k$ and $A' \subset \mathcal{C}_{k'}$ for some $k \neq k'$. In order to show the smallest cluster containing $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ are disjoint with high probability, it suffices to show the following statement.

- Let A and A' be two connected subsets of $\{p > \lambda'\}$ and belong to two distinct connected components of $\{p > \lambda^*\}$. Then the smallest cluster containing $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ are disjoint with high probability.

Note that this observation reduce the original statement which concerns with generic δ-separated sets to the current statement which only concerns with one level near the split level. Since there are finitely many split levels, a simple union bound will suffice to show the δ consistency of the cluster tree returned by Algorithm 2.

The proof will be completed by the following two claims.

Claim 1. If A is a connected subset of $\{p > \lambda'\}$, then $A \cap \{X_i\}_{i=1}^n$ is in the same connected component of

$$\widehat{L}(\lambda' - a_n) := \bigcup_{\{X_j : \widehat{D}(\lambda' - a_n)\}} B(X_j, 2h). \tag{29}$$

Proof. It suffices to show that

$$\{p > \lambda'\} \subset \widehat{L}(\lambda' - a_n). \tag{30}$$

Since for large n,

$$a_n + (4h/c_S)^\alpha \le \delta_0,$$

By **C2** there exists $\mathcal{N}_h \subset \{p > \lambda'\}$ with $card(\mathcal{N}_h) \le A_c(h)^{-d}$ such that $\mathcal{N}_h$ is a h cover. Since $\{p > \lambda'\}$ satisfies the inner cone condition **C1**,

$$P(B(x,h) \cap \{p > \lambda'\}) \ge \lambda^* c_I V_d h^d \ge \lambda_0 c_I V_d h^d.$$

So there exists c_I' only depending on d and c_I such that

$$P(\{\{X_i\}_{i=1}^n \cap B(x,h) \cap \{p > \lambda'\} = \emptyset\}) \le (1 - \lambda_0 c_I V_d h^d)^n \le \exp(-c_I' \lambda_0 n^{2\alpha/(\alpha+d)}) = o(n^{-2}),$$

where the second inequality follows from $h = O(n^{1/(2\alpha+d)})$ and the equality follows from $\lambda_0 n^{2\alpha/(2\alpha+d)}/\log(n) \to \infty$ and n being large enough. Consider the event

$$\mathcal{A} = \{\{X_i\}_{i=1}^n \cap B(x,h) \cap \{p > \lambda'\} \ne \emptyset \text{ for all } x \in \mathcal{N}_h\}.$$

By the union bound

$$P(\mathcal{A}^c) \le card(\mathcal{N}_h) \exp(-c_I' \lambda_0 n^{2\alpha/(2\alpha+d)}) = A_c h^{-d} \exp(-c_I' \lambda_0 n^{2\alpha/(2\alpha+d)}) = o(1). \tag{31}$$

So for any $y \in \{p > \lambda'\}$, there exists $x \in \mathcal{N}_h$ such that $|y - x| \le h$. Under event $\mathcal{A}$ there exists $X_j \in \{p > \lambda'\}$ such that $|X_j - x| \le h$. Therefore $y \in B(X_j, 2h)$. Since

$$X_j \in \{X_i\}_{i=1}^n \cap \{p > \lambda'\} \subset \widehat{D}(\lambda' - a_n),$$

the claim follows. □

To finish the proof of the theorem, we still need to show at level $\lambda' - a_n$ the data points $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ are contained in distinct clusters. Therefore the following claim finish the proof.

Claim 2. There exists a partition $\{S_i\}_{i=1}^I$ of $\widehat{D}(\lambda' - a_n)$ such that $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$ belong to distinct subsets of the partition and that data points in distinct subsets of the partition are mutually disconnected.

Proof. Let $\{B_i\}_{i=1}^I$ be the collection of connected components of $\{p \geq (4h/c_S)^\alpha + \lambda^*\}$. Since A and A' belong to distinct connected components of $\{p > \lambda^*\}$, and $\lambda^* < (4h/c_S)^\alpha + \lambda^*$, A and A' are contained in distinct elements of $\{B_i\}_{i=1}^I$. From condition **S** ,

$$\min_{i \neq j} d(B_i, B_j) \geq 4h. \tag{32}$$

Note that $\widehat{D}(\lambda' - a_n) \subset \{p \geq (4h/c_S)^\alpha + \lambda^*\}$ as a consequence of event $\mathcal{B}$. Thus $S_i = B_i \cap \widehat{D}(\lambda' - a_n)$ form a partition of $\widehat{D}(\lambda' - a_n)$. Let

$$L_i = \bigcup_{X_j \in S_i} B(X_j, 2h).$$

By (32), $L_i \cap L_j = \emptyset$ if $i \neq j$. This shows that data points in distinct subsets of the partition $\{S_i\}_{i=1}^I$ are mutually disconnected at the graph $\mathbb{C}(h, \lambda' - a_n)$. □

□

Proof of Proposition 2.
Step 1. In this step we show that condition **S(2)** holds. Consider an arbitrary split level λ, and two connected components C_1, C_2. If

$$\inf_{\delta > 0} d(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) > 0$$

then we have $d(C_1 \cap \{p \geq \lambda\}, C_2 \cap \{p \geq \lambda\}) > 0$, and the thesis is trivial. Thus assume that

$$\inf_{\delta > 0} d(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) = 0,$$

i.e.

$$\lim_{\delta \to 0} d(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) = 0.$$

Thus there exists $y_0 \in \{p = \lambda\}$, and points $y_{1,2}^\delta \in C_{1,2} \cap \{p \geq \lambda + \delta\}$ such that

$$y_{1,2}^\delta \overset{\delta \to 0}{\to} y_0, \qquad |y_1^\delta - y_2^\delta| = d(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}).$$

It is straightforward to check that $p(y_1^\delta) = p(y_2^\delta) = \lambda + \delta$.

The thesis is now rewritten as $|y_1^\delta - y_2^\delta| \geq c_S \delta^{1/2}$ for some constant $c_S > 0$ and all sufficiently small δ. Since split levels are also critical, $\nabla p(y_0) = 0$; since p is a Morse function, $\nabla^2 p(y_0)$ is non-degenerate. By Taylor formula we have

$$\delta = p(y_j^\delta) - p(y_0) = (y_j^\delta - y_0)^T \nabla^2 p(y_0)(y_j^\delta - y_0)/2 + O(|y_j^\delta - y_0|^3), \qquad j = 1, 2, \tag{33}$$

and, as $\nabla^2 p(y_0)$ is non-degenerate, it follows $|y_j^\delta - y_0| = O(\delta^{1/2})$, i.e. there exist constants $c_1, c_2, \delta_0 > 0$ such that

$$c_1 \delta^{1/2} \leq |y_j^\delta - y_0| \leq c_2 \delta^{1/2} \qquad \text{for all } \delta \in (0, \delta_0).$$

We can estimate c_2 from below: denoting by

$$a := \max\{|e_1(y_0)|, |e_2(y_0)|\}, \qquad e_1(y_0), e_2(y_0) = \text{eigenvalues of } \nabla^2 p(y_0),$$

(33) gives

$$(y_j^\delta - y_0)^T \nabla^2 p(y_0)(y_j^\delta - y_0) \leq a c_2^2 |y_j^\delta - y_0|^2,$$

hence $c_2 \geq \sqrt{2/a}$. By the Lipschitz regularity of the gradient, i.e. hypothesis

$$|\nabla p(x) - \nabla p(y)| \leq L|x-y|$$

for some $L > 0$, we have

$$|\nabla p(y_1^\delta) - \nabla p(y_0)| = |\nabla p(y_1^\delta)| \leq L|y_1^\delta - y_0| \leq L c_2 \delta^{1/2}. \tag{34}$$

Consider now the segment $[y_1^\delta, y_2^\delta]$ between y_1^δ and y_2^δ: since $y_j^\delta \in C_j \cap \{p \geq \lambda + \delta\}$ $(j = 1, 2)$, and $C_j \cap \{p \geq \lambda + \delta\}$ are disconnected for all $\delta > 0$, there exists some point $z \in [y_1^\delta, y_2^\delta]$ such that $p(z) < \lambda + \delta/2$. By Taylor's formula we then have

$$p(z) = p(y_1^\delta) + \nabla p(y_1^\delta) \cdot (z - y_1^\delta) + (z - y_1^\delta)^T \nabla^2 p(y_1^\delta)(z - y_1^\delta)/2 + O(|z - y_1^\delta|^3)$$

If inequality $|y_1^\delta - y_2^\delta| \leq k\delta^{1/2}$ were to holds for some $k > 0$, then since the domain is compact and $\nabla^2 p \in C^2$, denoting by

$$A := \sup_x \Big(\max\{|e_1(x)|, |e_2(x)|\} \Big), \qquad e_1(x), e_2(x) = \text{eigenvalues of } \nabla^2 p(x),$$

we have

$$\begin{aligned} |p(z) - p(y_1^\delta)| &\leq |\nabla p(y_1^\delta)| \cdot |z - y_1^\delta| + |\nabla^2 p(y_1^\delta)| \cdot |z - y_1^\delta|^2/2 \\ &\leq |\nabla p(y_1^\delta)| \cdot |y_1^\delta - y_2^\delta| + |\nabla^2 p(y_1^\delta)| \cdot |y_1^\delta - y_2^\delta|^2/2 \overset{(34)}{\leq} (Lkc_2 + k^2 A/2)\delta. \end{aligned}$$

Since $p(y_1^\delta) = \lambda + \delta$, and $p(z) < \lambda + \delta/2$, we need $Lkc_2 + k^2A/2 > 1/2$, hence

$$k \geq A^{-1}(\sqrt{L^2 c_2^2 + A} - Lc2),$$

i.e.

$$\begin{aligned} |y_1^\delta - y_2^\delta| &= d(C_1 \cap \{p \geq \lambda + \delta\}, C_2 \cap \{p \geq \lambda + \delta\}) \\ &\geq A^{-1}(\sqrt{L^2 c_2^2 + A} - Lc_2)\delta^{1/2} \geq A^{-1}(\sqrt{2L^2/a + A} - L\sqrt{2/a})\delta^{1/2}. \end{aligned}$$

Step 2. In this step we show that condition **C** holds.

Proof of C1. Since a Morse function has only isolated non degenerate critical points, and an isolated set in a compact domain is also finite, we infer that $\nabla p(x) = 0$ only for finitely many x. In particular, since λ^* are split levels, and $\{p = \lambda^*\}$ contains a critical point, there exist sufficiently small δ_1, $\delta_2 > 0$ such that $\{\lambda^* + \delta_1 \leq p \leq \lambda^* + \delta_2\}$ contains no critical points (since there are only finitely many critical points). Since the level sets are orthogonal to the gradient, we infer that $\{p = \lambda^* + \delta_1\}$ is smooth. In particular, $\{p = \lambda^* + \delta_1\}$ it satisfies the inner cone property with $c_I = 1/2$.

The key difficulty in extending the above argument to $\{p > \lambda^*\}$ (instead of just $\{p \geq \lambda^* + \delta_1\}$ with $\delta_1 > 0$) is that the norm of gradient $|\nabla p|$ can approach zero as $\delta_1 \to 0$, since $\{p = \lambda^*\}$ is a split level, hence it contains critical points.

The Morse function requirement, however, gives the "bare minimum" regularity to ensure C1. We aim to prove, by contradiction, that $\{p \geq \lambda^*\}$ also satisfies C1, i.e. the boundary $\{p = \lambda^*\}$ does not exhibit cusps. If a cusp were to appear, then there exist arc-length parameterized curves $\gamma_j : [0, \epsilon] \longrightarrow \Omega$, $j = 1, 2$, such that $x_0 = \gamma_1(0) = \gamma_2(0)$ and the angle $\angle\gamma_1(s)x_0\gamma_2(s) \to 0$ as $s \to 0$.

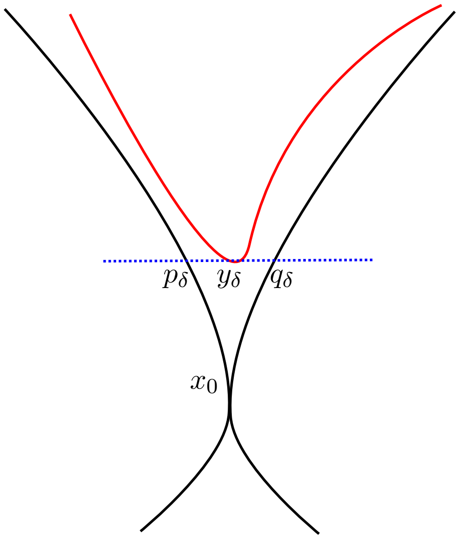

Figure 3: Construction in the proof of Morse function case.

Consider a level set $\{p = \lambda^* + \delta\}$, for small $\delta > 0$. Let $y_\delta \in \{p = \lambda^* + \delta\}$ be the point on $\{p = \lambda^* + \delta\}$ closest to x_0, i.e. $|x_0 - y_\delta| = \min_{y \in \{p=\lambda^*+\delta\}} |x_0 - y|$, and we proved that $|x_0 - y_\delta| = O(\sqrt{\delta})$. Let p_δ, q_δ be the intersection between $\{p = \lambda^*\}$ and the tangent line to $\{p = \lambda^* + \delta\}$ through y_δ. Clearly, as $\{p = \lambda^*\}$ has a cusp at x_0, we get $\lim_{\delta \to 0} \angle p_\delta x_0 q_\delta = 0$. Thus

$$d(y_\delta, \{p = \lambda^*\}) \leq |p_\delta - y_\delta| = o(\sqrt{\delta}).$$

Since $\nabla p(x_0) = 0$, and the gradient ∇p is L-Lipschitz continuous for some constant L, we infer $|\nabla p(y)| \leq A\sqrt{\delta}$ for some $A > 0$ and all y on the segment $[p_\delta, y_\delta]$. Thus it follows

$$\delta = |p(p_\delta) - p(y_\delta)| \leq AL\sqrt{\delta}|p_\delta - y_\delta| = o(\delta).$$

This is a contradiction.

Proof of C2. Let $U = \{p \geq \lambda^* + \delta\}$. Fix an arbitrary r. Clearly $U \subseteq \bigcup_{x \in U} B(x, r/3)$. Since $U = \{p \geq \lambda^* + \delta\}$ is closed, and the domain Ω is compact, we infer $U = \{p \geq \lambda^* + \delta\}$ is also compact. Thus we can extract a covering $U \subseteq \bigcup_{i=1}^{C_r} B(x_i, r/3)$ with finitely many balls. By Vitali covering lemma, we can further extract mutually disjoint balls $B(x_{i_j}, r/3)$

such that

$$U \subseteq \bigcup_{j=1}^{C'_r} B(x_{i_j}, r)$$

Since $B(x_{i_j}, r/3) \subset \Omega_{r/3}$, and $\{B(x_{i_j}, r/3)\}_{j=1}^{C'_r}$ are pairwise disjoint, we have

$$V_d C'_r (r/3)^{-d} \leq \mathcal{L}^d(\Omega_{r/3}).$$

Thus we can choose $\mathcal{N}_r = \{x_{i_j}\}$, $j = 1, \cdots, C'_r$.

□

B.2.1. Proofs in Section 4.5

Proof of lemma 2. Let $\lambda > 0$ be given. Later in the proof, it can be seen that $\lambda = C_0$, being the common upper bound of $f_i \in F$. Define $a > 0$ to be such that

$$56\lambda \cdot 8^{d-1} a^d = 1. \tag{35}$$

Consider

$$f(x) = \begin{cases} \lambda, & x \in [0, 56a] \times [0, 8a]^{d-1} = \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

Let $b = \left(\frac{\log(32)}{n\lambda V_d}\right)^{1/d}$. For $0 < \alpha < 1$, define

$$g(r) = \begin{cases} 0, & 0 \leq r \leq b \\ \mathcal{K}\left(\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} - |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}|\right)^{\alpha}, & |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \leq \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ 0, & \text{otherwise,} \end{cases}$$

and for $\alpha \geq 1$,define

$$g(r) = \begin{cases} 0, & 0 \leq r \leq b \\ 2^{1-\alpha}\delta - \mathcal{K}|r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}|^{\alpha}, & 0 \leq |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \leq \frac{1}{2}\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ \left(\delta^{1/\alpha} - \mathcal{K}^{1/\alpha}|r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}|\right)^{\alpha}, & \frac{1}{2}\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \leq |r - b - \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}| \leq \left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{K} \leq 1$ is chosen so that $g \in \Sigma(L, \alpha)$. By construction $0 \leq g(r) \leq \delta$.

Consider the inequality

$$\begin{aligned} b + 2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} &= \left(\frac{\log(32)}{n\lambda V_d}\right)^{1/d} + 2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \\ &\leq \left(\frac{1}{4^d 8\lambda}\right)^{1/d} + 2\left(\frac{1}{16(7\lambda)^{1/d}}\right) \\ &\leq \frac{1}{4}\left(\frac{1}{7}\right)^{1/d}\lambda^{-1/d} + \frac{1}{8}\left(\frac{1}{7}\right)^{1/d}\lambda^{-1/d} = 3a \end{aligned} \tag{37}$$

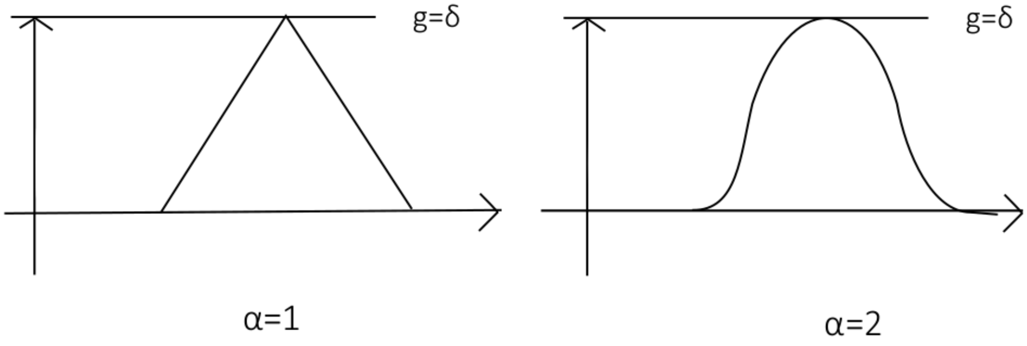

Figure 4: The radial function g for $\alpha = 1, 2$.

where the first inequality follows from $n \geq 4^d \frac{8\log(32)}{V_d}$ and $\delta \leq \left(\frac{\mathcal{K}}{16^{\alpha}(7\lambda)^{\alpha/d}}\right)$. So by construction, 9 disjoint ball $\left\{B\left(x_i, b + 2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}\right)\right\}_{i=0}^{8}$ can be placed in Ω. For $i = 1, \ldots, 8$, let $f_i = f(x) - g(|x - x_i|) + g(|x - x_0|)$. Thus $f_i \in \Sigma(L, \alpha)$ for $, i = 1, \ldots, 8$ and the common upper bound of f_i is $\|f_i\|_\infty = \lambda$. Since $\int f = 1$, by symmetry each f_i also integrates to 1. The fact that $f_i \geq 0$ follows from $0 \leq g(r) \leq \delta \leq p_{\max}$. Since for any $1 \leq i, j \leq 8$, $f_j(x) = \lambda$ for any $x \in B(x_i, b)$.

$$\begin{aligned} P(\text{There exists a point in } B(x_i, b) \text{ for any } i) &\geq 1 - (1 - \lambda V_d b^d)^n \\ &\geq 1 - \exp(-V_d b^d \lambda n) = 1 - 1/32 \end{aligned}$$

where $b = \left(\frac{\log(32)}{n\lambda V_d}\right)^{1/d}$ is used in the last equality. Thus

$$P(\text{There exists a point in every } \ B(x_i, b) \text{ for } \ i = 1 \ldots 8) \geq 3/4.$$

Suppose the family $F = \{f_i\}_{i=1}^{8}$ is given ahead. One wants to show that any algorithm being δ consistent with probability $3/4$ can identify f_i with probability at least $1/2$. To begin consider $B_i = \{f_i \geq \lambda\}$. B_i has exactly two connected components and one is $B(x_i, b)$. Denote the other connected component of B_i by V_i. Thus $B_i = V_i \cup B(x_i, b)$, where $V_i \cap B(x_i, b) = \emptyset$. Define the three events $\mathcal{E}_1$, $\mathcal{E}_1$ and $\mathcal{E}_3$ as following

$$\begin{aligned} &\mathcal{E}_1 = \{\text{There exists a point in every } \ B(x_i, b) \text{ for } \ i = 1 \ldots 8\} \\ &\mathcal{E}_2 = \{\text{The algorithm is } (\delta, \epsilon) \text{ consistent }\} \\ &\mathcal{E}_3 = \{\text{The algorithm can indentify the true density}\} \end{aligned} \tag{38}$$

Then one has $\mathcal{E}_1 \cap \mathcal{E}_2 \subset \mathcal{E}_3$. This is because if an algorithm is δ, consistent and every $B(x_i, b)$ contains at least one point , the algorithm will assign points in $\cup_{j\neq i} B(x_j, b)$ and points $B(x_i, b)$ into different clusters before joining them into the same cluster. In this way, the algorithm can identify the true density. Since $P(\mathcal{E}_1) \geq 3/4$ and $P(\mathcal{E}_2) \geq 3/4$, $P(\mathcal{E}_3) \geq 1/2$

It remains to compute the KL divergent between f_1 and f_2 and apply Fano's lemma. Using spherical coordinate centering at x_1 and x_2, the KL divergent is given by

$$\begin{aligned}
\mathrm{KL}(f_1, f_2) &= dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} (\lambda) \log\left(\frac{\lambda}{\lambda - g(r)}\right) r^{d-1} + (\lambda - g(r)) \log\left(\frac{\lambda - g(r)}{\lambda}\right) r^{d-1} dr \\
&= dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \log\left(\frac{\lambda}{\lambda - g(r)}\right) r^{d-1} dr \\
&= dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \log\left(1 + \frac{g(r)}{\lambda - g(r)}\right) r^{d-1} dr \\
&\leq dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \frac{g(r)}{\lambda - g(r)} r^{d-1} dr \leq dV_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} g(r) \frac{g(r)}{\lambda} r^{d-1} dr \\
&\leq d\lambda^{-1}\delta^2 V_d \int_b^{b+2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}} r^{d-1} dr \\
&\leq \frac{d\delta^2 V_d}{\lambda d} \left(b + 2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}\right)^d
\end{aligned}$$

Thus by Fano's lemma

$$n \geq \frac{(1/2)\log_2(8) - 1}{\mathrm{KL}(f_1, f_2)} = \frac{1}{2\mathrm{KL}(f_1, f_2)} \tag{39}$$

and this implies

$$\left(\frac{\lambda}{2\delta^2 V_d n}\right)^{1/d} \leq 2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} + \left(\frac{\log(32)}{n\lambda V_d}\right)^{1/d}. \tag{40}$$

Since $\delta \leq \lambda/(2^{d/2+1})$, this gives

$$\frac{\lambda}{2^{d+1}\delta^2 V_d n} \geq \frac{\log(32)}{n\lambda V_d}. \tag{41}$$

Combines equation (40) and equation (41) one has

$$2\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} \geq \left(\frac{\lambda}{2\delta^2 V_d n}\right)^{1/d} (1 - \frac{1}{2}) \tag{42}$$

This gives

$$n \geq \frac{\lambda \mathcal{K}^{d/\alpha}}{C(d)\delta^{2+d/\alpha}}, \tag{43}$$

where $C(d) = 2^{2d+1} V_d$.

To justify that the collection of functions $\{f_i\}_{i=1}^8$ constructed in the previous proof satisfies condition **C** and $\mathbf{S}(\alpha)$, observe that λ is the only split level of f_i for all $1 \le i \le 8$. The case of $\alpha > 1$ is only provided as the case of $\alpha < 1$ is simpler. Straight forward computations shows that for any $t \le 2^{-\alpha}$, $\{x : f_i(x) \ge \lambda + t\}$ has two connected components: $B\left(x_i, b_0 + \frac{1}{2}\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha} - \left(\frac{t}{\mathcal{K}}\right)^{1/\alpha}\right)$ and $\left(B\left(x_i, b_0 + \frac{1}{2}\left(\frac{\delta}{\mathcal{K}}\right)^{1/\alpha}\right)\right)^c \cap \Omega$. Therefore condition **C** and $\mathbf{S}(\alpha)$ are trivially satisfied. □

B.3. Proofs in Section 4.6

Poof proposition 3. By (9), with probability at least γ, we have

$$\|\widehat{p}_h - p\|_\infty \le a_n.$$

Step 1. In this step, we show that for any split level λ^* satisfying (18), there exists $\widehat{\lambda^*}$ being Δ-significant and that

$$|\widehat{\lambda^*} - \lambda^*| \le \Delta$$

for large n. Let $\mathcal{C}$ and $\mathcal{C}'$ be two sets split at λ^*. Thus there exists $\mathcal{B}$ being the connected component of $\{p \ge \lambda^*\}$ containing both $\mathcal{C}$ and $\mathcal{C}'$.
By (18), for large n, neither $\{X_i\}_{i=1}^n \cap \mathcal{C} \cap \{p \ge \lambda^* + 2\Delta\}$ nor $\{X_i\}_{i=1}^n \cap \mathcal{C}' \cap \{p \ge \lambda^* + 2\Delta\}$ is empty. Let $X_i \in \mathcal{C} \cap \{p \ge \lambda^* + 2\Delta\}$ and $X_j \in \mathcal{C}' \cap \{p \ge \lambda^* + 2\Delta\}$.

- By the same argument that gives (30)

$$\{p \ge \lambda^*\} \subset \widehat{L}(\lambda^* - a_n) := \bigcup_{\{X_j : \widehat{D}(\lambda^* - a_n)\}} B(X_j, 2h). \tag{44}$$

Since $\mathcal{B} \subset \{p \ge \lambda^*\}$ and that $\mathcal{B}$ is connected, X_i and X_j have the same label in $\mathbb{C}(h, \lambda^* - a_n)$.

- Since $X_i \in \mathcal{C}, X_j \in \mathcal{C}'$ and that $\mathcal{C}$ and $\mathcal{C}'$ are split exactly at λ^*, X_i, X_j are contained in the distinct connected components of $\{p \ge \lambda^* + \Delta\}$. By **Claim 2** in the proof of Theorem 1, X_i and X_j belong to distinct connected components of $\mathbb{C}(h, \lambda^* + \Delta - a_n)$. Let $\widehat{\lambda^*}$ be defined as in (17). By the above two bullet points,

$$\lambda^* - a_n \le \widehat{\lambda^*} \le \lambda^* + \Delta - a_n.$$

The fact that $\widehat{\lambda^*}$ is Δ-significant follows from the observation that

$$X_i, X_j \in \mathbb{C}(h, \lambda^* + 2\Delta - a_n).$$

Step 2. In this step, we show that if $\widehat{\lambda^*}$ is a Δ-significant level of the cluster tree constructed using modified DBSCAN, then there exists λ^* being a split level of p such that

$$|\widehat{\lambda^*} - \lambda^*| \le \Delta.$$

So suppose X_i ,X_j and $\widehat{\lambda^*}$ satisfies (17) and that $X_i, X_j \in \mathbb{C}(h, \widehat{\lambda^*} + \Delta)$. Let

$$\lambda^* := \sup\{\lambda \geq 0 : X_i \text{ and } X_j \text{ are in the same connected component of } \{p \geq \lambda\}\}.$$

- By (44), X_i and X_j have the same label in $\mathbb{C}(h, \lambda^* - a_n)$. Therefore,

$$\lambda^* - a_n \leq \widehat{\lambda^*}.$$

- For the sake of contradiction, suppose that

$$\widehat{\lambda^*} > \lambda^* + \Delta.$$

Then by **Claim 2** in the proof of Theorem 1, X_i and X_j belong to distinct connected components of $\mathbb{C}(h, \lambda^* + \Delta - a_n)$. By definition of $\widehat{\lambda^*}$, this implies

$$\lambda^* + \Delta - a_n \geq \widehat{\lambda^*},$$

which is a contradiction. This finishes the proof.

□

B.4. A Side result: consistency of the KDE tree

As a side result, we also compute the upper bound of cluster tree estimators generated by kernel density estimators. We acknowledge that KDE clustering algorithms have been studied by many authors including Rinaldo and Wasserman (2010), Rigollet and Vert (2009) and Kim et al. (2016). For completeness, we provide δ-consistency results for KDE cluster tree returned by Algorithm 3, but we do not claim the novelty of these results.

Algorithm 3 Clustering based on connected components

```
INPUT: i.i.d sample $\{X_i\}_{i=1}^n$, the kernel $K : \mathbb{R}^d \to \mathbb{R}$, the level $\lambda$ and $h > 0$
  1. Compute $\widehat{L}(\lambda) = \{x : \widehat{p}_h(x) \geq \lambda\}$.
  2. Construct a graph $\mathbb{G}_{h,k}$ with nodes
         $\widehat{D}(\lambda) = \{X_i\}_{i=1}^n \cap \widehat{L}(\lambda)$
     and edges $(X_i, X_j)$ if $X_i$ and $X_j$ belong to the same connected component of $\widehat{L}(\lambda)$.
  3. Compute $\mathbb{C}(h, \lambda)$, the graphical connected components of $\mathbb{G}_{h,\lambda}$.
OUTPUT: $\widehat{T}_n = \{\mathbb{C}(h, \lambda), \lambda \geq 0\}$.
```

We start by showing that for generic $\alpha > 0$, if $p \in \Sigma(L, \alpha)$, level sets of KDE estimator are good approximations of the corresponding population quantities.

Lemma 6. *Assume that $p \in \Sigma(L, \alpha)$, where $\alpha > 0$, and let K be a α-valid kernel. Then, there exist constants C_1 and C_2, depending on $\|p\|_\infty$, K, L and d such that if $h = C_1 \frac{1}{n^{1/(2\alpha+d)}}$, then with probability $1 - 1/n$, uniformly over all $\lambda > 0$,*

$$\left\{x \colon p(x) \geq \lambda + C_2 \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)}\right\} \subset \{x \colon \widehat{p}_h(x) \geq \lambda\} \subset \left\{x \colon p(x) \geq \lambda - C_2 \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)}\right\}. \tag{45}$$

As a direct corollary of Lemma 6, we show that algorithm 3 is consistent with the optimal rate.

Corollary 3. *Let h be chosen as in Lemma 6. Under the assumptions of Lemma 6, the cluster tree returned by Algorithm 3 is δ-consistent with probability at least $1-1/n$, where*

$$\delta \geq 3C_2 \left(\frac{\log n}{n} \right)^{\alpha/(2\alpha+d)}, \tag{46}$$

with $C_2 = C_2(\|p\|_\infty, K, L, d)$ a constant independent of n and δ.

We remark that Algorithm 3 is computationally infeasible even in small dimensions. This is mainly because it requires to compute the level set $\{x : \widehat{p}_h(x) \geq \lambda\}$ exactly to determine the clustering structure of the data points. However Algorithm 3 does not require additional regularity conditions such as $\mathbf{S}(\alpha)$ and $\mathbf{C}$ to attain the minimax optimal rates.

B.4.1. Proofs in Appendix B.4

Proof of lemma 6. For any $x \in \mathbb{R}^d$, with probability at least $1-1/n$

$$\begin{aligned} |\hat{p}_h(x) - p(x)| \leq & |\hat{p}_h(x) - p_h(x)| + |p_h(x) - p(x)| \\ \leq & C_1(K, d, \|p\|_\infty) \sqrt{\frac{\log n}{nh^d}} + C_2(K, \alpha, L) h^\alpha \end{aligned} \tag{47}$$

where the second inequality follows from proposition 7 and standard calculations for the bias. By taking

$$h = h_n = \Theta \left(\frac{\log n}{n} \right)^{1/(2\alpha+d)}$$

in (47),

$$\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p(x)| \leq C(\|p\|_\infty, K, L, \alpha, d) \left(\frac{\log(n)}{n} \right)^{\alpha/(2\alpha+d)}$$

This completes the proof. □

Proof of Corollary 3. Let A and A' be two given connected subsets of $\mathbb{R}^d$. Suppose $\lambda > 0$ satisfies $\lambda + 3\delta = \inf_{x \in A \cup A'} f(x)$ and that A and A' are contained in two distinct connected components of $\{p > \lambda\}$. It suffices to show that the estimate cluster tree at $\{\hat{p}_h \geq \lambda + 2\delta\}$ gives correct labels to $A \cap \{X_i\}_{i=1}^n$ and $A' \cap \{X_i\}_{i=1}^n$, where

$$h = h_n = \Theta \left(\frac{\log n}{n} \right)^{1/(2\alpha+d)}$$

- Since A, A' are connected and
$$A,\ A' \subset \{p \geq 3\delta + \lambda\} \subset \{\hat{p}_h \geq 2\delta + \lambda\},$$
A and A' each belongs to the connected component of $\{\hat{p}_h \geq 2\delta + \lambda\}$. Therefore the cluster tree at $\{\hat{p}_h \geq 2\delta + \lambda\}$ will assign $A \cap \{X_i\}_{i=1}^n$ the same label. This is also true for $A' \cap \{X_i\}_{i=1}^n$.
- It remains to show that A and A' are in the two distinct connected components of $\{\hat{p}_h \geq 2\delta + \lambda\}$. For the sake of contradiction, suppose that A and A' are in the same connected components of $\{\hat{p}_h \geq 2\delta + \lambda\}$. Since
$$\{\hat{p}_h \geq \lambda + 2\delta\} \subset \{p \geq \lambda + \delta\} \subset \{p > \lambda\},$$
A and A' are in the same connected components of $\{p > \lambda\}$. This is a contradiction.

□

Appendix C. Proofs from Section 5

C.1. Proof of Proposition 4

We first prove two simple technical lemmas, that will also be used in the proof of Proposition 6.

Lemma 7. *Suppose that $\epsilon > 2a_n$, where*
$$\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| \leq a_n,$$
and let $\lambda \in (\lambda_* + a_n, \lambda^* - a_n)$. Then,
$$S_{-h} \cap \{X_i\}_{i=1}^n \subset \hat{D}_h(\lambda) \subset S_h,$$
where $S = \{p \geq \lambda^*\}$ and
$$\hat{D}_h(\lambda) = \{x : \hat{p}_h(x) \geq \lambda\} \bigcap \{X_i\}_{i=1}^n.$$

Proof of lemma 7. For the first inclusion, suppose $X_j \in S_{-h} \cap \{X_i\}_{i=1}^n$. Then $B(X_j, h) \subset S$. Since K is supported on $B(0,1)$,
$$p_h(X_j) = \frac{1}{V_d h^d} \int_{B(X_j,h)} p(y)dy \geq \lambda^*. \tag{48}$$
As a result,
$$\hat{p}_h(X_j) \geq p_h(X_j) - a_n \geq \lambda^* - a_n \geq \lambda,$$
which implies that $X_j \in \hat{D}_h(\lambda)$. For the second inclusion, if $X_j \in \hat{D}_h(\lambda)$, then $\hat{p}_h(X_j) \geq \lambda$. So
$$p_h(X_j) \geq \hat{p}_h(X_j) - a_n \geq \lambda - a_n > \lambda_*$$
However, for any point $x \in S_h^c$, since $B(x,h) \subset S^c$, $p_h(x) \leq \lambda_*$ (see (48)). So $X_j \in \hat{D}_h(\lambda)$ implies $X_j \in S_h$. □

Lemma 8. *Under the same assumption as in Lemma 7, suppose further that* $\lambda^* > a_n$. *Let* $\hat{L}(\lambda) = \bigcup_{X_i \in \hat{D}_h(\lambda)} B(X_i, h)$ *and* $\mathcal{C}$ *be any connected components of* S. *Then* $\mathcal{C}_{-2h} \subset \hat{L}(\lambda)$.

Proof of lemma 8. Let $x \in \mathcal{C}_{-2h}$. Then, $B(x,h) \subset S$, which implies, by (48), that $p_h(x) \geq \lambda^*$ and therefore that

$$\hat{p}_h(x) \geq p_h(x) - a_n \geq \lambda^* - a_n > 0.$$

Therefore, $B(x,h) \cap \{X_i\}_{i=1}^n$ is not empty – otherwise $\hat{p}_h(x) = 0$ – so that there exists a sample point, say X_j, in $B(x,h)$. Since $B(x,h) \subset S_{-h}$, we conclude that $X_j \in S_{-h}$. By lemma 7 we then have that $X_j \in \hat{D}_h(\lambda)$. This shows that if $x \in \mathcal{C}_{-2h}$, then there exists some $X_j \in \hat{D}_h(\lambda)$ such that $x \in B(X_j, h)$. This finishes the lemma. □

Proof of Proposition 4. Let

$$a_n = C_1 \sqrt{\frac{\log(n) + \log(1/h)}{nh^d}}$$

be defined as in (7). Then by (6),

$$P\left(\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| \leq a_n\right) \geq 1 - 1/n. \tag{49}$$

Denote $h = C(\frac{\log(n)}{n\epsilon^2})^{1/d}$ where C is chosen such that $3a_n \leq \epsilon$.

Step 1. Suppose $A_{2h} \subset \mathcal{C}_i$. Then $A \subset \mathcal{C}_{i,-2h}$. Then by Lemma 8, one has $A \subset \mathcal{C}_{i,-2h} \subset \hat{L}(\lambda)$. Since data points in connected components of $\hat{L}(\lambda)$ have the same label, and A is contained in only one connected components of $\hat{L}(\lambda)$, points in $A \cap \{X_i\}_{i=1}^n$ have the same labels.

Step 2. Suppose A1 holds. Since $d(\mathcal{C}_i, \mathcal{C}_j) > 4h$, $\{\mathcal{C}_{i,2h}\}_{i=1}^I$ are pairwise disjoint. Since $\hat{D}_h(\lambda_k) \subset \bigcup_{i=1}^I \mathcal{C}_{i,h}$, this means for any i, j there is no edges connect $\hat{D}_h(\lambda_k) \cap \mathcal{C}_{i,h}$ and $\hat{D}_h(\lambda_k) \cap \mathcal{C}_{j,h}$. Since A and A' belong to distinct members of $\{\mathcal{C}_i\}_{i=1}^I$, labels in $A \cap \{X_i\}_{i=1}^n$ and in $A' \cap \{X_i\}_{i=1}^n$ are different.

□

C.2. Proof of Proposition 5

Proof of Proposition 5. Let

$$a_n = C_1 \sqrt{\frac{\log(n) + \log(1/h)}{nh^d}}$$

be defined in (9). Then by (6).

$$P\left(\sup_{x \in \mathbb{R}^d} |\hat{p}_h(x) - p_h(x)| \leq a_n\right) \geq 1 - 1/n. \tag{50}$$

Denote $h = C(\frac{\log(n)}{n\epsilon^2})^{1/d}$ where C is chosen such that $3a_n \leq \epsilon$. Denote $\lambda_k = \frac{k}{nh^d V_d}$. Therefore $\lambda^* - \lambda_* \geq 3a_n$.
Consequently λ is well defined and one has

$$\lambda_* + a_n \leq \lambda < \lambda^* - a_n.$$

By lemma 7, the nodes of $\mathbb{G}_{h,k}$ are contained in S_h. Thus

$$\bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h) \subset S_{2h}. \tag{51}$$

By lemma 8,

$$S_{-2h} \subset \bigcup_{X_j \in \mathbb{G}_{h,k}} B(X_j, h). \tag{52}$$

Since $h_0 \geq h$ using assumption A3 then

$$\mathcal{L}(\hat{S} \triangle S) \leq \mathcal{L}(S_{2h} \backslash S) + \mathcal{L}(S_h \backslash S_{-2h}) \leq C_0 h.$$

□

C.3. Proof of Proposition 6

We begin by constructing of a well-behaved class of sets satisfying the boundary regularity condition **(R)**. These sets will then be used to define high-density clusters in the proof of Proposition 6. Sets satisfying the properties given in the next definition are well known in the literature on support estimation: see, e.g., Korostelev and Tsybakov (1993). For completeness we also show that they satisfy the boundary regularity condition **(R)**.

Definition 13. *Denote by $\mathcal{G}_d(L)$ the class of all domains in $[0,1]^d$ satisfying*

$$\Big\{(x_1, \ldots, x_d) : (x_1, \ldots, x_{d-1}) \in [0,1]^{d-1}, \quad 0 \leq x_d \leq g(x_1, \ldots, x_{d-1})\Big\},$$

where $g : \mathbb{R}^{d-1} \to \mathbb{R}$ satisfies

- $1/2 \leq |g(x)| \leq 3/2$ *for all* $x \in [0,1]^{d-1}$
- $|g(x) - g(x')| \leq L|x - x'|$ *for all* $x, x' \in \mathbb{R}^{d-1}$.

Lemma 9. *There exist a constants h_0 only depending only on L such that for any $\Omega \in \mathcal{G}_d(L)$, one has for any $0 \leq h \leq h_0$,*

$$\mathcal{L}(\Omega_h \backslash \Omega_{-h}) \leq C_0 h,$$

where $\Omega_h = \bigcup_{x \in \Omega} B(x,h)$, $\Omega_{-h} = \{x \in \Omega : B(x,h) \subset \Omega\}$ and C_0 is some constant depending on d.

Proof of lemma 9. Given $\Omega \in \mathcal{G}_d(L)$, let g be the corresponding map as in definition 13. Denote $\underline{x}$ be a generic point in $\mathbb{R}^{d-1}$. Consider the change of coordinate map $\phi : \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$\phi(\underline{x}, x_d) = (\underline{x},\ x_d g(\underline{x})).$$

The inverse map $\phi^{-1} : \mathbb{R}^d \to \mathbb{R}^d$ where $\phi^{-1}(\underline{x}, x_d) = (\underline{x},\ x_d/g(\underline{x}))$ is also well defined as $g > 0$.

Observe that $\phi([0,1]^d) = \Omega$, and there exists a constant $C(d)$ depending only on d such that $[0,1]^d$ satisfies condition A3 with $h_0 = 1/2$ and $C_0 = C(d)$. Thus in order to justify the lemma, it suffices to show that the maps ϕ and ϕ^{-1} only distort the distance and volume by factors depending on L only.

To be more precise, it suffices to show that for some constant L' depending on L and some absolute constant C,

$$|\phi^{-1}(x) - \phi^{-1}(x')| \le L'|x - x'| \text{ and } |\phi(x) - \phi(x')| \le L'|x - x'| \text{ for all } x, x' \in [-2,2]^d$$
$$\mathcal{L}(\phi^{-1}(B)) \le C\mathcal{L}(B) \text{ and } \mathcal{L}(\phi(B)) \le C\mathcal{L}(B) \text{ for any } B \subset [-2,2]^d\ .$$

Since the calculations of ϕ are similar to that of ϕ^{-1}, only the former one is shown in this case.

Step 1. To show that $\phi(x)$ is Lipschitz, it suffices to bound $\|\nabla\phi\|_{op}$.

$$\nabla\phi(\underline{x}, x_d) = (\frac{\partial \phi_i}{\partial x_j}) = \begin{bmatrix} 1 & 0 & \dots & 0 & x_d \frac{\partial g(\underline{x})}{\partial x_1} \\ 0 & 1 & \dots & 0 & x_d \frac{\partial g(\underline{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & x_d \frac{\partial g(\underline{x})}{\partial x_{d-1}} \\ 0 & 0 & \dots & 0 & g(\underline{x}) \end{bmatrix}$$

A straight forward calculations shows that for any $(\underline{x}, x^d) \in [-2,2]^d$,

$$\|\nabla\phi\|_{op} \le 1 + x_0\|\nabla g(\underline{x})\|_2 + g(\underline{x}) \le 5/2 + 2L.$$

Step 2. The change of variables equations gives

$$\mathcal{L}(\phi(B)) = \int_{\phi(B)} d\mathcal{L} = \int_B |\det(\nabla\phi(x))| dx.$$

Since $\det(\nabla\phi(\underline{x}, x_d)) = g(\underline{x})$ which is bounded above by $3/2$, one has $\mathcal{L}(\phi(B)) \le (3/2)\mathcal{L}(B)$.

□

Proof of Proposition 6. Let $0 < \delta \le 1/16$ be depending on ϵ which will be specified later. For some constant depending $C(d)$ only depending on d, it is desired to construct $\{S_i\}_{i=1}^M \in \mathcal{G}_d(C(d))$ such that $\mathcal{L}(S_i \triangle S_j) \ge \delta$ and that M is of order δ^{-d+1}.

Step 1. Consider a hyper rectangle $[0,2\delta]\times[0,\delta]^{d-2}$ in $\mathbb{R}^{d-1}$. One can place $N=\lfloor\delta^{-1}\rfloor^{d-1}/2$ such hyper rectangles into $[0,1]^{d-1}$ without having any two intersect. Denote theses hype rectangles by $\{R_i\}_{i=1}^N$. R_i is composed of two hypercube of dimension $[0,\delta]^{d-1}$, which are denoted as R_i^0 and R_i^1.

Denote $\underline{x}$ to be a generic point in $\mathbb{R}^{d-1}$. One can defined a map $g:[-\delta/2,\delta/2]^{d-1}\to\mathbb{R}$ by

$$g(\underline{x})=\begin{cases}C(d)\left(\delta/2-\|\underline{x}\|_{\mathbb{R}^{d-1}}\right),\ \text{if}\ \|\underline{x}\|_{\mathbb{R}^{d-1}}\le\delta/2\\ 0,\ \text{otherwise}\ .\end{cases}$$

The region

$$\mathcal{C}=\{(\underline{x},x_d):\underline{x}\in[-\delta/2,\delta/2]^{d-1},0\le x_d\le g(\underline{x})\}$$

defines a region of hyper cone in $\mathbb{R}^d$ and $C(d)$ is set so that the cone volume

$$\int_{[-\delta/2,\delta/2]^{d-1}}g(\underline{x})d\underline{x}=\delta^d.$$

Let g_i^0 and g_i^1 be the corresponding map on R_i^0 and R_i^1, as the later ones are copies of $[-\delta/2,\delta/2]^{d-1}$.

Step 2. Let $W=\{w=(w_1,\dots,w_N),w_j\in\{0,1\}\}$. By Varshamov-Gilbert lemma, there exist $w^1,\dots,w^M\in W$ such that (i) $M\ge 2^{N/8}$ (ii) $H(w^i,w^j)\ge N/8$, where H denote the hamming distance.
For $1\le j\le M$, let $G_j:[0,1]^{d-1}\to\mathbb{R}^d$ be defined as

$$G_j(\underline{x})=1/2+\sum_{i=1}^N g_i^{w_i^j}(\underline{x}).$$

Consider

$$S_j=\{(\underline{x},x_d):\underline{x}\in[0,1]^{d-1},0\le x_d\le G_j(\underline{x})\}$$

Thus by construction $G_j\in\mathcal{G}_d(C(d))$ in definition 13. For $l=0,1$ and $1\le i\le N$, define

$$\mathcal{C}_i^l=\{(\underline{x},x_d):\underline{x}\in R_i^l,0\le x_d\le g_i^l(\underline{x})\}.$$

So $\mathcal{C}_j^l$ are non-overlapping cones with volume being δ^d, which are indexical copies of $\mathcal{C}$.

Step 3. Let $\{f_j\}_{j=1}^M$ be such that

$$f_j=\begin{cases}1/4,\ \text{if}\ x\in[0,a]^d\backslash S_j\\ 1/4+\epsilon,\ \text{if}\ x\in S_j\\ 0,\ \text{otherwise}\end{cases}$$

Since $1=\int f_j=a^d/4+(1/4+\epsilon)(3/4)^d\le a^d/4+(1/2)(3/4)^d$, a has to be greater than 1. Thus $S_j\subset[0,a]^d$ and so S_j can be viewed as the support of f_i at the gap.

Step 5. For any i and j Since f_i and f_j are only possibly different on $\{\mathcal{C}_k^0 \cup \mathcal{C}_k^1\}_{k=1}^N$. Also $f_i \neq f_j$ within $\mathcal{C}_k^0 \cup \mathcal{C}_k^1$ if and only if $w_k^i \neq w_k^j$. Thus the $KL(f_i, f_j)$ is determined by

$$KL(f_i, f_j) = \sum_{k=1}^N \int_{\mathcal{C}_k^0 \cup \mathcal{C}_k^1} f_i \log\left(\frac{f_i}{f_j}\right) = \sum_{k: w_k^i \neq w_k^j} \int_{\mathcal{C}_k^0 \cup \mathcal{C}_k^1} f_i \log\left(\frac{f_i}{f_j}\right)$$

Suppose $w_k^i \neq w_k^j$, then

$$\int_{\mathcal{C}_k^0 \cup \mathcal{C}_k^1} f_i \log\left(\frac{f_i}{f_j}\right) = \int_{\mathcal{C}} (1/4+\epsilon) \log\left(\frac{1/4+\epsilon}{1/4}\right) + (1/4)\log\left(\frac{1/4}{1/4+\epsilon}\right) \leq 4\delta^d\epsilon^2.$$

So $KL(f_i, f_j) \leq H(w^i, w^j)\delta^d\epsilon^2 \leq N\delta^d\epsilon^2$.

Step 6. To apply Fano's lower bound lemma (see Tsybakov (2009)) it suffices to have

$$\max_{i\neq j} KL(P_i, P_j) \leq \frac{\log M}{16n}$$

Since $M \geq 2^N/8$ it suffices to have $nN\delta^d\epsilon^2 \leq N\log(2)/128$. Thus it suffices to have $\delta^d = \min\{a\frac{1}{n\epsilon^2}, 1/16\}$ for some absolute constant a.

Step 7. By Fano's lemma, the minimax rate is bounded from below above by

$$\mathcal{L}(S_i \triangle S_j) = H(w^i, w^j) 2\mathcal{L}(\mathcal{C}) \geq (N/8)2\delta^d = c\delta,$$

for some absolute constant c.

□

C.4. Lower bounds of clustering at the Gap

We cite Theorem VI.1 of Chaudhuri et al. (2014) to demonstrate that the scaling in Proposition 4 is minimax optimal.

Proposition 8. *Consider a finite family of density functions $F = \{f_j\}$. Suppose all $f_j \in F$ have gap of size $\epsilon > 0$ at level λ_*. This means that for any j, $\{f_j \geq \lambda_* + \epsilon\} \cup \{f_j \leq \lambda_*\} = \mathbb{R}^d$. For any j, let $\{\mathcal{C}_j^i\}_{i=1}^{I_j}$ be the connected components of $\{f_j \geq \lambda_* + \epsilon\}$ and $d(\mathcal{C}_j^i, \mathcal{C}_j^{i'}) \geq \sigma$ for $i \neq i'$.*
There exists subsets A_j and A_j' for density f_j such that $A_{j,\sigma} \subset \mathcal{C}_j^i$ and $A'_{j,\sigma} \subset \mathcal{C}_j^{i'}$ with the following additional property.
Consider any algorithm that is given $n \geq 100$ i.i.d. samples $\{X_i\}_{i=1}^n$ from some $f_j \in F$ and, with probability at least 3/4, outputs a tree in which the smallest cluster containing $A_j \cap \{X_i\}_{i=1}^n$ is disjoint from the smallest cluster containing $A_j' \cap \{X_i\}_{i=1}^n$. Then there exists a constant $C(d)$ only depending on d such that

$$n \geq \frac{C(d)}{\sigma^d\lambda^*\epsilon^2} \log\frac{1}{\sigma^d\lambda^*}. \tag{53}$$

The proof of the proposition can be found in Theorem VI.1 of Chaudhuri et al. (2014). We omit the details of the proof for brevity.

Appendix D. A Cluster Consistency Resut for General Densities

The type of cluster consistency result we have obtained for densities with gaps can be easily generalized to arbitrary densities. To that end, we will introduce the notion of (ϵ, σ)-separated clusters and of h-thick clusters, where ϵ, σ and h are positive numbers..

Definition 14. *Let ϵ and σ be positive numbers. Two connected subsets A and A' of the support of P are said to be (ϵ, σ)-separated when*

- *they belong to different connected components of $L(\lambda^* - \epsilon)$, where $\lambda^* = \inf_{x \in A \cup A} f(x) > \epsilon$, and*
- $\min_{k \neq l} \operatorname{dist}(\mathcal{C}_k, \mathcal{C}_l) > \sigma$*, where $\mathcal{C}_1, \ldots, \mathcal{C}_m$ are the connected components of $L(\lambda^* - \epsilon)$.*

In addition to being well-separated, we further require the clusters to be *thick*, in a sense made precise below.

Definition 15. *A subset A is h-thick if $A_{-h} \neq \emptyset$.*

Remark 2. Comparison with the separation criterion of Chaudhuri et al. (2014). *The notion of (ϵ, σ)-separated clusters is analogous to the corresponding notion of separated clusters introduced in Chaudhuri et al. (2014), with ϵ and σ quantifying the degree of "vertical" and "horizontal" separation among clusters. There are however, two main differences. First, in our definition the parameter ϵ is on the same scale as the density p and, therefore, represents vertical separation among level sets in an additive and not multiplicative way. Secondly, we use different parameters to measure the degree of horizontal separation among clusters (σ) and the degrees of thickness of the clusters (h); in contrast, Chaudhuri et al. (2014) rely on just one parameter (σ in their notation) to express both separation and thickness.*

With these general notions of thick and well-separated clusters in place, we provide the following uniform consistency results for clustering, which applies to arbitrary densities.

Corollary 4. *Let $\{X_1, \ldots, X_n\}$ be an i.i.d. sample from a probability distribution P with an arbitrary density p and let $\epsilon > 0$ and $\sigma > 0$. Set $a_n = C_1 \sqrt{\frac{\log(n) + \log(1/h)}{nh^d}}$ as in (7) and suppose the input parameters h satisfying*

$$\sigma/4 \geq h \geq C \left(\frac{\log(n)}{n\epsilon^2} \right)^{1/d} \quad \textit{and}, \tag{54}$$

for any $C > 0$ such that $2a_n < \epsilon$. Then with probability at least $1 - 1/n$, uniformly over all clusters A' and A that are h thick and (ϵ, σ) separated,

i. $A_{-h} \cap \{\, X_1, \ldots, X_n\}$ and $A'_{-h} \cap \{\, X_1, \ldots, X_n\}$, if non-empty, belong to distinct connected components of $\mathbb{G}_{k,h}$;

ii. all the sample points in A_{-h}, if any, belong to the same connected component of $\mathbb{G}_{k,h}$

where $k = \lceil nh^d V_d \lambda \rceil$, with $\lambda \in (\lambda^* - \epsilon + a_n, \lambda^* - a_n]$ and $\lambda^* = \inf_{xA \cup A'} f(x)$.

The proof of the previous corollary is almost the identical to the proof of Chaudhuri et al. (2014) and is omitted.

Remark 3. Optimality. *The scaling of the parameters* (ϵ, σ) *is minimax optimal, since the construction used in Chaudhuri et al. (2014) yields* (ϵ, σ)*-separated clusters that are also* h *thick with* $h = \sigma/4$*, so that the resulting lower bound applies to this case as well. Thus, DBSCAN delivers nearly minimax optimal cluster consistency with respect to our definitions of separated and thick clusters.*