CONSISTENT PROCEDURES FOR CLUSTER TREE ESTIMATION AND PRUNING

By Kamalika Chaudhuri[‡], Sanjoy Dasgupta[*,‡], Samory Kpotufe[§,¶] and Ulrike von Luxburg[†,‖]

University of California, San Diego[‡], Toyota Technological Institute-Chicago[§], Max-Planck Institute for Intelligent Systems[¶], and University of Hamburg[‖]

For a density f on $\mathbb{R}^d$, a *high-density cluster* is any connected component of $\{x : f(x) \geq \lambda\}$, for some $\lambda > 0$. The set of all high-density clusters forms a hierarchy called the *cluster tree* of f. We present two procedures for estimating the cluster tree given samples from f. The first is a robust variant of the single linkage algorithm for hierarchical clustering. The second is based on the k-nearest neighbor graph of the samples. We give finite-sample convergence rates for these algorithms which also imply consistency, and we derive lower bounds on the sample complexity of cluster tree estimation. Finally, we study a tree pruning procedure that guarantees, under milder conditions than usual, to remove clusters that are spurious while recovering those that are salient.

1. Introduction. We consider the problem of hierarchical clustering in a "density-based" setting, where a cluster is formalized as a region of high density. Given data drawn i.i.d. from some unknown distribution with density f in $\mathbb{R}^d$, the goal is to estimate the "hierarchical cluster structure" of the density, where a cluster is defined as a connected subset of an f-level set $\{x \in \mathcal{X} \ : \ f(x) \geq \lambda\}$. These subsets form an infinite tree structure as $\lambda \geq 0$ varies, in the sense that each cluster at some level λ is contained in a cluster at a lower level $\lambda' < \lambda$. This infinite tree is called the *cluster tree* of f and is illustrated in Figure 1.

Our formalism of the cluster tree (Section 2.2) and our notion of consistency follow early work on clustering, in particular that of Hartigan (1981). Much subsequent work has been devoted to estimating the connected com-

*Sanjoy Dasgupta is grateful to the National Science Foundation for support under grant IIS-0347646.

†Ulrike von Luxburg acknowledges funding of the German Research Foundation (individual grant LU1718/1-1 and Research Unit 1735 "Structural Inference in Statistics: Adaptation and Efficiency".

AMS 2000 subject classifications: Primary 62G; secondary 68W

Keywords and phrases: Cluster tree estimation, Density level set estimation, Clustering

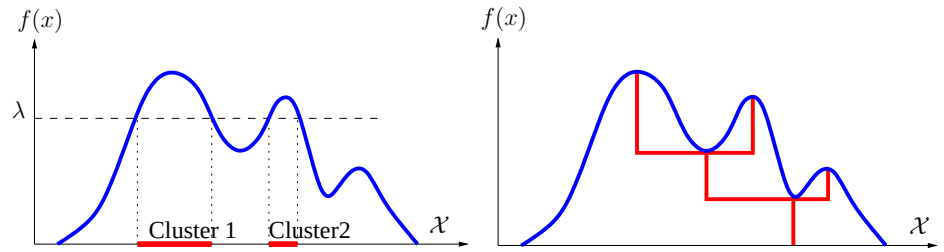


FIG 1. *Left: A probability density f on $\mathbb{R}$, and two clusters at a fixed level λ. Right: The same density, with the branching structure of the corresponding cluster tree.*

ponents of a single level set; see, for example, Polonik (1995), Tsybakov (1997) and, more recently, Maier et al. (2009), Rigollet and Vert (2009), Rinaldo and Wasserman (2010), and Singh et al. (2009). In contrast to these results, the present work is concerned with the simultaneous estimation of all level sets of an unknown density: recovering the cluster tree as a whole.

Are there hierarchical clustering algorithms which converge to the cluster tree? Previous theory work (Hartigan, 1981, Penrose, 1995) has provided partial consistency results for the well-known single-linkage clustering algorithm, while other work (Wishart, 1969) has suggested ways to overcome the deficiencies of this algorithm by making it more robust, but without proofs of convergence. In this paper, we propose a novel way to make single-linkage more robust, while retaining most of its elegance and simplicity (see Figure 3). We establish its finite-sample rate of convergence (Theorem 3.3); the centerpiece of our argument is a result on continuum percolation (Theorem 4.7). This also implies consistency in the sense of Hartigan.

We then give an alternative procedure based on the k-nearest neighbor graph of the sample (see Figure 4). Such graphs are widely used in machine learning, and interestingly there is still much to understand about their expressiveness. We show that by successively removing points from this graph, we can create a hierarchical clustering that also converges to the cluster tree, at roughly the same rate as the linkage-based scheme (Theorem 3.4).

Next, we use tools from information theory to give a lower bound on the problem of cluster tree estimation (Theorem 6.1), which matches our upper bounds in its dependence on most of the parameters of interest.

The convergence results for our two hierarchical clustering procedures nevertheless leave open the possibility that the trees they produce contain spurious branching. This is a well-studied problem in the cluster tree literature, and we address it with a pruning method (Figure 9) that preserves the consistency properties of the tree estimators while providing finite-sample

guarantees on the removal of false clusters (Theorem 7.5). This procedure is based on simple intuition that can carry over to other cluster tree estimators.

2. Definitions and previous work. Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$. We exclusively consider Euclidean distance on $\mathcal{X}$, denoted $\|\cdot\|$. Let $B(x, r)$ be the closed ball of radius r around x.

2.1. *Clustering.* We start by considering the more general context of clustering. While clustering procedures abound in statistics and machine learning, it remains largely unclear whether clusters in finite data—for instance, the clusters returned by a particular procedure—reveal anything meaningful about the underlying distribution from which the data is sampled. Understanding what statistical estimation based on a finite data set reveals about the underlying distribution is a central preoccupation of statistics and machine learning; however this kind of analysis has proved elusive in the case of clustering, except perhaps in the case of density-based clustering.

Consider for instance k-means, possibly the most popular clustering procedure in use today. If this procedure returns k clusters on an n-sample from a distribution f, what do these clusters reveal about f? Pollard (1981) proved a basic consistency result: if the algorithm always finds the global minimum of the k-means cost function (which, incidentally, is NP-hard and thus computationally intractable in general; see Dasgupta and Freund (2009), Theorem 3), then as $n \to \infty$, the clustering is the globally optimal k-means solution for f, suitably defined. Even then, it is unclear whether the best k-means solution to f is an interesting or desirable quantity in settings outside of vector quantization.

Our work, and more generally work on density-based clustering, relies on meaningful formalisms of how a clustering of data generalizes to unambiguous structures of the underlying distribution. The main such formalism is that of the cluster tree.

2.2. *The cluster tree.* We start with notions of connectivity. A *path* P in $S \subset \mathcal{X}$ is a continuous function $P : [0, 1] \to S$. If $x = P(0)$ and $y = P(1)$, we write $x \overset{P}{\rightsquigarrow} y$ and we say that x and y are connected in S. This relation – "connected in S" – is an equivalence relation that partitions S into its *connected components*. We say $S \subset \mathcal{X}$ is *connected* if it has a single connected component.

The cluster tree is a hierarchy each of whose levels is a *partition of a subset* of $\mathcal{X}$, which we will occasionally call a *subpartition* of $\mathcal{X}$. Write $\Pi(\mathcal{X}) = \{\text{subpartitions of } \mathcal{X}\}$.

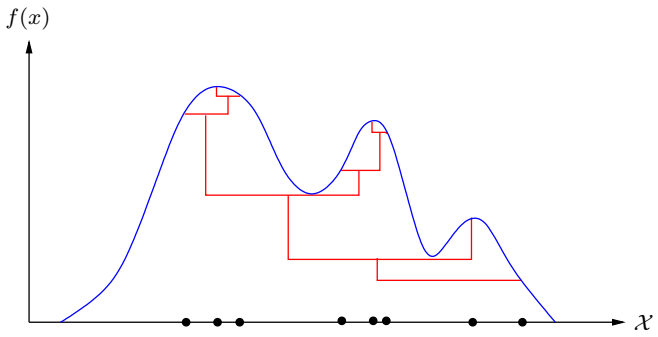


FIG 2. *A probability density f, and the restriction of $\mathbb{C}_f$ to a finite set of eight points.*

DEFINITION 2.1. *For any $f : \mathcal{X} \to \mathbb{R}$, the* cluster tree of f *is a function $\mathbb{C}_f : \mathbb{R} \to \Pi(\mathcal{X})$ given by $\mathbb{C}_f(\lambda) =$ connected components of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$. Any element of $\mathbb{C}_f(\lambda)$, for any λ, is called a* cluster *of f.*

For any λ, $\mathbb{C}_f(\lambda)$ is a set of disjoint clusters of $\mathcal{X}$. They form a hierarchy in the following sense.

LEMMA 2.2. *Pick any $\lambda' \leq \lambda$. Then:*

1. *For any $C \in \mathbb{C}_f(\lambda)$, there exists $C' \in \mathbb{C}_f(\lambda')$ such that $C \subseteq C'$.*
2. *For any $C \in \mathbb{C}_f(\lambda)$ and $C' \in \mathbb{C}_f(\lambda')$, either $C \subseteq C'$ or $C \cap C' = \emptyset$.*

We will sometimes deal with the restriction of the cluster tree to a finite set of points $x_1, \ldots, x_n$. Formally, the restriction of a subpartition $\mathbb{C} \in \Pi(\mathcal{X})$ to these points is defined to be $\mathbb{C}[x_1, \ldots, x_n] = \{C \cap \{x_1, \ldots, x_n\} : C \in \mathbb{C}\}$. Likewise, the restriction of the cluster tree is $\mathbb{C}_f[x_1, \ldots, x_n] : \mathbb{R} \to \Pi(\{x_1, \ldots, x_n\})$, where $\mathbb{C}_f[x_1, \ldots, x_n](\lambda) = \mathbb{C}_f(\lambda)[x_1, \ldots, x_n]$ (Figure 2).

2.3. *Notion of convergence and previous work.* Suppose a sample $X_n \subset \mathcal{X}$ of size n is used to construct a tree $\mathbb{C}_n$ that is an estimate of $\mathbb{C}_f$. Hartigan (1981) provided a sensible notion of consistency for this setting.

DEFINITION 2.3. *For any sets $A, A' \subset \mathcal{X}$, let A_n (resp, A'_n) denote the smallest cluster of $\mathbb{C}_n$ containing $A \cap X_n$ (resp, $A' \cap X_n$). We say $\mathbb{C}_n$ is* consistent *if, whenever A and A' are different connected components of $\{x : f(x) \geq \lambda\}$ (for some $\lambda > 0$), $\mathbb{P}(A_n$ is disjoint from $A'_n) \to 1$ as $n \to \infty$.*

It is well known that if X_n is used to build a uniformly consistent density estimate f_n (that is, $\sup_x |f_n(x) - f(x)| \to 0$), then the cluster tree $\mathbb{C}_{f_n}$

is consistent; see the appendix for details. The problem is that $\mathbb{C}_{f_n}$ is not easy to compute for typical density estimates f_n: imagine, for instance, how one might go about trying to find level sets of a mixture of Gaussians! Wong and Lane (1983) have an efficient procedure that tries to approximate $\mathbb{C}_{f_n}$ when f_n is a k-nearest neighbor density estimate, but they have not shown that it preserves the consistency of $\mathbb{C}_{f_n}$.

On the other hand, there is a simple and elegant algorithm that is a plausible estimator of the cluster tree: *single linkage* (or *Kruskal's algorithm*). Given a data set $x_1, \ldots, x_n \in \mathbb{R}^d$, it operates as follows.

1. For each i, set $r_2(x_i)$ to the distance from x_i to its nearest neighbor.
2. As r grows from 0 to ∞:
   (a) Construct a graph G_r with nodes $\{x_i : r_2(x_i) \leq r\}$. Include edge (x_i, x_j) if $\|x_i - x_j\| \leq r$.
   (b) Let $\mathbb{C}_n(r)$ be the connected components of G_r.

Hartigan (1981) has shown that single linkage is consistent in one dimension (that is, for $d = 1$). But he also demonstrates, by a lovely reduction to continuum percolation, that this consistency fails in higher dimension $d \geq 2$. The problem is the requirement that $A \cap X_n \subset A_n$: by the time the clusters are large enough that one of them contains all of A, there is a reasonable chance that this cluster will be so big as to also contain part of A'.

With this insight, Hartigan defines a weaker notion of *fractional consistency*, under which A_n (resp, A'_n) need not contain *all* of $A \cap X_n$ (resp, $A' \cap X_n$), but merely a sizeable chunk of it – and ought to be very close (at distance $\to 0$ as $n \to \infty$) to the remainder. He then shows that single linkage achieves this weaker consistency for any pair A, A' for which the ratio

$$\frac{\inf\{f(x) : x \in A \cup A'\}}{\sup\{\inf\{f(x) : x \in P\} : \text{paths } P \text{ from } A \text{ to } A'\}}$$

is sufficiently large. More recent work by Penrose (1995) closes the gap and shows fractional consistency whenever this ratio is > 1.

A more robust version of single linkage has been proposed by Wishart (1969): when connecting points at distance r from each other, only consider points that have at least k neighbors within distance r (for some $k > 2$). Thus initially, when r is small, only the regions of highest density are available for linkage, while the rest of the data set is ignored. As r gets larger, more and more of the data points become candidates for linkage. This scheme is intuitively sensible, but Wishart does not provide a proof of convergence. Thus it is unclear how to set k, for instance.

Several papers (Rigollet and Vert, 2009, Maier et al., 2009, Singh et al., 2009, Rinaldo and Wasserman, 2010) have recently considered the problem of recovering the connected components of $\{x : f(x) \geq \lambda\}$ for a user-specified λ: the *flat* version of our problem. Most similar to the work in this paper is the algorithm of Maier et al. (2009), which uses the k-nearest neighbor graph of the data. These level set results invariably require *niceness* conditions on the specific level set being recovered, often stated in terms of the smoothness of the boundary of clusters , and/or regularity conditions on the density f on clusters of the given level set. It is unclear whether these conditions hold for all level sets of a general density, in other words how restrictive these conditions are in the context of recovering the entire cluster tree. In contrast, under mild requirements on the distribution, our conditions on the recovered level sets hold for *any* level set as the sample size n increases. The main distributional requirement for consistency is that of continuity of the density f on a compact support $\mathcal{X}$.

A different approach is taken in a paper of Steinwart (2011), which does not require the user to specify a density level, but rather automatically determines the smallest λ at which $\mathbb{C}_f(\lambda)$ has two components. In Steinwart (2011) the continuity requirements on the density are milder than for other results in the literature, including ours. However it does restrict attention to bimodal densities due to technical hurdles of the flat case: different levels of the cluster tree are collapsed together in the flat case making it difficult to recover a given level from data especially in the case of multimodal densities. Interestingly, the hierarchical setting resolves some of the technical hurdles of the flat case since levels of the cluster tree would generally appear at different levels of a sensible hierarchical estimator. This makes it possible in this paper to give particularly simple estimators, and to analyze them under quite modest assumptions on the data.

A related issue that has received quite a lot of attention is that of *pruning* a cluster tree estimate: removing spurious clusters. A recent result of Rinaldo et al. (2012) gives meaningful statistical guarantees, but is based on the cluster tree of an empirical density estimate, which is algorithmically problematic as discussed earlier. Stuetzle and Nugent (2010) have an appealing top-down scheme for estimating the cluster tree, along with a post-processing step (called *runt pruning*) that helps identify modes of the distribution. The consistency of this method has not yet been established. We provide a consistent pruning procedure for both our procedures.

The present results are based in part on earlier conference versions, namely Chaudhuri and Dasgupta (2010) and Kpotufe and von Luxburg (2011). The result of Chaudhuri and Dasgupta (2010) analyzes the consistency of the

Algorithm 1

1. For each x_i set $r_k(x_i) = \min\{r : B(x_i, r) \text{ contains } k \text{ data points}\}$.
2. As r grows from 0 to ∞:
   (a) Construct a graph G_r with nodes $\{x_i : r_k(x_i) \leq r\}$. Include edge (x_i, x_j) if $\|x_i - x_j\| \leq \alpha r$.
   (b) Let $\mathbb{C}_n(r)$ be the connected components of G_r.

FIG 3. *An algorithm for hierarchical clustering. The input is a sample* $X_n = \{x_1, \ldots, x_n\}$ *from density* f *on* $\mathcal{X}$. *Parameters* k *and* α *need to be set. Single linkage is* $(\alpha = 1, k = 2)$. *Wishart suggested* $\alpha = 1$ *and larger* k.

first cluster tree estimator (see next section) but provides no pruning method for the estimator. The result of Kpotufe and von Luxburg (2011) analyzes the second cluster tree estimator and shows how to prune it. However the pruning method is tuned to this second estimator and works only under strict Hölder continuity requirements on the density. The present work first provides a unified analysis of both estimators using techniques developed in Chaudhuri and Dasgupta (2010). Second, building on insight from Kpotufe and von Luxburg (2011), we derive a new pruning method which proveably works for either estimator without Hölder conditions on the distribution. In particular, the pruned version of either cluster tree estimate remains consistent under mild uniform continuity assumptions. The main finite-sample pruning result of Theorem 7.5 requires even milder conditions on the density than required for consistency.

3. Algorithms and results. The first algorithm we consider in this paper is a generalization of Wishart's scheme and of single linkage, shown in Figure 3. It has two free parameters: k and α. For practical reasons, it is of interest to keep these as small as possible. We provide finite-sample convergence rates for all $1 \leq \alpha \leq 2$ and we can achieve $k \sim d \log n$ if $\alpha \geq \sqrt{2}$. Our rates for $\alpha = 1$ force k to be much larger, exponential in d. It is an open problem to determine whether the setting $(\alpha = 1, k \sim d \log n)$ yields consistency.

Conceptually, the algorithm creates a series of graphs $G_r = (V_r, E_r)$ satisfying a nesting property: $r \leq r' \quad \Rightarrow \quad V_r \subset V_{r'}$ and $E_r \subset E_{r'}$. A point is admitted into G_r only if it has k neighbors within distance r; when r is small, this picks out the regions of highest density, roughly. The edges of G_r are between all pairs of points within distance αr of each other.

In practice, the only values of r that matter are those corresponding to interpoint distances within the sample, and thus the algorithm is efficient.

Algorithm 2

1. For each x_i set $r_k(x_i) = \min\{r : B(x_i, r) \text{ contains } k \text{ data points}\}$.
2. As r grows from 0 to ∞:
   (a) Construct a graph G_r^{NN} with nodes $\{x_i : r_k(x_i) \leq r\}$.
   Include edge (x_i, x_j) if:

$$\begin{array}{ll} \|x_i - x_j\| \leq \alpha \max(r_k(x_i), r_k(x_j)) & k\text{-NN graph} \\ \|x_i - x_j\| \leq \alpha \min(r_k(x_i), r_k(x_j)) & \text{mutual } k\text{-NN graph} \end{array} .$$

   (b) Let $\mathbb{C}_n(r)$ be the connected components of G_r^{NN}.

FIG 4. *A cluster tree estimator based on the k-nearest neighbor graph.*

A further simplification is that the graphs G_r don't need to be explicitly created. Instead, the clusters can be generated directly using Kruskal's algorithm, as is done for single linkage.

The second algorithm we study (Figure 4) is based on the k-nearest neighbor graph of the samples. There are two natural ways to define this graph, and we will analyze the sparser of the two, the mutual k-NN graph, which we shall denote G^{NN}. Our results hold equally for the other variant.

One way to think about the second hierarchical clustering algorithm is that it creates the k-nearest neighbor graph on all the data samples, and then generates a hierarchy by removing points from the graph in decreasing order of their k-NN radius $r_k(x_i)$. The resulting graphs G_r^{NN} have the same nodes as the corresponding G_r but have potentially fewer edges: $E_r^{\text{NN}} \subset E_r$. This makes them more challenging to analyze.

Much of the literature on density-based clustering refers to clusters not by the radius r at which they appear, but by the "corresponding empirical density", which in our case would be $\lambda = k/(nv_d r^d)$, where v_d is the volume of the unit ball in $\mathbb{R}^d$. The reader who is more comfortable with the latter notation should mentally substitute $G[\lambda]$ whenever we refer to G_r. We like using r because it is directly observed rather than inferred. Consider, for instance, a situation in which the underlying density f is supported on a low-dimensional submanifold of $\mathbb{R}^d$. The two cluster tree algorithms continue to be perfectly sensible, as does r; but the inferred λ is misleading.

3.1. *A notion of cluster salience.* Suppose density f is supported on some subset $\mathcal{X}$ of $\mathbb{R}^d$. We will find that when Algorithms 1 and 2 are run on data drawn from f, their estimates are consistent in the sense of Definition 2.3. But an even more interesting question is, what clusters will be identified from a *finite* sample? To answer this, we need a notion of salience.

The first consideration is that a cluster is hard to identify if it contains a thin "bridge" that would make it look disconnected in a small sample. To control this, we consider a "buffer zone" of width σ around the clusters.

DEFINITION 3.1. *For $Z \subset \mathbb{R}^d$ and $\sigma > 0$, write $Z_\sigma = Z + B(0,\sigma) = \{y \in \mathbb{R}^d : \inf_{z\in Z} \|y - z\| \le \sigma\}$.*

Z_σ is a full-dimensional set, even if Z itself is not.

Second, the ease of distinguishing two clusters A and A' depends inevitably upon the separation between them. To keep things simple, we'll use the same σ as a separation parameter.

DEFINITION 3.2. *Let f be a density supported on $\mathcal{X} \subset \mathbb{R}^d$. We say that $A, A' \subset \mathcal{X}$ are (σ,ϵ)-**separated** if there exists $S \subset \mathcal{X}$ (*separator set*) such that (*i*) any path in $\mathcal{X}$ from A to A' intersects S, and (*ii*) $\sup_{x\in S_\sigma} f(x) < (1-\epsilon)\inf_{x\in A_\sigma \cup A'_\sigma} f(x)$.*

Under this definition, A_σ and A'_σ must lie within $\mathcal{X}$, otherwise the right-hand side of the inequality is zero. S_σ need not be contained in $\mathcal{X}$.

3.2. *Consistency and rate of convergence.* We start with a result for Algorithm 1, under the settings $\alpha \ge \sqrt{2}$ and $k \sim d\log n$. The analysis section also has results for $1 \le \alpha \le 2$ and $k \sim (2/\alpha)^d d\log n$. The result states general saliency conditions under which a given level λ of the cluster tree is recovered at level $r(\lambda)$ of the estimator. The mapping r is of the form $\left(\frac{k}{nv_d\lambda}\right)^{1/d}$ (see Definition 4.4), where v_d is the volume of the unit ball in $\mathbb{R}^d$.

THEOREM 3.3. *There is an absolute constant C such that the following holds. Pick any $0 < \delta, \epsilon < 1$, and run Algorithm 1 on a sample X_n of size n drawn from f, with settings*

$$\sqrt{2} \le \alpha \le 2 \quad \text{and} \quad k \ge C \cdot \frac{d\log n}{\epsilon^2} \cdot \log^2 \frac{1}{\delta}.$$

Then there is a mapping $r : [0,\infty) \to [0,\infty)$ such that the following holds with probability at least $1-\delta$. Consider any pair of connected subsets $A, A' \subset \mathcal{X}$ such that A, A' are (σ,ϵ)-separated for ϵ and some $\sigma > 0$. Let $\lambda = \inf_{x\in A_\sigma \cup A'_\sigma} f(x)$. If $n \ge \frac{k}{v_d(\sigma/2)^d\lambda}\left(1+\frac{\epsilon}{2}\right)$, then:

1. *Separation. $A \cap X_n$ is disconnected from $A' \cap X_n$ in $G_{r(\lambda)}$.*
2. *Connectedness. $A \cap X_n$ and $A' \cap X_n$ are each connected in $G_{r(\lambda)}$.*

The two parts of this theorem – separation and connectedness – are proved in Sections 4.1 and 4.2, respectively.

A similar result holds for Algorithm 2 under stronger requirements on k.

THEOREM 3.4. *Theorem 3.3 applies also to Algorithm 2, provided the following additional condition on k is met: $k \geq \frac{\Lambda}{\lambda} \cdot Cd \log n \cdot \log \frac{1}{\delta}$, where $\Lambda = \sup_{x \in \mathcal{X}} f(x)$.*

In the analysis section, we give a lower bound (Lemma 5.3) that shows why this dependence on Λ/λ is needed.

Finally, we point out that these finite-sample results imply consistency (Definition 2.3): as $n \to \infty$, take $k_n = (d \log n)/\epsilon_n^2$ with any schedule of $\{\epsilon_n\}$ such that $\epsilon_n \to 0$ and $k_n/n \to 0$. Under mild uniform continuity conditions, any two connected components A, A' of $\{f \geq \lambda\}$ are (σ, ϵ)-separated for some $\sigma, \epsilon > 0$ (see appendix); thus they are identified given large enough n.

4. Analysis of Algorithm 1.

4.1. *Separation.* Both cluster tree algorithms depend heavily on the radii $r_k(x)$: the distance within which x's nearest k neighbors lie (including x itself). The empirical probability mass of $B(x, r_k(x))$ is k/n. To show that $r_k(x)$ is meaningful, we need to establish that the mass of this ball under density f is also roughly k/n. The uniform convergence of these empirical counts follows from the fact that balls in $\mathbb{R}^d$ have finite VC dimension, $d+1$.

We also invoke uniform convergence over *half-balls*: each of these is the intersection of a ball with a halfspace through its center. Using uniform Bernstein-type bounds, we derive basic inequalities which we use repeatedly.

LEMMA 4.1. *Assume $k \geq d \log n$, and fix some $\delta > 0$. Then there exists a constant C_δ such that with probability $> 1 - \delta$, we have that, first, every ball $B \subset \mathbb{R}^d$ satisfies the following conditions:*

$$\begin{aligned} f(B) \geq \frac{C_\delta d \log n}{n} &\implies f_n(B) > 0 \\ f(B) \geq \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd \log n} &\implies f_n(B) \geq \frac{k}{n} \\ f(B) \leq \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd \log n} &\implies f_n(B) < \frac{k}{n} \end{aligned}$$

Here $f_n(B) = |X_n \cap B|/n$ is the empirical mass of B, while $f(B)$ is its probability under f. Second, for every half-ball $H \subset \mathbb{R}^d$:

$$f(H) \geq \frac{C_\delta d \log n}{n} \implies f_n(H) > 0.$$

We denote this uniform convergence over balls and half-balls as event E_o.

PROOF. See appendix. $C_\delta = 2C_o \log(2/\delta)$, where C_o is the absolute constant from Lemma C.2. □

We will typically preface other results by a statement like "Assume E_o." It is to be understood that E_o occurs with probability at least $1-\delta$ over the random sample X_n, where δ is henceforth fixed. The constant C_δ will keep reappearing through the paper.

For any cluster $A \subset \mathcal{X}$, there is a certain scale r at which every data point in A appears in G_r. What is this r?

LEMMA 4.2. *Assume E_o. Pick any set $A \subset \mathcal{X}$, and let $\lambda = \inf_{x \in A_\sigma} f(x)$. If $r < \sigma$ and $v_d r^d \lambda \geq \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd \log n}$, then G_r contains every point in $A_{\sigma-r} \cap X_n$.*

PROOF. Any point $x \in A_{\sigma-r}$ has $f(B(x,r)) \geq v_d r^d \lambda$; and thus, by Lemma 4.1, has at least k neighbors within radius r. □

In order to show that two separate clusters A and A' get distinguished in the cluster tree, we need to exhibit a scale r at which every point in A and A' is active, but there is no path from A to A'.

LEMMA 4.3. *Assume E_o. Suppose sets $A, A' \subset \mathcal{X}$ are (σ, ϵ)-separated by set S, and let $\lambda = \inf_{x \in A_\sigma \cup A'_\sigma} f(x)$. Pick $0 < r < \sigma$ such that*

$$\frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd \log n} \leq v_d r^d \lambda < \left(\frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd \log n}\right) \cdot \frac{1}{1-\epsilon}.$$

Then:

(a) G_r contains all points in $(A_{\sigma-r} \cup A'_{\sigma-r}) \cap X_n$.
(b) G_r contains no points in $S_{\sigma-r} \cap X_n$.
(c) If $r < 2\sigma/(\alpha+2)$, then $A \cap X_n$ is disconnected from $A' \cap X_n$ in G_r.

PROOF. Part (a) is directly from Lemma 4.2. For (b), any point $x \in S_{\sigma-r}$ has $f(B(x,r)) < v_d r^d \lambda(1-\epsilon)$; and thus, by Lemma 4.1, has strictly fewer than k neighbors within distance r.

For (c), since points in $S_{\sigma-r}$ are absent from G_r, any path from A to A' in that graph must have an edge across $S_{\sigma-r}$. But any such edge has length at least $2(\sigma - r) > \alpha r$ and is thus not in G_r. □

DEFINITION 4.4. *Define $r(\lambda)$ to be the value of r for which $v_d r^d \lambda = \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd \log n}$.*

COROLLARY 4.5. *The conditions of Lemma 4.3 are satisfied by $r = r(\lambda)$ if $r(\lambda) < 2\sigma/(\alpha+2)$ and $k \geq 4C_\delta^2(d/\epsilon^2)\log n$.*

4.2. *Connectedness.* We need to show that points in A (and similarly A') are connected in $G_{r(\lambda)}$. First we state a simple bound (proved in the appendix) that works if $\alpha = 2$ and $k \sim d \log n$; later we consider smaller α.

LEMMA 4.6. *Assume E_o. Let A be a connected set in $\mathcal{X}$ with $\lambda = \inf_{x \in A_\sigma} f(x)$. Suppose $1 \leq \alpha \leq 2$. Then $A \cap X_n$ is connected in G_r whenever $r \leq 2\sigma/(2+\alpha)$ and*

$$v_d r^d \lambda \quad \geq \quad \max\left\{ \left(\frac{2}{\alpha}\right)^d \frac{C_\delta d \log n}{n}, \quad \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd \log n} \right\}.$$

Comparing this to the definition of $r(\lambda)$, we see that choosing $\alpha = 1$ would entail $k \geq 2^d$, which is undesirable. We can get a more reasonable setting of $k \sim d \log n$ by choosing $\alpha = 2$, but we'd like α to be as small as possible. A more refined argument shows that $\alpha \approx \sqrt{2}$ is enough.

THEOREM 4.7. *Assume E_o. Let A be a connected set in $\mathcal{X}$ with $\lambda = \inf_{x \in A_\sigma} f(x)$. Suppose $\alpha \geq \sqrt{2}$. Then $A \cap X_n$ is connected in G_r whenever $r \leq \sigma/2$ and*

$$v_d r^d \lambda \quad \geq \quad \max\left\{ \frac{4C_\delta d \log n}{n}, \quad \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd \log n} \right\}.$$

PROOF. Recall that a *half-ball* is the intersection of an open ball and a halfspace through the center of the ball. Formally, it is defined by a center μ, a radius r, and a unit direction u:

$$\{z \in \mathbb{R}^d : \|z - \mu\| < r, (z - \mu) \cdot u > 0\}.$$

We will describe any such set as "the half of $B(\mu, r)$ in direction u". If the half-ball lies entirely in A_σ, its probability mass is at least $(1/2)v_d r^d \lambda$. By uniform convergence bounds (Lemma 4.1), if $v_d r^d \lambda \geq (4C_\delta d \log n)/n$, then every such half-ball within A_σ contains at least one data point.

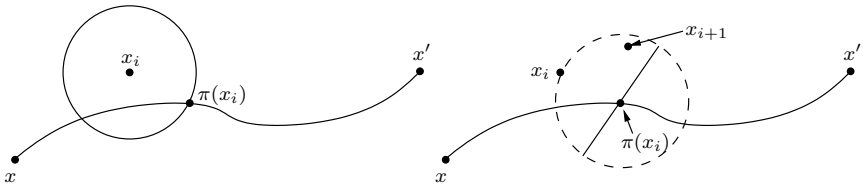


FIG 5. *Left: P is a path from x to x', and $\pi(x_i)$ is the point furthest along the path that is within distance r of x_i. Right: The next point, $x_{i+1} \in X_n$, is chosen from the half-ball of $B(\pi(x_i), r)$ in the direction of $x_i - \pi(x_i)$.*

Pick any $x, x' \in A \cap X_n$; there is a path P in A with $x \overset{P}{\rightsquigarrow} x'$. We'll identify a sequence of data points $x_0 = x, x_1, x_2, \ldots$, ending in x', such that for every i, point x_i is active in G_r and $\|x_i - x_{i+1}\| \leq \alpha r$. This will confirm that x is connected to x' in G_r.

To begin with, recall that P is a continuous function from $[0,1]$ into A. For any point $y \in \mathcal{X}$, define $N(y)$ to be the portion of $[0,1]$ whose image under P lies in $B(y,r)$: that is, $N(y) = \{0 \leq z \leq 1 : P(z) \in B(y,r)\}$. If y is within distance r of P, then $N(y)$ is nonempty. Define $\pi(y) = P(\sup N(y))$, the furthest point along the path within distance r of y (Figure 5, left).

The sequence $\{x_i\}$ is defined iteratively; $x_0 = x$, and for $i = 0, 1, 2, \ldots$:

- If $\|x_i - x'\| \leq \alpha r$, set $x_{i+1} = x'$ and stop.
- By construction, x_i is within distance r of path P and hence $N(x_i) \neq \emptyset$.
- Let B be the open ball of radius r around $\pi(x_i)$. The half of B in direction $x_i - \pi(x_i)$ contains a data point; this is x_{i+1} (Figure 5, right).

The process eventually stops since each $\pi(x_{i+1})$ is further along path P than $\pi(x_i)$; formally, $\sup N(x_{i+1}) > \sup N(x_i)$. This is because $\|x_{i+1} - \pi(x_i)\| < r$, so by continuity of the function P, there are points further along P (beyond $\pi(x_i)$) whose distance to x_{i+1} is still $< r$. Thus x_{i+1} is distinct from $x_0, x_1, \ldots, x_i$. Since there are finitely many data points, the process must terminate, so the sequence $\{x_i\}$ constitutes a path from x to x'.

Each x_i lies in $A_r \subseteq A_{\sigma - r}$ and is thus active in G_r under event E_o (Lemma 4.2). Finally, the distance between successive points is $\|x_i - x_{i+1}\|^2$

$$\begin{aligned} &= \|x_i - \pi(x_i) + \pi(x_i) - x_{i+1}\|^2 \\ &= \|x_i - \pi(x_i)\|^2 + \|\pi(x_i) - x_{i+1}\|^2 - 2(x_i - \pi(x_i)) \cdot (x_{i+1} - \pi(x_i)) \\ &\leq 2r^2 \quad \leq \quad \alpha^2 r^2, \end{aligned}$$

where the second-last inequality is from the definition of half-ball. □

To complete the proof of Theorem 3.3, take $k \geq 4C_\delta^2(d/\epsilon^2)\log n$. The relationship that defines $r = r(\lambda)$ (Definition 4.4) then implies

$$\frac{k}{n} \;\leq\; v_d r^d \lambda \;\leq\; \frac{k}{n}\left(1+\frac{\epsilon}{2}\right).$$

This shows that clusters at density level λ emerge when the growing radius r of the cluster tree algorithm reaches roughly $(k/(\lambda v_d n))^{1/d}$. In order for (σ,ϵ)-separated clusters to be distinguished, the one additional requirement of Lemma 4.3 and Theorem 4.7 is that $r = r(\lambda)$ be at most $\sigma/2$; this is what yields the final lower bound on n.

5. Analysis of Algorithm 2. The second cluster tree estimator (Figure 4), based on the k-nearest neighbor graph of the data points, satisfies the same guarantees as the first, under a more generous setting of k.

Let G_r^{NN} be the k-NN graph at radius r. We have already observed that G_r^{NN} has the same vertices as G_r, and a subset of its edges. Therefore, if clusters are separated in G_r, they are certainly separated in G_r^{NN}: the separation properties of Lemma 4.3 carry over immediately to the new estimator. What remains is to establish a connectedness property, an analogue of Theorem 4.7, for these potentially much sparser graphs.

5.1. *Connectivity properties.* As before, let f be a density on $\mathcal{X} \subset \mathbb{R}^d$. Let $\Lambda = \sup_{x\in\mathcal{X}} f(x)$; then the smallest radius we expect to be dealing with is roughly $(k/(nv_d\Lambda))^{1/d}$. To be safe, let's pick a value slightly smaller than this, and define $r_o = (k/(2nv_d\Lambda))^{1/d}$.

We'll first confirm that r_o is, indeed, a lower bound on the radii $r_k(\cdot)$.

LEMMA 5.1. *Assume E_o. If $k \geq 4C_\delta^2 d\log n$, then $r_k(x) > r_o$ for all x.*

PROOF. Pick any x and consider the ball $B(x, r_o)$. By definition of r_o,

$$f(B(x,r_o)) \;\leq\; v_d r_o^d \Lambda \;=\; \frac{k}{2n} \;\leq\; \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}$$

where the last inequality is from the condition on k. Under E_o (Lemma 4.1), we then get $f_n(B(x,r_o)) < k/n$; therefore $r_k(x) > r_o$. □

Now we present an analogue of Theorem 4.7.

THEOREM 5.2. *Assume E_o. Let A be a connected set in $\mathcal{X}$, with $\lambda = \inf_{x\in A_\sigma} f(x)$. Suppose $\alpha \geq \sqrt{2}$. Then $A \cap X_n$ is connected in G_r^{NN} whenever $r + r_o \leq \sigma$ and*

$$v_d r^d \lambda \quad\geq\quad \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}$$

and

$$k \;\geq\; \max\left\{\frac{\Lambda}{\lambda}\cdot 8C_\delta d\log n, 4C_\delta^2 d\log n\right\}.$$

PROOF. We'll consider events at two different scales: a small radius r_o, and the potentially larger radius r from the theorem statement.

Let's start with the small scale. The lower bound on k yields

$$v_d r_o^d \lambda \;=\; \frac{k}{2n}\cdot\frac{\lambda}{\Lambda} \;\geq\; \frac{4C_\delta d\log n}{n}.$$

As in Theorem 4.7, this implies that every half-ball of radius r_o within A_σ contains at least one data point.

Let x and x' be any two points in $A\cap X_n$. As in Theorem 4.7, we can find a finite sequence of data points $x = x_0, x_1, \ldots, x_p = x'$ such that for each i, two key conditions hold: (i) $\|x_i - x_{i+1}\| \leq \alpha r_o$ and (ii) x_i lies within distance r_o of A.

Now let's move to a different scale $r \leq \sigma - r_o$. Since each x_i lies in $A_{r_o} \subseteq A_{\sigma-r}$, we know from Lemma 4.2 that all x_i are active in G_r^{NN} given the lower bound on $v_d r^d\lambda$. The edges (x_i, x_{i+1}) are also present, because

$$\|x_i - x_{i+1}\| \leq \alpha r_o \leq \alpha\min(r_k(x_i), r_k(x_{i+1}))$$

using Lemma 5.1 and the bound on k. Hence x is connected to x' in G_r^{NN}. □

It is straightforward to check that $r(\lambda)$ is always $\geq r_o$, and Theorem 3.4 follows immediately.

5.2. *A lower bound on neighborhood cardinality.* The result for k-nearest neighbor graphs requires a larger setting of k than our earlier result; in particular, k needs to exceed the ratio Λ/λ. We now show that this isn't just a looseness in our bound, but in fact a necessary condition for these types of graphs.

Recall that the mutual k-NN graph contains all the data points, and puts an edge between points x and x' if $\|x - x'\| \leq \alpha\min(r_k(x), r_k(x'))$ (the α is our adaptation). We will assume $1 \leq \alpha \leq 2$, as is the case in all our upper bounds.

LEMMA 5.3. *Pick any $\lambda > 0$, any $\Lambda > 32\lambda$, and any $k \leq \Lambda/(64\lambda)$. Then there is a density f on $\mathcal{X}\subset\mathbb{R}$ with $\lambda \leq f(x) \leq \Lambda$ for all $x\in\mathcal{X}$, and with the following property: for large enough n, when n samples are drawn i.i.d. from f, the resulting mutual k-NN graph (with $1\leq\alpha\leq 2$) is disconnected with probability at least $1/2$.*

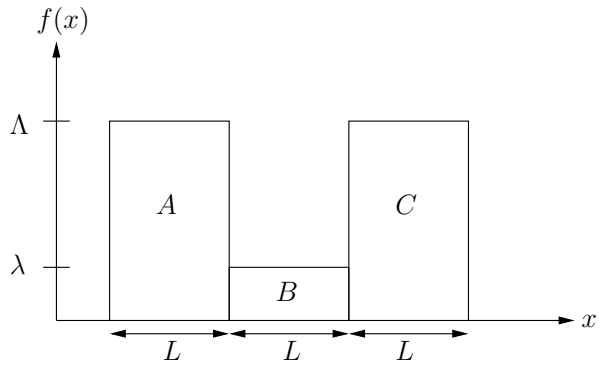


FIG 6. *A density that illustrates why k-NN graphs require $k \geq \Lambda/\lambda$ for connectivity.*

PROOF. Consider the density shown in Figure 6, consisting of two dense regions, A and C, bridged by a less dense region B. Each region is of width $L = 1/(\lambda + 2\Lambda)$. We'll show that the mutual k-NN graph of a sample from this distribution is likely to be disconnected. Specifically, with probability at least $1/2$, there will be no edges between A and $B \cup C$.

To this end, fix any $n \geq \Lambda/\lambda$, and define $\Delta = 1/(4n\lambda) < L$. Consider the leftmost portion of B of length Δ. The probability that a random draw from f falls in this region is $\Delta\lambda = 1/(4n)$. Therefore, the probability that no point falls in this region is $(1 - 1/(4n))^n \geq 3/4$. Call this event E_1.

Next, divide A into intervals of length $\Delta, 2\Delta, 4\Delta$, and so on, starting from the right. We'll show that with probability at least $3/4$, the right half of each such interval contains at least $k + 1$ points; call this event E_2. To see why, let's focus on one particular interval, say that of length $2^i\Delta$. The probability that a random point falls in the right half of this interval is $2^{i-1}\Delta\Lambda \geq 2^{i+3}k/n$. Therefore, the number of points in this region is $\geq 2^{i+3}k$ in expectation, and by a Chernoff bound, is $\geq k+1$ except with probability $< \exp(-2^{i+1}k)$. Taking a union bound over all the intervals yields an overall failure probability of at most $1/4$.

With probability at least $1/2$, events E_1 and E_2 both occur. Whereupon, for any point in A, its nearest neighbor in $B \cup C$ is at least twice as far as its k nearest neighbors in A. Thus the mutual k-NN graph has no edges between A and $B \cup C$. □

This constraint on k is unpleasant, and it would be interesting to either find mild smoothness assumptions on f, or better, modified notions of k-NN graph, that render it unnecessary.

6. Lower bound. We have shown that the two cluster tree algorithms distinguish pairs of clusters that are (σ, ϵ)-separated. The number of samples required to capture clusters at density $\geq \lambda$ is, by Theorem 3.3,

$$O\left(\frac{d}{v_d(\sigma/2)^d\lambda\epsilon^2}\log\frac{d}{v_d(\sigma/2)^d\lambda\epsilon^2}\right),$$

We'll now show that this dependence on σ, λ, and ϵ is optimal. The only room for improvement, therefore, is in constants involving d.

THEOREM 6.1. *Pick any $0 < \epsilon < 1/2$, any $d > 1$, and any $\sigma, \lambda > 0$ such that $\lambda v_{d-1}\sigma^d < 1/120$. Then there exist: an input space $\mathcal{X} \subset \mathbb{R}^d$; a finite family of densities $F = \{f_i\}$ on $\mathcal{X}$; subsets $A_i, A_i', S_i \subset \mathcal{X}$ such that A_i and A_i' are (σ, ϵ)-separated by S_i for density f_i, and $\inf_{x\in A_{i,\sigma}\cup A_{i,\sigma}'} f_i(x) \geq \lambda$, with the following additional property.*

Consider any algorithm that is given $n \geq 100$ i.i.d. samples X_n from some $f_i \in F$ and, with probability at least $3/4$, outputs a tree in which the smallest cluster containing $A_i \cap X_n$ is disjoint from the smallest cluster containing $A_i' \cap X_n$. Then

$$n \;\geq\; \frac{C_2}{v_d\sigma^d\lambda\epsilon^2 d^{1/2}}\log\frac{1}{v_d\sigma^d\lambda d^{1/2}}$$

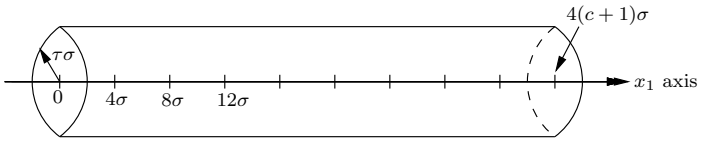
for some absolute constant C_2.

PROOF. Given the parameters $d, \sigma, \epsilon, \lambda$, we will construct a space $\mathcal{X}$ and a finite family of densities $F = \{f_i\}$ on $\mathcal{X}$. We will then argue that any cluster tree algorithm that is able to distinguish (σ, ϵ)-separated clusters must be able, when given samples from some f_i, to determine the identity of I. The sample complexity of this latter task can be lower-bounded using Fano's inequality (Appendix D): it is $\Omega((\log|F|)/\theta)$, for

$$\theta = \max_{i\neq j} K(f_i, f_j),$$

where $K(\cdot,\cdot)$ is Kullback-Leibler divergence.

The support $\mathcal{X}$. The support $\mathcal{X}$ is made up of two disjoint regions: a cylinder $\mathcal{X}_0$, and an additional region $\mathcal{X}_1$ which serves as a repository for excess probability mass. $\mathcal{X}_1$ can be chosen as any Borel set disjoint from $\mathcal{X}_0$. The main region of interest is $\mathcal{X}_0$ and is described as follows in terms of a constant $c > 1$ to be specified. Pick $1 < \tau < \min\left\{\sigma, (2c\cdot v_d\lambda)^{-1/d}\right\} + 1$, such that $\tau^{d-1} \leq 2$. Let B_{d-1} be the unit ball in $\mathbb{R}^{d-1}$, and let $\tau\sigma B_{d-1}$ be this same ball scaled to have radius $\tau\sigma$. The cylinder $\mathcal{X}_0$ stretches along the x_1-axis; its cross-section is $\tau\sigma B_{d-1}$ and its length is $4(c+1)\sigma$ for some $c > 1$ to be specified: $\mathcal{X}_0 = [0, 4(c+1)\sigma] \times \sigma B_{d-1}$. Here is a picture of it:



A family of densities on $\mathcal{X}$. The family F contains $c-1$ densities $f_1, \ldots, f_{c-1}$ which coincide on most of the support $\mathcal{X}$ and differ on parts of $\mathcal{X}_0$. Each density f_i is piecewise constant as described below. Items (ii) and (iv) describe the pieces that are common to all densities in F.

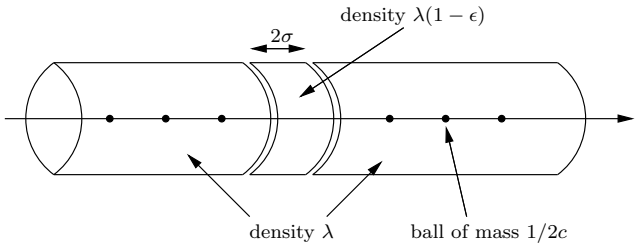
(i) Density $\lambda(1-\epsilon)$ on $(4\sigma i + \sigma, 4\sigma i + 3\sigma) \times \tau\sigma B_{d-1}$.

(ii) Balls of mass $1/(2c)$ centered at locations $4\sigma, 8\sigma, \ldots, 4c\sigma$ along the x_1-axis: each such ball is of radius $\tau - 1 < \sigma$, and the density on these balls is $1/(2c \cdot v_d(\tau-1)^d) \geq \lambda$. We refer to these as *mass balls*.

(iii) Density λ on the remainder of $\mathcal{X}_0$: this is the union of the cylinder segments $[0, 4\sigma i + \sigma] \times \tau\sigma B_{d-1}$ and $[4\sigma i + 3\sigma, 4(c+1)\sigma] \times \tau\sigma B_{d-1}$ minus the mass balls. Since the cross-sectional area of the cylinder is $v_{d-1}(\tau\sigma)^{d-1}$, the total mass here is at most $\lambda\tau^{d-1}v_{d-1}\sigma^d(4(c+1)-2)$.

(iv) The remaining mass is at least $1/2 - 8\lambda v_{d-1}\sigma^d(c+1)$; we will be careful to choose c so that this is nonnegative. The remaining mass is placed on $\mathcal{X}_1$ in some fixed manner that does not vary between densities in F.

Here is a sketch of f_i. The low-density region of width 2σ is centered at $4\sigma i + 2\sigma$ on the x_1-axis, and contains no mass balls.



For any $i \neq j$, the densities f_i and f_j differ only on the cylindrical sections $(4\sigma i+\sigma, 4\sigma i+3\sigma)\times\sigma B_{d-1}$ and $(4\sigma j+\sigma, 4\sigma j+3\sigma)\times\sigma B_{d-1}$, which are disjoint, contain no mass ball, and each have volume $2\tau^{d-1}v_{d-1}\sigma^d$. Thus

$$\begin{aligned} K(f_i, f_j) &= 2\tau^{d-1}v_{d-1}\sigma^d \left(\lambda \log \frac{\lambda}{\lambda(1-\epsilon)} + \lambda(1-\epsilon)\log\frac{\lambda(1-\epsilon)}{\lambda}\right) \\ &= 2\tau^{d-1}v_{d-1}\sigma^d\lambda(-\epsilon\log(1-\epsilon)) \quad \leq \quad \frac{8}{\ln 2}v_{d-1}\sigma^d\lambda\epsilon^2 \end{aligned}$$

(using $\ln(1-x) \geq -2x$ for $0 < x \leq 1/2$). This is an upper bound on the θ in the Fano bound.

Clusters and separators. Now define the clusters and separators as follows: for each $1 \leq i \leq c-1$,

- A_i is the tubular segment $[\sigma, 4\sigma i] \times (\tau-1)\sigma$,
- A_i' is the tubular segment $[4\sigma(i+1), 4(c+1)\sigma - \sigma] \times (\tau-1)\sigma$, and
- $S_i = \{4\sigma i + 2\sigma\} \times \sigma B_{d-1}$ is the cross-section of the cylinder at location $4\sigma i + 2\sigma$.

Thus A_i and A_i' are d-dimensional sets while S_i is a $(d-1)$-dimensional set. It can be seen that, for density f_i, A_i and A_i' are (σ, ϵ)-separated, and $\inf_{x \in A_{i,\sigma} \cup A'_{i,\sigma}} f_i(x) \geq \lambda$.

Now that the various structures are defined, we still need to argue that if an algorithm is given a sample X_n from some f_i (where i is unknown), and is able to separate $A_i \cap X_n$ from $A_i' \cap X_n$, then it can effectively infer the identity of i. This has sample complexity $\Omega((\log c)/\theta)$.

Let's set c to be a small constant, say $c = 6$. Then, even a small sample X_n of $n \geq 100$ points is likely (with probability at least $3/4$, say), to contain points from all of the c mass balls, each of which has mass $1/(2c)$. Suppose the algorithm even knows in advance that the underlying density is one of the $c-1$ choices in F, and is subsequently able (with probability at least $3/4$) to separate A_i from A_i'. To do this, it must connect all the points from mass balls within A_i, and all the points from mass balls within A_i', and yet keep these two groups apart. In short, this algorithm must be able to determine (with overall probability at least $1/2$) the segment $(4\sigma i + \sigma, 4\sigma i + 3\sigma)$ of lower density, and hence the identity of i.

We can thus apply Fano's inequality to conclude that we need

$$n \;>\; \frac{\frac{1}{2}\log(c-1) - 1}{\theta} \;\geq\; \frac{(\frac{1}{2}\log 5 - 1)\ln 2}{8 v_{d-1} \sigma^d \lambda \epsilon^2} \;\geq\; \frac{C_2}{v_d \sigma^d \lambda \epsilon^2 d^{1/2}}$$

for some absolute constant C_2. The last equality comes from the formula $v_d = \pi^{d/2}/\Gamma((d/2)+1)$, whereupon $v_{d-1} = O(v_d d^{1/2})$.

This is almost the bound in the theorem statement, short a logarithmic term. To finish up, we now switch to a larger value of c:

$$c \;=\; \left\lfloor \frac{1}{16 v_{d-1} \sigma^d \lambda} - 1 \right\rfloor,$$

and apply the same construction. We have already established that we need $n = \Omega(c/\epsilon^2)$ samples, so assume n is at least this large. Then, for small enough ϵ, it is very likely that when the underlying density is f_i, the sample X_n will contain the four point masses at 4σ, $4\sigma i$, $4\sigma(i+1)$, and $4(c+1)\sigma$. Therefore, the clustering algorithm must connect the point at 4σ to

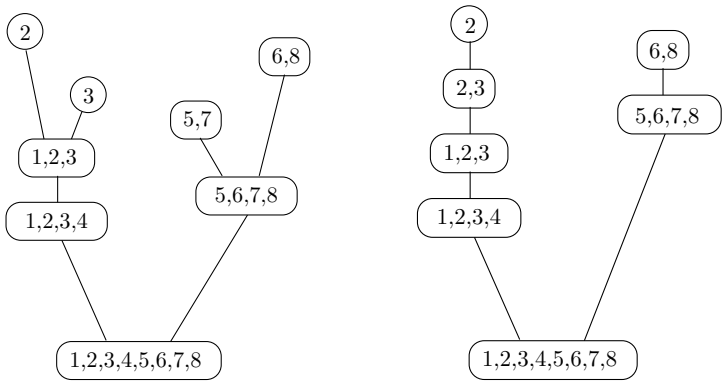


FIG 7. *A sample from the density of Figure 6 contains four points* $(1, 2, 3, 4)$ *from cluster* A *and four* $(5, 6, 7, 8)$ *from cluster* C*. Left: A cluster tree that correctly distinguishes* A *from* C*. Right: A better alternative that avoids fragmenting* A *and* C*.*

that at $4\sigma i$ and the point at $4\sigma(i+1)$ to that at $4(c+1)\sigma$, while keeping the two groups apart. Therefore, this algorithm can determine i. Applying Fano's inequality gives $n = \Omega((\log c)/\theta)$, which is the bound in the theorem statement. □

7. Pruning. Hartigan's notion of consistency (Definition 2.3) requires distinct clusters to be distinguished, but does not guard against fragmentation within a cluster. Consider, for instance, the density shown in Figure 6. Under Hartigan-consistency, in the limit, the cluster tree must include a cluster that contains all of A and a separate, disjoint cluster that contains all of C. But the tree is allowed to break A into further subregions. To be concrete, suppose we draw a sample from that density and receive four points from each of A and C. Figure 7, left, shows a possible cluster tree on these samples that meets the consistency requirement. However, we'd prefer the one on the right. Formally we want to avoid or remove *false clusters* as defined below.

DEFINITION 7.1. *Let* A_n *and* A'_n *be the vertices of two separate connected components (potentially at different levels) in the cluster tree returned by an algorithm. We call* A_n *and* A'_n false clusters *if they are part of the same connected component of the level set* $\{x : f(x) \geq \min_{x' \in A_n \cup A'_n} f(x')\}$.

This problem is generally addressed in the literature by making assumptions about the *size* of true clusters. Real clusters are assumed to be large in some sense, for instance in terms of their mass (Maier et al., 2009), or

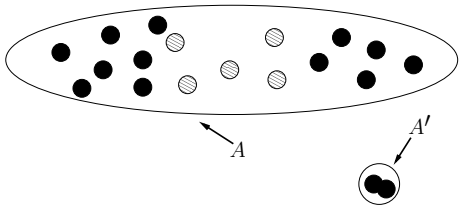


FIG 8. *Depicted are samples from two connected components A and A' of some level set. Suppose that only the black samples appear at level r in the empirical tree: the grey samples have $r_k(x) > r$. Then in G_r, samples from A appear as two large clusters. If we were to just go by size, we would be tempted to preserve this spurious partition and possibly to remove the smaller cluster of samples from A'. Our pruning method, however, will connect the two groups from A and maintain the cluster from A'.*

Pruning of level r.

- Set
$$\tilde{\lambda}_r = \frac{1}{v_d r^d}\left(\frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}\right) - \tilde{\epsilon}.$$
- Connect any two components of $\mathbb{C}_n(r)$ that belong to the same connected component in $\mathbb{C}_n(r(\max(\tilde{\lambda}_r, 0)))$.

FIG 9. *An algorithm for pruning G_r or G_r^{NN}, applied for every r. It assumes a tuning parameter $\tilde{\epsilon} > 0$. Recall from Definition 4.4 that for any $\lambda > 0$, we take $r(\lambda)$ to be the value of r for which $v_d r^d \lambda = \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}$.*

excess mass[1] (Stuetzle and Nugent, 2009). However, relying on size can be misleading in practice, as is illustrated in Figure 8. It turns out that, building on the results of the previous sections, there is a simple way to treat spurious clusters independent of their size.

7.1. *Intuition.* The pruning procedure of Figure 9 consists of a simple lookup: it reconnects components at level r if they are part of the same component at some level $r' > r$, where r' is a function of a tuning parameter $\tilde{\epsilon} \geq 0$. The larger $\tilde{\epsilon}$ is, the more aggressive the pruning.

The pruning procedure builds upon the same intuition as for the procedure of Kpotufe and von Luxburg (2011), however it differs in its ability to handle either cluster-tree algorithms, and works under significantly milder conditions than that of Kpotufe and von Luxburg (2011). The intuition is the following. Suppose $A_n, A'_n \subset X_n$ are not connected at some level r in the empirical tree (before pruning), but ought to be: they belong to the same

[1]The excess mass of a component A at level λ is generally defined as $\int_A (f(x) - \lambda)\, dx$.

component A of $\mathbb{C}(\lambda)$, where $\lambda = \min\{f(x) : x \in A_n \cup A'_n\}$. Then, key sample points from A that would have connected them are missing at level r in the empirical tree (Figure 8). These points have $r_k(x)$ greater than r, but probably not much greater. Looking at a nearby level $r' > r$, we will find A_n, A'_n connected and thus detect the situation.

The above intuition is likely to extend to cluster tree procedures other than the ones discussed here. The main requirement on the cluster tree estimate is that points in A (as discussed above) be connected at some nearby level in the tree.

7.2. *Separation.* The pruning procedure increases connectivity, but we must make sure that it isn't too zealous in doing so: clusters that are sufficiently separated should not be merged. We now will require a bit more separation between two sets A and A' in order to keep them apart in the empirical tree. As might be expected, how much more separation depends on the pruning parameter $\tilde{\epsilon}$. The higher $\tilde{\epsilon}$, the more aggressive the pruning, and the greater the separation requirement for detecting distinct clusters. The following lemma builds on Corollary 4.5.

LEMMA 7.2. *Assume E_o. Consider two sets $A, A' \subset \mathcal{X}$, and let $\lambda = \inf_{x \in A_\sigma \cup A'_\sigma} f(x)$. Suppose there exists a separator set S such that*

- *Any path in $\mathcal{X}$ from A to A' intersects S.*
- $\sup_{x \in S_\sigma} f(x) < (1-2\epsilon)\lambda - \tilde{\epsilon}$.

Then $A \cap X_n$ and $A' \cap X_n$ are in separate components of $\mathbb{C}_n(r(\lambda))$ after pruning, provided $k \geq 4C_\delta^2(d/\epsilon^2)\log n$ and

$$v_d\,(2\sigma/(\alpha+2))^d\,((1-\epsilon)\lambda - \tilde{\epsilon}) > \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}.$$

PROOF. Let r denote $r(\lambda)$ and recall from the definitions of $r(\lambda)$ and $\tilde{\lambda}_r$ (Figure 9) that

$$\begin{aligned}\tilde{\lambda}_r &= \lambda\left(\frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}\right)\left(\frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}\right)^{-1} - \tilde{\epsilon} \\ &\geq \left(1 - 2\frac{C_\delta\sqrt{d\log n}}{\sqrt{k}}\right)\lambda - \tilde{\epsilon} \geq (1-\epsilon)\lambda - \tilde{\epsilon}.\end{aligned}$$

The final term, call it λ', is ≥ 0 by the hypotheses of the lemma. Since $\tilde{\lambda}_r \geq \lambda'$ we have $r(\lambda') \geq r(\tilde{\lambda}_r)$. Thus we just have to show that $A \cap X_n$ and $A' \cap X_n$ are in separate components of $\mathbb{C}_n(r(\lambda'))$. To this end, notice

that, under our assumptions on A and A', these two sets belong to separate components of $\{x \in \mathcal{X} : f(x) \geq \lambda'\}$; in fact

$$\sup_{x \in S_\sigma} f(x) \leq (1-2\epsilon)\lambda - \tilde{\epsilon} \leq (1-\epsilon)\lambda'.$$

Moreover, the final requirement of the lemma statement can be rewritten as $r(\lambda') < 2\sigma/(\alpha+2)$. The argument of Lemma 4.3(c) then implies that $A \cap X_n$ is disconnected from $A' \cap X_n$ in $\mathbb{C}_n(r(\lambda'))$ and thus in $\mathbb{C}_n(r(\tilde{\lambda}_r))$, and hence also at level r after pruning. □

7.3. *Connectedness.* We now turn to the main result of this section, namely that the pruning procedure reconnects incorrectly fragmented clusters. Recall the intuition detailed above. We first have to argue that points with similar density make their first appearance at nearby levels r of the empirical tree. From the analysis of the previous sections, we know that a point x is present at level $r(f(x))$, roughly speaking. We now need to show that it cannot appear at a level too much smaller than this.

These assertions about single points are true only if the density doesn't vary too dramatically in their vicinity. In what follows, we will quantify the smoothness at scale σ by the constant

$$L_\sigma \;=\; \sup_{\|x-x'\| \leq \sigma} \left| f(x) - f(x') \right|.$$

LEMMA 7.3. *Assume E_o. Pick any x and let $f_\sigma(x) = \inf_{x' \in B(x,\sigma)} f(x')$. Suppose*

$$v_d(\sigma/2)^d(f_\sigma(x) + L_\sigma) \geq \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n},$$

we then have

$$v_d r_k^d(x)(f_\sigma(x) + L_\sigma) \geq \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}.$$

PROOF. Consider any r such that

$$v_d r^d(f_\sigma(x) + L_\sigma) < \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n} \leq v_d(\sigma/2)^d(f_\sigma(x) + L_\sigma).$$

Then $r \leq \sigma/2$, implying $f(B(x,r)) \leq v_d r^d(f_\sigma(x) + L_\sigma)$. Using the first inequality and Lemma 4.1, we have $f_n(B(x,r)) < k/n$, that is $r < r_k(x)$. □

Next, by combining the above lower-bound on $r_k(x)$ with our previous results on connectedness for both types of algorithms, we obtain the following pruning guarantees.

LEMMA 7.4. *Assume that event E_o holds, and that $\tilde{\epsilon} \geq L_\sigma$. Let A_n and A'_n denote the vertices of two disjoint components of G_r or G_r^{NN} after pruning, for some $r > 0$. Define $\lambda = \inf_{x \in A_n \cup A'_n} f(x)$. Then A_n and A'_n belong to separate components of the level set $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ whenever the following two conditions hold: first,*

$$v_d(\sigma/2)^d(\lambda - L_\sigma) \;\geq\; \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n},$$

and second,

$$k \;\geq\; \begin{cases} 4C_\delta d \log n & \text{for } G_r \\ \max(4C_\delta^2 d\log n, (\Lambda/\lambda)8C_\delta d \log n) & \text{for } G_r^{\mathrm{NN}} \end{cases}$$

PROOF. Let A be any component of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$. We'll show that $A \cap X_n$ is connected in G_r (or G_r^{NN}) after pruning, from which the lemma follows immediately.

Define $\lambda_\sigma = \inf_{x \in A_\sigma} f(x) \geq \inf_{x \in A} f(x) - L_\sigma \geq \lambda - L_\sigma$. Recall from Definition 4.4 that $r(\lambda_\sigma)$ is the value of r for which $v_d r^d \lambda_\sigma = \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}$. The first condition in the lemma statement thus implies that $r(\lambda_\sigma) \leq \sigma/2$. The second condition, together with Theorem 4.7 or Theorem 5.2, implies that $A \cap X_n$ is connected at level $r(\lambda_\sigma)$ of G or G^{NN}.

Next we show that $r(\lambda_\sigma) \leq r(\tilde{\lambda})$, i.e. $\lambda_\sigma \geq \tilde{\lambda}$. Again by the first condition on k, Lemma 7.3 holds for every $x \in A$, implying with little effort that

$$\lambda_\sigma \geq \frac{1}{v_d r^d}\left(\frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}\right) - L_\sigma \geq \frac{1}{v_d r^d}\left(\frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}\right) - \tilde{\epsilon} = \tilde{\lambda}.$$

Thus $A \cap X_n$ is connected at level $r(\tilde{\lambda}_r) \geq r(\lambda_\sigma)$ of G (or G^{NN}), and thus is reconnected when pruning at level r. □

The separation and connectedness results of this section can now be combined into the following theorem.

THEOREM 7.5. *There is an absolute constant C such that the following holds. Pick any $0 < \delta, \epsilon < 1$ and $\tilde{\epsilon} > 0$. Assume Algorithm 1 or 2 is run on a sample X_n of size n drawn from f, with settings*

$$\sqrt{2} \leq \alpha \leq 2 \quad \text{and} \quad k \geq C \cdot \frac{d\log n}{\epsilon^2} \cdot \log^2 \frac{1}{\delta},$$

followed by the pruning procedure with parameter $\tilde{\epsilon}$.

Then the following holds with probability at least $1 - \delta$. Define

$$\lambda_o \;=\; \frac{k}{nv_d(\sigma/2)^d} \cdot \frac{1+\epsilon}{1-\epsilon} + \frac{\tilde{\epsilon}}{1-\epsilon},$$

or in the case of Algorithm 2, the maximum of this quantity and $(\Lambda/k)Cd\log n \cdot \log(1/\delta)$*, where* $\Lambda = \sup_{x\in\mathcal{X}} f(x)$*.*

Recovery of true clusters: *Consider any two sets* $A, A' \subset \mathcal{X}$*, and suppose* $\lambda = \inf_{x\in A_\sigma \cup A'_\sigma} f(x) \geq \lambda_o$*. Suppose there exists a set* S *such that*

- *Any path in* $\mathcal{X}$ *from* A *to* A' *intersects* S*.*
- $\sup_{x\in S_\sigma} f(x) < (1-2\epsilon)\lambda - \tilde{\epsilon}$*.*

Then $A \cap X_n$ *and* $A' \cap X_n$ *are individually connected in* $\mathbb{C}_n(r(\lambda))$*, but lie in two separate components.*

Removal of false clusters: *Assume the pruning parameter satisfies* $\tilde{\epsilon} \geq 2\sup_{\|x-x'\|\leq\sigma} |f(x) - f(x')|$*. Let* A_n *and* A'_n *denote the vertices of two disjoint components in* $\mathbb{C}_n(r)$*, for any* $r > 0$*. If* $\lambda = \inf_{x\in A_n\cup A'_n} f(x) \geq \lambda_o$*, then the two sets of points* A_n *and* A'_n *belong to separate components of the level set* $\{x \in \mathcal{X} : f(x) \geq \lambda\}$*.*

The first part of the above theorem (recovery of true clusters) implies that the pruned tree remains a consistent estimator of the cluster tree, under the same asymptotic conditions as those for Theorem 3.3 and Theorem 3.4, and the additional condition that $\tilde{\epsilon} \to 0$.

The second part of the theorem states some general conditions on $\tilde{\epsilon}$ and λ under which false clusters are removed. To better understand these conditions, let's consider the simple case when f is Hölder-smooth:

$$\exists L, \beta > 0 \text{ such that } \forall x, x' \in \mathcal{X}, \quad \left|f(x) - f(x')\right| \leq L\left\|x - x'\right\|^\beta.$$

Consider $\sigma = (\tilde{\epsilon}/L)^{1/\beta}$ so that we have $\sup_{\|x-x'\|\leq\sigma} |f(x) - f(x')| \leq \tilde{\epsilon}$. Consider $0 < \epsilon < 1/3$. Then any $\lambda > 4\tilde{\epsilon}$ is $\geq \lambda_o$ if k is in the range

$$\frac{\Lambda}{4\tilde{\epsilon}} \cdot C \cdot \frac{d\log n}{\epsilon^2} \cdot \log^2\frac{1}{\delta} \leq k \leq 2^{-d} \cdot v_d \cdot L^{-d/\beta} \cdot \tilde{\epsilon}^{(\beta+d)/\beta} \cdot n.$$

Note that, without knowing the Hölder parameters L and β, we can ensure k is in the above range for any particular $0 < \epsilon < 1/3$, provided n and $k = k(n)$ are sufficiently large, by choosing $\tilde{\epsilon}$ as a function of k (e.g. $k = \Theta(\log^3 n)$ and $\tilde{\epsilon} = \Theta(1/\sqrt{k})$).

Finally, remark that under the above smoothness assumption and choice of σ, ϵ, we can further guarantee that *all* false clusters are removed! We only need to reconnect all components at levels where the minimum f value is at most $4\tilde{\epsilon}$. By Lemma 7.3, for k in the above range, we have $r_k(x) \geq (k/10nv_d\tilde{\epsilon})^{1/d}$ when $f(x) \leq 4\tilde{\epsilon}$. Thus, we just need to reconnect all components at levels $r > (k/10nv_d\tilde{\epsilon})^{1/d}$, and prune all other levels as discussed above. This then guarantees that all false clusters are removed with high probability, while also ensuring that the estimator remains consistent.

8. Final remarks. Both cluster tree algorithms are variations on standard estimators, but carefully control the neighborhood size k and make use of a novel parameter α to allow more edges at every scale r. The analysis relies on α being at least $\sqrt{2}$, and on k being at least $d \log n$. Is it possible to dispense with α (that is, to use $\alpha = 1$) while maintaining this setting of k?

There remains a discrepancy of 2^d between the upper and lower bounds on the sample complexity of building a hierarchical clustering that distinguishes all (σ, ϵ)-separated clusters. Can this gap be closed, and if so, what is needed, a better analysis or a better algorithm?

Finally, unlike with plug-in estimators of the cluster tree, our algorithms encode no knowledge of the dimension of the support. It is therefore likely that our results extend to settings where the distribution is supported on a low-dimensional subspace of $\mathbb{R}^d$.

APPENDIX A: PLUG-IN ESTIMATION OF THE CLUSTER TREE

One way to build a cluster tree is to return $\mathbb{C}_{f_n}$, where f_n is a uniformly consistent density estimate.

LEMMA A.1. *Suppose estimator f_n of density f (on space $\mathcal{X}$) satisfies $\sup_{x \in \mathcal{X}} |f_n(x) - f(x)| \leq \epsilon_n$. Pick any two disjoint sets $A, A' \subset \mathcal{X}$ and define $\Xi = \inf_{x \in A \cup A'} f(x)$ and $\xi = \sup_{A \overset{P}{\rightsquigarrow} A'} \inf_{x \in P} f(x)$. If $\Xi - \xi > 2\epsilon_n$ then A, A' lie entirely in disjoint connected components of $\mathbb{C}_{f_n}(\Xi - \epsilon_n)$.*

PROOF. A and A' are each connected in $\mathbb{C}_{f_n}(\Xi - \epsilon_n)$. But there is no path from A to A' in $\mathbb{C}_{f_n}(\lambda)$ for $\lambda > \xi + \epsilon_n$. □

The problem, however, is that computing the level sets of f_n is usually not an easy task. Hence we adopt a different approach in this paper.

APPENDIX B: CONSISTENCY

The following is a straightforward exercise in analysis.

LEMMA B.1. *Suppose density $f : \mathbb{R}^d \to \mathbb{R}$ is continuous and is zero outside a compact subset $\mathcal{X} \subset \mathbb{R}^d$. Suppose further that for some λ, $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ has finitely many connected components, among them $A \neq A'$. Then there exist $\sigma, \epsilon > 0$ such that A and A' are (σ, ϵ)-separated.*

PROOF. Let $A_1, A_2, \ldots, A_k$ be the connected components of $\{f \geq \lambda\}$, with $A = A_1$ and $A' = A_2$.

First, each A_i is closed and thus compact. To see this, pick any $x \in \mathcal{X} \setminus A_i$. There must be some x' on the shortest path from x to A_i with $f(x') < \lambda$ (otherwise $x \in A_i$). By continuity of f, there is some ball $B(x', r)$ on which $f < \lambda$; thus this ball doesn't touch A_i. Then $B(x, r)$ doesn't touch A_i.

Next, for any $i \neq j$, define $\Delta_{ij} = \inf_{x \in A_i, y \in A_j} \|x - y\|$ to be the distance between A_i and A_j. We'll see that $\Delta_{ij} > 0$. Specifically, define $g : A_i \times A_j \to \mathbb{R}$ by $g(a, a') = \|a - a'\|$. Since g has compact domain, it attains its infimum for some $a \in A_i, a' \in A_j$. Thus $\Delta_{ij} = \|a - a'\| > 0$.

Let $\Delta = \min_{i \neq j} \Delta_{ij} > 0$, and define S to be the set of points at distance exactly $\Delta/2$ from A: $S = \{x \in \mathcal{X} : \inf_{y \in A} \|x - y\| = \Delta/2\}$. S separates A from A'. Moreover, it is closed by continuity of $\|\cdot\|$, and hence is compact. Define $\lambda_o = \sup_{x \in S} f(x)$. Since S is compact, f (restricted to S) is maximized at some $x_o \in S$. Then $\lambda_o = f(x_o) < \lambda$.

To finish up, set $\delta = (\lambda - \lambda_o)/3 > 0$. By uniform continuity of f, there is some $\sigma > 0$ such that f doesn't change by more than δ on balls of radius σ. Then $f(x) \leq \lambda_o + \delta = \lambda - 2\delta$ for $x \in S_\sigma$ and $f(x) \geq \lambda - \delta$ for $x \in A_\sigma \cup A'_\sigma$. Thus S is a $(\sigma, \delta/(\lambda - \delta))$-separator for A, A'. □

APPENDIX C: PROOF DETAILS

C.1. Proof of Lemma 4.1. We start with a standard generalization result due to Vapnik and Chervonenkis; the following version is a paraphrase of Theorem 5.1 of Bousquet et al. (2004).

THEOREM C.1. *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ to $\{0, 1\}$ with VC dimension $d < \infty$, and $\mathbb{P}$ a probability distribution on $\mathcal{X}$. Let $\mathbb{E}$ denote expectation with respect to $\mathbb{P}$. Suppose n points are drawn independently at random from $\mathbb{P}$; let $\mathbb{E}_n$ denote expectation with respect to this sample. Then for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $g \in \mathcal{G}$:*

$$-\min(\beta_n\sqrt{\mathbb{E}_n g}, \beta_n^2 + \beta_n\sqrt{\mathbb{E} g}) \;\leq\; \mathbb{E}g - \mathbb{E}_n g \;\leq\; \min(\beta_n^2 + \beta_n\sqrt{\mathbb{E}_n g}, \beta_n\sqrt{\mathbb{E} g}),$$

where $\beta_n = \sqrt{(4/n)(d \ln 2n + \ln(8/\delta))}$.

By applying this bound to the class $\mathcal{G}$ of indicator functions over balls (or half-balls), we get the following:

LEMMA C.2. *Suppose X_n is a sample of n points drawn independently at random from a distribution f over $\mathcal{X}$. For any set $Y \subset \mathcal{X}$, define $f_n(Y) = |X_n \cap Y|/n$. There is a universal constant $C_o > 0$ such that for any $\delta > 0$,*

with probability at least $1-\delta$*, for any ball (or half-ball)* $B \subset \mathbb{R}^d$*,*

$$f(B) \geq \frac{C_o}{n}\left(d\log n + \log\frac{1}{\delta}\right) \quad \Longrightarrow \quad f_n(B) > 0$$

$$f(B) \geq \frac{k}{n} + \frac{C_o}{n}\left(d\log n + \log\frac{1}{\delta} + \sqrt{k\left(d\log n + \log\frac{1}{\delta}\right)}\right) \quad \Longrightarrow \quad f_n(B) \geq \frac{k}{n}$$

$$f(B) < \frac{k}{n} - \frac{C_o}{n}\left(d\log n + \log\frac{1}{\delta} + \sqrt{k\left(d\log n + \log\frac{1}{\delta}\right)}\right) \quad \Longrightarrow \quad f_n(B) < \frac{k}{n}$$

PROOF. The VC dimension of balls in $\mathbb{R}^d$ is $d+1$, while that of half-balls (each the intersection of a ball and a halfspace) is $O(d)$. The following statements apply to either class.

The bound $f(B) - f_n(B) \leq \beta_n\sqrt{f(B)}$ from Theorem C.1 yields $f(B) > \beta_n^2 \implies f_n(B) > 0$. For the second bound, we use $f(B) - f_n(B) \leq \beta_n^2 + \beta_n\sqrt{f_n(B)}$. It follows that

$$f(B) \geq \frac{k}{n} + \beta_n^2 + \beta_n\sqrt{\frac{k}{n}} \quad \Longrightarrow \quad f_n(B) \geq \frac{k}{n}.$$

For the last bound, we rearrange $f(B) - f_n(B) \geq -(\beta_n^2 + \beta_n\sqrt{f(B)})$ to get

$$f(B) < \frac{k}{n} - \beta_n^2 - \beta_n\sqrt{\frac{k}{n}} \quad \Longrightarrow \quad f_n(B) < \frac{k}{n}.$$

□

Lemma 4.1 now follows immediately, by taking $k \geq d\log n$. Since the uniform convergence bounds have error bars of magnitude $(d\log n)/n$, it doesn't make sense, when using them, to take k any smaller than this.

C.2. Proof of Lemma 4.6. Consider any $x, x' \in A \cap X_n$. Since A is connected, there is a path P in A with $x \overset{P}{\rightsquigarrow} x'$. Fix any $0 < \gamma < 1$. Because the density of A_σ is lower bounded away from zero, it follows by a volume and packing-covering argument that A, and thus P, can be covered by a finite number of balls of diameter γr. Thus we can choose finitely many points $z_1, z_2, \ldots, z_k \in P$ such that $x = z_0$, $x' = z_k$ and $\|z_{i+1} - z_i\| \leq \gamma r$.

Under E_o (Lemma 4.1), any ball centered in A with radius $(\alpha-\gamma)r/2$ contains at least one data point if

$$v_d\left(\frac{(\alpha-\gamma)r}{2}\right)^d \lambda \;\geq\; \frac{C_\delta d\log n}{n}. \tag{C.1}$$

Assume for the moment that this holds. Then, every ball $B(z_i, (\alpha-\gamma)r/2)$ contains at least one point; call it x_i.

By the upper bound on r, each such x_i lies in $A_{\sigma-r}$; therefore, by Lemma 4.1, the x_i are all active in G_r. Moreover, consecutive points x_i are close together:

$$\|x_{i+1} - x_i\| \;\leq\; \|x_{i+1} - z_{i+1}\| + \|z_{i+1} - z_i\| + \|z_i - x_i\| \;\leq\; \alpha r.$$

Thus all edges (x_i, x_{i+1}) exist in G_r, whereby x is connected to x' in G_r.

All this assumes that equation (C.1) holds for some $\gamma > 0$. Taking $\gamma \to 0$ gives the lemma.

APPENDIX D: FANO'S INEQUALITY

Consider the following game played with a predefined, finite class of distributions $F = \{f_1, \ldots, f_\ell\}$, defined on a common space $\mathcal{X}$:

- Nature picks $I \in \{1, 2, \ldots, \ell\}$.
- Player is given n i.i.d. samples $X_1, \ldots, X_n$ from f_i.
- Player then guesses the identity of I.

Fano's inequality (Cover and Thomas, 2005, Yu, 1997) gives a lower bound on the number of samples n needed to achieve a certain success probability. It depends on how similar the distributions f_i are: the more similar, the more samples are needed. Define $\theta = \frac{1}{\ell^2}\sum_{i,j=1}^{\ell} K(f_i, f_j)$ where $K(\cdot)$ is KL divergence. Then n needs to be $\Omega((\log \ell)/\theta)$. Here's the formal statement.

THEOREM D.1 (Fano). *Let $g : \mathcal{X}^n \to \{1, 2, \ldots, \ell\}$ denote Player's computation. If Nature chooses I uniformly at random from $\{1, 2, \ldots, \ell\}$, then for any $0 < \delta < 1$,*

$$n \leq \frac{(1-\delta)(\log_2 \ell) - 1}{\theta} \quad \Longrightarrow \quad \Pr(g(X_1, \ldots, X_n) \neq I) \geq \delta.$$

REFERENCES

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 3176:169–207, 2004.

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

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 2005.

S. Dasgupta and Y. Freund. Random projection trees for vector quantization. *IEEE Transactions on Information Theory*, 55(7):3229–3242, 2009.

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

S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. In *International Conference on Machine Learning*, 2011.

M. Maier, M. Hein, and U. von Luxburg. Optimal construction of k-nearest neighbor graphs for identifying noisy clusters. *Theoretical Computer Science*, 410:1749–1764, 2009.

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

D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1):135–140, 1981.

W. Polonik. Measuring mass concentrations and estimating density contour clusters – an excess mass approach. *Annals of Statistics*, 23(3):855–881, 1995.

P. Rigollet and R. Vert. Fast rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009.

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

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

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

I. Steinwart. Adaptive density level set clustering. In *24th Annual Conference on Learning Theory*, 2011.

W. Stuetzle and R. Nugent. Clustering with confidence: a binning approach. In *Conference of the International Federation of Classification Societies*, 2009.

W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2):397–418, 2010.

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

D. Wishart. Mode analysis: a generalization of nearest neighbor which reduces chaining effects. In *Proceedings of the Colloquium on Numerical Taxonomy held in the University of St. Andrews*, pages 282–308, 1969.

M.A. Wong and T. Lane. A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society Series B*, 45(3):362–368, 1983.

B. Yu. Assouad, Fano and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997.

9500 Gilman Drive
La Jolla, CA 92093, USA
E-mail: kamalika@cs.ucsd.edu
dasgupta@cs.ucsd.edu

6045 S. Kenwood Ave.
Chicago, Illinois 60637 USA
E-mail: samory@ttic.edu

Vogt-Klln-Str. 30
22527 Hamburg, Germany
luxburg@informatik.uni-hamburg.de