JMLR: Workshop and Conference Proceedings vol 40:1–19, 2015

Beyond Hartigan Consistency: Merge Distortion Metric for Hierarchical Clustering

Justin Eldridge ELDRIDGE@CSE.OHIO-STATE.EDU
Mikhail Belkin MBELKIN@CSE.OHIO-STATE.EDU
Yusu Wang YUSU@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277, USA

Abstract

Hierarchical clustering is a popular method for analyzing data which associates a tree to a dataset. Hartigan consistency has been used extensively as a framework to analyze such clustering algorithms from a statistical point of view. Still, as we show in the paper, a tree which is Hartigan consistent with a given density can look very different than the correct limit tree. Specifically, Hartigan consistency permits two types of undesirable configurations which we term *over-segmentation* and *improper nesting*. Moreover, Hartigan consistency is a limit property and does not directly quantify difference between trees.

In this paper we identify two limit properties, *separation* and *minimality*, which address both over-segmentation and improper nesting and together imply (but are not implied by) Hartigan consistency. We proceed to introduce a *merge distortion metric* between hierarchical clusterings and show that convergence in our distance implies both separation and minimality. We also prove that uniform separation and minimality imply convergence in the merge distortion metric. Furthermore, we show that our merge distortion metric is stable under perturbations of the density.

Finally, we demonstrate applicability of these concepts by proving convergence results for two clustering algorithms. First, we show convergence (and hence separation and minimality) of the recent robust single linkage algorithm of Chaudhuri and Dasgupta (2010). Second, we provide convergence results on manifolds for topological split tree clustering.

Keywords: hierarchical clustering, Hartigan consistency, metric distortion

1. Introduction

Hierarchical clustering is an important class of techniques and algorithms for representing data in terms of a certain tree structure (Jain and Dubes, 1988). When data are sampled from a probability distribution, one needs to study the relationship between trees obtained from data samples to the infinite tree of the underlying probability density. This question was first explored in Hartigan (1975), which introduced the notion of *high-density clusters*. Specifically, given density function $f : \mathcal{X} \to \mathbb{R}$, the high-density clusters are defined to be the connected components of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ for some λ. The set of all clusters forms a hierarchical structure known as the *density cluster tree* of f. The natural notion of consistency for finite density estimators is is to require that any two high density clusters are also separate in the finite tree given enough samples. This notion was introduced in Hartigan (1981) and is known as *Hartigan consistency*. Still, while clearly desirable, it is well known that Hartigan consistency does not fully capture the properties of convergence that one would *a priori* expect. In particular, it does not exclude trees which are very different from the underlying probability distribution.

In this paper we identify two distinct undesirable configuration types permitted by Hartigan consistency, *over-segmentation* (identified as the problem of *false clusters* in Chaudhuri et al., 2014) and *improper nesting*, and show how both of these result from clusters merging at the wrong level. To address these issues we propose two basic properties for hierarchical cluster convergence: *minimality* and *separation*. Together they imply Hartigan consistency and, furthermore, rule out "improper" configurations. We proceed to introduce a *merge distortion* metric on clustering trees and show that convergence in the metric implies both separation and minimality. Moreover, we demonstrate that uniform versions of these properties are in fact equivalent to metric convergence. We note that the introduction of a quantifiable *merge distortion* metric also addresses another issue with Hartigan consistency, which is a limit property of clustering algorithms and is not quantifiable as such. We also prove that the merge distortion metric is robust to small perturbations of the density.

Still, attempts to formulate some intuitively desirable properties of clustering have led to well-known impossibility results, such as those proven by Kleinberg (2003). In order to show that our definitions correspond to actual objects, and, furthermore, to realistic algorithms, we analyze the robust single linkage clustering proposed by Chaudhuri and Dasgupta (2010). We prove convergence of that algorithm under our merge distortion metric and hence show that it satisfies separation and minimality conditions. We also propose a topological split tree algorithm for hierarchical clustering (based on the algorithm introduced by Chazal et al. (2013) for flat clustering) and demonstrate its convergence on Riemannian manifolds.

Previous work. The problem of devising an algorithm which provably converges to the true density cluster tree in the sense of Hartigan has a long history. Hartigan (1981) proved that single linkage clustering is *not* consistent in dimensions larger than one. Previous to this, Wishart (1969) had introduced a more robust version of single linkage, but its consistency had not been known. Stuetzle and Nugent (2010) introduced another generalization of single-linkage designed to estimate the density cluster tree, but again consistency was not established. Recently, however, two distinct consistent algorithms have been introduced: The robust single linkage algorithm of Chaudhuri and Dasgupta (2010), and the tree pruning method of Kpotufe and Luxburg (2011). Both algorithms are analyzed together, along with a pruning extension, in Chaudhuri et al. (2014). The robust single linkage algorithm was generalized in Balakrishnan et al. (2013) to densities supported on a Riemannian submanifold of $\mathbb{R}^d$. We analyze the algorithm of Chaudhuri and Dasgupta (2010) in Section 6. Chaudhuri and Dasgupta (2010) provide several theorems which make precise the sense in which clusters are connected and separated at each step of the robust single linkage algorithm. This paper translates their results to our formalism, thereby proving that robust single linkage converges to the density cluster tree in the merge distortion metric.

A central contribution of this paper will be to introduce notions which extend Hartigan consistency, and are desirable properties of any algorithm which estimates the density cluster tree. In a related direction, Kleinberg (2003) outlined three desirable of a clustering method, and proved that no method satisfying all three exists. Carlsson and Mémoli (2010) showed that, surprisingly, if one allows clustering methods to return a hierarchical decomposition rather than a fixed partition of a space, a unique method satisfying three axioms related to Kleinberg's exists, and is in fact single linkage agglomerative clustering. Additionally, Carlsson and Mémoli (2010) proved that single linkage clustering recovers the support of a density in the limit.

2. Preliminaries and definitions

A *clustering* $\mathcal{C}$ of a set X is the organization of its elements into a collection of subsets of X called *clusters*. In general, clusters may overlap or be disjoint. If the collection of clusters exhibits nesting behavior (to be made precise below), the clustering is called *hierarchical*. The nesting property permits us to think of a hierarchical clustering as a tree of clusters, henceforth called a *cluster tree*.

Definition 1 (Cluster tree) *A* cluster tree *(hierarchical clustering) of a set X is a collection $\mathcal{C}$ of subsets of X s.t. $X \in \mathcal{C}$ and $\mathcal{C}$ has hierarchical structure. That is, if $C, C' \in \mathcal{C}$ such that $C \neq C'$, then $C \cap C' = \emptyset$, or $C \subset C'$ or $C' \subset C$. Each element C of $\mathcal{C}$ is called a* cluster*. Each cluster C is a node in the tree. The* descendants *of C are those clusters $C' \in \mathcal{C}$ such that $C' \subset C$. Every cluster in the tree except for X itself is a descendant of X, hence X is called the* root *of the cluster tree.*

Note that our definition of a cluster tree does not assume that either the set of objects X or the collection of clusters $\mathcal{C}$ are finite or even countable. Hierarchical clustering is commonly formulated as a sequence of nested partitions of X (Jain and Dubes, 1988, see), culminating in the partition of X into singleton clusters. Our formulation differs in that it is a sequence of nested partitions of *subsets* of X. Notably, we don't impose the requirement that $\{x\}$ appear as a cluster for every x.

Given a density f supported on $\mathcal{X} \subset \mathbb{R}^d$, a natural way to cluster $\mathcal{X}$ is into regions of high density. Hartigan (1975) made this notion precise by defining a *high-density cluster* of f to be a connected component of the superlevel set $\{f \geq \lambda\} := \{x \in \mathcal{X} : f(x) \geq \lambda\}$ for any $\lambda \geq 0$. It is clear that this clustering exhibits the nesting property: If C is a connected component of $\{f \geq \lambda\}$, and C' is a connected component of $\{f \geq \lambda'\}$, then either $C \subseteq C'$, $C' \subseteq C$, or $C \cap C' = \emptyset$. We can therefore interpret the set of all high-density clusters of a density f as a cluster tree:

Definition 2 (Density cluster tree of f) *Let $\mathcal{X} \subset \mathbb{R}^d$ and consider any $f : \mathcal{X} \to \mathbb{R}$. The* density cluster tree *of f, written $\mathcal{C}_f$, is the cluster tree whose nodes (clusters) are the connected components of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ for some $\lambda \geq 0$.*

We note that the density cluster tree of f is closely related to the so-called *split tree* studied in the computational geometry and topology literature as a variant of the *contour tree*; see e.g, (Carr et al., 2003). We discuss a split tree-based approach to estimating the density cluster tree in Section 7.

In practice we do not have access to the true density f, but rather a finite collection of samples $X_n \subset \mathcal{X}$ drawn from f. We may attempt to recover the structure of the density cluster tree $\mathcal{C}_f$ by applying a hierarchical clustering algorithm to the sample, producing a discrete cluster tree $\hat{\mathcal{C}}_{f,n}$ whose clusters are subsets of X_n. In order to discuss the sense in which the discrete estimate $\hat{\mathcal{C}}_{f,n}$ is consistent with the density cluster tree $\mathcal{C}_f$ in the limit $n \to \infty$, Hartigan (1981) introduced a notion of convergence which has since been referred to as *Hartigan consistency*. We follow Chaudhuri and Dasgupta (2010) in defining Hartigan consistency in terms of the density cluster tree:

Definition 3 (Hartigan consistency) *Suppose a sample $X_n \subset \mathcal{X}$ of size n is used to construct a cluster tree $\hat{\mathcal{C}}_{f,n}$ that is an estimate of $\mathcal{C}_f$. For any sets $A, A' \subset \mathcal{X}$, let A_n (respectively A'_n) denote the smallest cluster of $\hat{\mathcal{C}}_{f,n}$ containing $A \cap X_n$ (respectively, $A' \cap X_n$). We say $\hat{\mathcal{C}}_{f,n}$ is consistent if, whenever A and A' are different connected components of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ for some $\lambda > 0$, $\Pr(A_n$ is disjoint from $A'_n) \to 1$ as $n \to \infty$.*

In what follows, it will be useful to talk about the "height" at which two points in a clustering merge. To motivate our definition, consider the two points a and a' which sit on the surface of the density depicted in Figure 1. Intuitively, a sits at height $f(a)$ on the surface, while a' sits at $f(a')$. If we look at the superlevel set $\{f \geq f(a)\}$, we see that a and a' lie in two different high-density clusters. As we sweep $\lambda < f(a)$, the disjoint components of $\{f \geq \lambda\}$ containing a and a' grow, until they merge at height μ. We therefore say that the *merge height* of a and a' is μ.

We may also interpret the situation depicted in Figure 1 in the language of the density cluster tree. Let A be the connected component of $\{f \geq f(a)\}$ which contains a, and let A be the component of $\{f \geq f(a')\}$ containing a'. Recognize that A and A' are nodes in the density cluster tree. As we walk the unique path from A to the root, we eventually come across a node M which contains both a and a'. Note that M is a connected component of the superlevel set $\{f \geq \mu\}$. It is desirable to assign a height to the entire cluster M, and a natural choice is therefore μ.

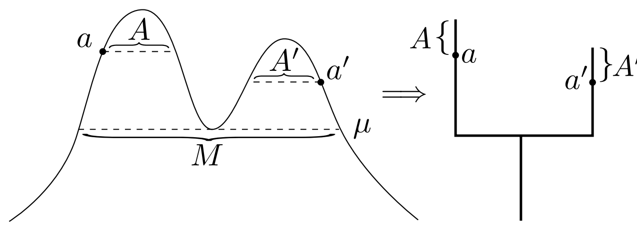

Figure 1

We extend this intuition to cluster trees which may not, in general, be associated with a density f by introducing the concept of a height function:

Definition 4 (Cluster tree with height function) *A cluster tree with a height function is a triple* $\mathsf{C} = (X, \mathcal{C}, h)$*, where* X *is a set of objects,* $\mathcal{C}$ *is a cluster tree of* X*, and* $h : X \to \mathbb{R}$ *is a height function mapping each point in* X *to a "height". Furthermore, we define the height of a cluster* $C \in \mathcal{C}$ *to be the lowest height of any point in the cluster. That is,* $h(C) = \inf_{x \in C} h(x)$*. Note that the nesting property of* $\mathcal{C}$ *implies that if* C' *is a descendant of* C *in the cluster tree, then* $h(C') \geq h(C)$*.*

We will be consistent in using C_f to denote the density cluster tree of f equipped with height function f. That is, $\mathsf{C}_f = (\mathcal{X}, \mathcal{C}_f, f)$. Armed with these definitions, we may precisely discuss the sense in which points – and, by extension, clusters – are connected at some level of a tree:

Definition 5 *Let* $\mathsf{C} = (X, \mathcal{C}, h)$ *be a hierarchical clustering of* X *equipped with height function* h*.*

1. *Let* $x, x' \in X$*. We say that* x *and* x' *are* connected at level λ *if there exists a* $C \in \mathcal{C}$ *with* $x, x \in C$ *such that* $h(C) \geq \lambda$*. Otherwise,* x *and* x' *are* separated at level λ.
2. *A subset* $S \subset X$ *is* connected at level λ *if for any* $s, s' \in S$*,* s *and* s' *are connected at level* λ*.*
3. *Let* $S \subset X$ *and* $S' \subset X$*. We say that* S *and* S' *are* separated at level λ *if for any* $s \in S$*,* $s' \in S'$*,* s *and* s' *are separated at level* λ*.*

We can now formalize the notion of *merge height*:

Definition 6 (Merge height) *Let* $\mathsf{C} = (X, \mathcal{C}, h)$ *be a hierarchical clustering equipped with a height function. Let* $x, x' \in X$*, and suppose that* M *is the smallest cluster of* $\mathcal{C}$ *containing both* x *and* x'*. That is, if* M' *is a proper subset of* M*, then* $x \notin M'$ *or* $x' \notin M'$*. We define the* merge height *of* x *and* x' *in* C*, written* $m_{\mathsf{C}}(x, x')$*, to be the height of the cluster* M *in which the two points merge, i.e.,* $m_{\mathsf{C}}(x, x') = h(M)$*. If* $S \subset X$*, we define the* merge height *of* S *to be the* $\inf_{(s,s') \in S \times S} m_{\mathsf{C}}(s, s')$*.*

In what follows, we argue that a natural and advantageous definition of convergence to the true density cluster tree is one which requires that, for any two points x, x', the merge height of x and x' in an estimate, $m_{\hat{\mathsf{C}}_{f,n}}(x, x')$, approaches the true merge height $m_{\mathsf{C}_f}(x, x')$ in the limit $n \to \infty$.

3. Notions of consistency

In this section we argue that while Hartigan consistency is a desirable property, it is not sufficient to guarantee that an estimate captures the true cluster tree in a sense that matches our intuition. We first illustrate the issue by giving an example in which an algorithm is Hartigan consistent, yet produces results which are very different from the true cluster tree. We then introduce a new, stronger notion of consistency which directly addresses the weaknesses of Hartigan's definition.

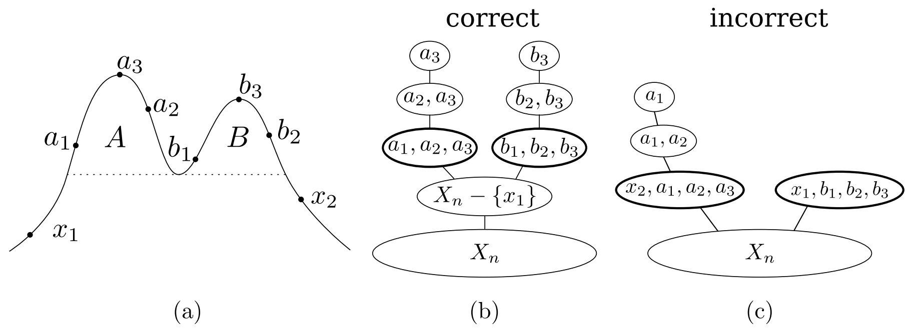

Figure 2: (c) is Hartigan consistent, yet looks rather different than the true tree.

The insufficiency of Hartigan consistency. An algorithm which is Hartigan consistent can nevertheless produce results which are quite different than the true cluster tree. Figure 2 illustrates the issue. Figure 2(a) depicts a two-peaked density f from which the finite sample X_n is drawn. The two disjoint clusters A and B are also shown. The two trees to the right represent possible outputs of clustering algorithms attempting to recover the hierarchical structure of f. Figure 2(b) depicts what we would intuitively consider to be an ideal clustering of X_n, whereas Figure 2(c) shows an undesirable clustering which does not match our intuition behind the density cluster tree of f.

First, note that while the two clusterings are very different, both satisfy Hartigan consistency. Hartigan's notion requires only separation: The smallest empirical cluster containing $A \cap X_n$ must be disjoint from the smallest empirical cluster containing $B \cap X_n$ in the limit. The smallest empirical cluster containing $A \cap X_n$ in the undesirable clustering is $A_n := \{x_2, a_1, a_2, a_3\}$, whereas the smallest containing $B \cap X_n$ is $B_n := \{x_1, b_1, b_2, b_3\}$. A_n and B_n are clearly disjoint, and so Hartigan consistency is not violated. In fact, the undesirable tree separates any pair of disjoint clusters of f, and therefore represents a possible output of an algorithm which is Hartigan consistent despite being quite different from the true tree.

We will show that the undesirable configurations of Figure 2(c) arise because Hartigan consistency does not place strong demands on the level at which a cluster should be connected. Consider a cluster A occurring at level λ of the true density, and let A_n be the smallest empirical cluster containing all of $A \cap X_n$. In the ideal case, an algorithm would perfectly recover A such that $A_n = A \cap X_n$. It is much more likely, however, that A_n contains "extra" points from outside of A. Hartigan consistency places one constraint on the nature of these extra points: They may not belong to some other disjoint cluster of f. However, Hartigan's notion allows A_n to contain points from clusters which are *not* disjoint from A. By their nature, these points must be of density less than λ. If A_n contains such extra points, then $A \cap X_n$ is *separated* at level λ, and in fact only becomes connected at level $\min_{a \in A_n} f(a) < \delta$. Therefore, permitting $A \cap X_n$ to become connected at a level lower than λ is equivalent to allowing "extra" points of density $< \lambda$ to be contained within A_n.

The undesirable configurations depicted in Figure 2(c) can be divided into two distinct categories, which we term *over-segmentation* and *improper nesting*. Either of these issues may exist independently of the other, and both are symptoms of allowing clusters to become connected at lower levels than what is appropriate.

Over-segmentation occurs when an algorithm fragments a true cluster, returning empirical clusters which are disjoint at level λ but are in actuality part of the same connected component of $\{f \geq \lambda\}$. The problem is recognized in the literature by Chaudhuri et al. (2014), who refer to it as the presence of *false clusters*. Figure 2(c) demonstrates over-segmentation by including the clusters $A_n := \{x_2, a_1, a_2, a_3\}$ and $B_n := \{x_1, b_1, b_2, b_3\}$. A_n and B_n are disjoint at level $f(x_1)$, though both are in actuality contained within the same connected component of $\{f \geq f(x_1)\}$.

It is clear that over-segmentation is a direct result of clusters connecting at the incorrect level. The severity of the issue is determined by the difference between the levels at which the cluster connects in the density cluster tree and the estimate. That is, if A is connected at λ in the density cluster tree, but $A \cap X_n$ is only connected at $\lambda - \delta$ in the empirical clustering, then the larger δ the greater the extent to which A is over-segmented.

Improper nesting occurs when an empirical cluster C_n is the smallest cluster containing a point x, and $f(x) > \min_{c \in C_n} f(c)$. The clustering in Figure 2(c) displays two instances of improper nesting. First, the left branch of the cluster tree has improperly nested the cluster $\{a_1, a_2\}$, as it is the smallest cluster containing a_2, yet $f(a_1) < f(a_2)$. The right branch of the same tree has also been improperly nested in a decidedly "lazier" fashion: the cluster $\{x_1, b_1, b_2, b_3\}$ is the smallest empirical cluster containing each of b_1, b_2, and b_3, despite each being of density greater than $f(x_1)$. Improper nesting is considered undesirable because it breaks the intuition we have about the containment of clusters in the density cluster tree; Namely, if $A \subset A'$ and $a \in A$, $a' \in A'$, then $f(a) \geq f(a')$.

Note that like over-segmentation, improper nesting is caused by a cluster becoming connected at a lower level than is appropriate. For instance, suppose C_n is improperly nested; That is, it is the smallest empirical cluster containing some point x such that $f(x) > \min_{c \in C_n} f(c)$. Let $\tilde{C}$ be the connected component of $\{f \geq f(x)\}$, and let $\tilde{C}_n$ be the smallest empirical cluster containing all of $\tilde{C} \cap X_n$. Then $C_n \subset \tilde{C}_n$ such that $\min_{c \in \tilde{C}_n} f(c) < f(x)$. In other words, $\tilde{C} \cap X_n$ is connected only below $f(x)$.

As previously mentioned, it is not reasonable to demand that a cluster A be perfectly recovered by a clustering algorithm. Rather, if A is connected at level λ in the density cluster tree, we should allow $A \cap X_n$ to be first connected at a level $\lambda - \delta$ in the estimate, for some small positive δ. We make this notion precise with the following definition:

Definition 7 (δ-minimal) *Let A be a connected component of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$, and let $\hat{\mathsf{C}}_{f,n}$ be an estimate of the density cluster tree of f computed from finite sample X_n. A is* δ-minimal *in $\hat{\mathsf{C}}_{f,n}$ if $A \cap X_n$ is connected at level $\lambda - \delta$ in $\hat{\mathsf{C}}_{f,n}$.*

Intuitively, each cluster of the density cluster tree should be δ-minimal in an empirical clustering for as small of a δ as possible. For example, take any sample $x \in X_n$ and let C be the connected component of $\{f \geq f(x)\}$ containing x. Some examination shows that C is 0-minimal in the ideal clustering depicted in Figure 2(b). As the ideal clustering is free from over-segmentation and improper nesting, it stands to reason that a cluster can only exhibit these issues to the extent that it is δ-minimal; The larger δ, the more severely a cluster may be over-segmented or improperly nested.

Minimality and separation. We have identified two senses – over-segmentation and improper nesting – in which a hierarchical clustering method can produce results which are inconsistent with the density cluster tree, but which are not prevented by Hartigan consistency. We have shown that both are symptoms of clusters becoming connected at the improper level, and argued that the extent to which a cluster is δ-minimal controls the amount in which it is over-segmented or improperly nested. With more and more samples, we'd like the extent to which a clustering exhibits over-segmentation and improper nesting to shrink to zero. We therefore introduce a notion of consistency which requires any cluster to be δ-minimal with $\delta \to 0$ as $n \to \infty$.

In the following, suppose a sample $X_n \subset \mathcal{X}$ of size n is used to construct a cluster tree $\hat{\mathcal{C}}_{f,n}$ that is an estimate of $\mathcal{C}_{f,n}$, and let $\hat{\mathsf{C}}_{f,n}$ be $\hat{\mathcal{C}}_{f,n}$ equipped with f as height function. Furthermore, it is assumed that each definition holds with probability approaching one as $n \to \infty$.

Definition 8 (Minimality) *We say that $\hat{\mathsf{C}}_{f,n}$ ensures* minimality *if given any connected component A of the superlevel set $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ for some $\lambda > 0$, $A \cap X_n$ is connected at level $\lambda - \delta$ in $\hat{\mathsf{C}}_{f,n}$ for any $\delta > 0$ as $n \to \infty$.*

Minimality concerns the level at which a cluster is connected – it says nothing about the ability of an algorithm to distinguish pairs of disjoint clusters. For this, we must complement minimality with an additional notion of consistency which ensures separation. Hartigan consistency is sufficient, but does not explicitly address the level at which two clusters are separated. We will therefore introduce a slightly different notion, which we term *separation*:

Definition 9 (Separation) *We say that $\hat{\mathsf{C}}_{f,n}$ ensures* separation *if when A and B are two disjoint connected components of $\{f \geq \lambda\}$ merging at $\mu = m_{\mathsf{C}_f}(A \cup B)$, $A \cap X_n$ and $B \cap X_n$ are separated at level $\mu + \delta$ in $\hat{\mathsf{C}}_{f,n}$ for any $\delta > 0$ as $n \to \infty$.*

It is interesting to note that Hartigan consistency contains some weak notion of connectedness, as it requires the two sets $A \cap X_n$ and $B \cap X_n$ to be connected into clusters A_n and B_n at the same level at which they are separated. Our notion only requires that $A \cap X_n$ and $B \cap X_n$ be disjoint at this level. We "factor out" Hartigan consistency's idea of connectedness, leaving separation, and replace it with a stronger notion of minimality.

Taken together, minimality and separation imply Hartigan consistency.

Theorem 10 (Minimality and separation $\Longrightarrow$ Hartigan consistency) *If a hierarchical clustering method ensures both separation and minimality, then it is Hartigan consistent.*

Proof Let A and A' be disjoint connected components of the superlevel set $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ merging at level μ. Pick any $\lambda - \mu > \delta > 0$. Definitions 8 and 9 imply that there exists an N such that for all $n \geq N$, $A \cap X_n$ and $A' \cap X_n$ are separated and individually connected at level $\mu + \delta$. Assume $n \geq N$. Let A_n be the smallest cluster containing all of $A \cap X_n$, and A'_n be the smallest cluster containing all of $A' \cap X_n$. Suppose for a contradiction that there is some $x \in X_n$ such that $x \in A_n \cap A'_n$. Then either $A_n \subset A'_n$ or $A'_n \subset A_n$. In either case, there is some cluster C such that $h(C) \geq \mu + \delta$, $A_n \subset C$, and $A'_n \subset C$. Since $A \cap X_n \subset A_n$ and $A' \cap X_n \subset A'_n$, this contradicts the assumption that $A \cap X_n$ and $A' \cap X_n$ are separated at level $\mu + \delta$. Hence $A_n \cap A'_n = \emptyset$. ■

Minimality and separation have been defined as properties which are true for all clusters in the limit. In addition, we may define stronger versions of these concepts which require that all clusters approach minimality and separation uniformly:

Definition 11 (Uniform minimality and separation) *$\hat{\mathsf{C}}_{f,n}$ ensures* uniform minimality *if given any $\delta > 0$ there exists an N depending only on δ such that for all $n \geq N$ and all λ, any cluster $A \in \{x \in \mathcal{X} : f(x) \geq \lambda\}$ is connected at level $\lambda - \delta$. $\hat{\mathsf{C}}_{f,n}$ is said to ensure* uniform separation *if given any $\delta > 0$ there exists an N depending only on δ such that for all $n \geq N$ and all μ, any two disjoint connected components merging in $\{x \in \mathcal{X} : f(x) \geq \mu\}$ are separated at level $\mu + \delta$.*

The uniform versions of minimality and separation are equivalent to the weaker versions under some assumptions on the density. The proof of the following theorem is given in Appendix A.1.

Theorem 12 *If the density f is bounded from above and is such that $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ contains finitely many connected components for any λ, then any algorithm which ensures minimality also* *ensures uniform minimality on f, and any algorithm which ensures separation also ensures uniform separation.*

In the next section, we will introduce a distance between hierarchical clusterings, and show that convergence in this metric implies these consistency properties.

4. Merge distortion metric

The previous section introduced the notions of minimality and separation, which are desirable properties for a hierarchical clustering algorithm estimating the density cluster tree. Like Hartigan consistency, minimality and separation are limit properties, and do not directly quantify the disparity between the true density cluster tree and an estimate. We now introduce a *merge distortion metric* on cluster trees (equipped with height functions) which will allow us to do just that.

We make our definitions specifically so that convergence in the merge distortion metric implies the desirable properties of minimality and separation. Specifically, consider once again the density depicted in Figure 1. Suppose we run a cluster tree algorithm on a finite sample X_n drawn from f, obtaining a hierarchical clustering $\hat{\mathcal{C}}_{f,n}$. Let $\hat{\mathsf{C}}_{f,n}$ be this clustering equipped with f as a height function. We may then talk about the height at which two points merge in $\hat{\mathsf{C}}_{f,n}$, and of the level at which clusters are connected and separated in $\hat{\mathsf{C}}_{f,n}$. These are the concepts required to discuss minimality and separation.

Suppose that the algorithm ensures minimality and separation in the limit. What can we say about the merge height of a and a' in $\hat{\mathsf{C}}_{f,n}$ as $n \to \infty$? First, minimality will suggest that $M \cap X_n$ be connected in $\hat{\mathsf{C}}_{f,n}$ at level $\mu - \delta$, with $\delta \to 0$. This implies that the merge height of a and a' is bounded below by $\mu - \delta$, with $\delta \to 0$. On the other hand, separation implies that $A \cap X_n$ and $A' \cap X_n$ be separated at level $\mu + \delta$, with $\delta \to 0$. Therefore the merge height of a and a' is bounded above by $\mu + \delta$, with $\delta \to 0$. Hence in the limit $n \to \infty$, the merge height of a and a' in $\hat{\mathsf{C}}_{f,n}$, written $m_{\hat{\mathsf{C}}_{f,n}}(a, a')$, must converge to μ, which is otherwise known as $m_{\mathsf{C}_f}(a, a')$: the merge height of a and a' in the true density cluster tree.

With this as motivation, we'll work backwards, defining our distance between clusterings in such a way that convergence in the metric implies that the merge height between any two points in the estimated tree converges to the merge height in the true density cluster tree. We'll then show that this entails minimality and separation, as desired.

Merge distortion metric. Let $\mathsf{C}_1 = (\mathcal{X}_1, \mathcal{C}_1, h_1)$ and $\mathsf{C}_2 = (\mathcal{X}_2, \mathcal{C}_2, h_2)$ be two cluster trees equipped with height functions. Recall from Definition 6 that each cluster tree is associated with its own merge height function which summarizes the level at which pairs of points merge. We define the distance between C_1 and C_2 in terms of the distortion between merge heights. In general, C_1 and C_2 cluster different sets of objects, so we will use the distortion with respect to a *correspondence*[1] between these sets.

Definition 13 (Merge distortion metric) *Let $\mathsf{C}_1 = (X_1, \mathcal{C}_1, h_1)$ and $\mathsf{C}_2 = (X_2, \mathcal{C}_2, h_2)$ be two hierarchical clusterings equipped with height functions. Let $S_1 \subset X_1$ and $S_2 \subset X_2$. Let $\gamma \subset S_1 \times S_2$ be a correspondence between S_1 and S_2. The merge distortion distance between C_1 and C_2 with respect to γ is defined as*

$$d_\gamma(\mathsf{C}_1, \mathsf{C}_2) = \max_{(x_1,x_2),(x_1',x_2') \in \gamma} |m_{\mathsf{C}_1}(x_1, x_1') - m_{\mathsf{C}_2}(x_2, x_2')|.$$

1. Recall that a correspondence γ between sets S and S' is a subset of $S \times S'$ such that for $\forall s \in S, \exists s' \in S'$ such that $(s, s') \in \gamma$, and $\forall s' \in S', \exists s \in S$ such that $(s, s') \in \gamma$.

The above definition is related to the standard notion of the distortion of a correspondence between two metric spaces (Burago et al., 2001). We note that if $X_1 = X_2$ and γ is a correspondence between X_1 and X_2, then $d_\gamma(\mathsf{C}_1, \mathsf{C}_2) = 0$ implies that $\mathsf{C}_1 = \mathsf{C}_2$ in the sense that the two trees $\mathcal{C}_1$ and $\mathcal{C}_2$ are isomorphic and the height function for corresponding nodes are identical.

Now consider the special case of the distance between the true density cluster tree $\mathsf{C}_f = (\mathcal{X}, \mathcal{C}_f, f)$ and a finite estimate. Suppose we run a hierarchical clustering algorithm on a sample $X_n \subset \mathcal{X}$ of size n drawn from f, obtaining a cluster tree $\hat{\mathcal{C}}_{f,n}$. Denote by $\hat{\mathsf{C}}_{f,n} = (X_n, \hat{\mathcal{C}}_{f,n}, f)$ the cluster tree equipped with height function f. Then the natural correspondence is induced by identity in X_n: That is, $\hat{\gamma}_n = \{(x,x) : x \in X_n\}$. We then define our notion of convergence to the density cluster tree with respect to this correspondence:

Definition 14 (Convergence to the density cluster tree) *We say that a sequence of cluster trees $\{\hat{\mathsf{C}}_{f,n}\}$ converges to the high density cluster tree C_f of f, written $\hat{\mathsf{C}}_{f,n} \to \mathsf{C}_f$, if for any $\varepsilon > 0$ there exists an N such that for all $n \geq N$, $d_{\hat{\gamma}_n}(\hat{\mathsf{C}}_{f,n}, \mathsf{C}_f) < \varepsilon$.*

5. Properties of the merge distortion metric

We now prove various useful properties of our merge distortion metric. First, we show that convergence in the distance implies both uniform minimality and uniform separation. We then show that the converse is also true. We conclude by discussing stability properties of the distance.

Theorem 15 $\hat{\mathsf{C}}_{f,n} \to \mathsf{C}_f$ *implies 1) uniform minimality and 2) uniform separation.*

Proof Our proof consists of two parts.

Part I: $\hat{\mathsf{C}}_{f,n} \to \mathsf{C}_f$ implies uniform minimality. Pick any $\delta > 0$ and let n be large enough that $d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) < \delta$. Let A be a connected component of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ for arbitrary λ. Let $a, a' \in A \cap X_n$. Then $m_{\hat{\mathsf{C}}_{f,n}}(a,a') > m_{\mathsf{C}_f}(a,a') - \delta$. But a and a' are elements of A, such that $m_{\mathsf{C}_f}(a,a') \geq \lambda$. Hence $m_{\hat{\mathsf{C}}_{f,n}}(a,a') > \lambda - \delta$. Since a and a' were arbitrary, it follows that $A \cap X_n$ is connected at level $\lambda - \delta$.

Part II: $\hat{\mathsf{C}}_{f,n} \to \mathsf{C}_f$ implies uniform separation. Pick any $\delta > 0$ and let n be large enough that $d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) < \delta$. Let A and A' be disjoint connected components of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ for arbitrary λ. Let $\mu := m_{\mathsf{C}_f}(A \cup A')$ be the merge height of A and A' in the density cluster tree. Take any $a \in A \cap X_n$ and $a' \in A' \cap X_n$. Then $m_{\hat{\mathsf{C}}_{f,n}}(a,a') < m_{\mathsf{C}_f}(a,a') + \delta = \mu + \delta$. Therefore a and a' are separated at level $\mu + \delta$. Since a and a' were arbitrary, it follows that $A \cap X_n$ and $A' \cap X_n$ are separated at level $\mu + \delta$. ■

The converse is also true. In other words, convergence in our metric is equivalent to the combination of uniform minimality and uniform separation.

Theorem 16 *If $\hat{\mathsf{C}}_{f,n}$ ensures uniform separation and uniform minimality, then $\hat{\mathsf{C}}_{f,n} \to \mathsf{C}_f$.*

Proof Take any $\delta > 0$. Uniform separation and minimality imply that there exists an N such that for all λ any cluster $A \in \{x \in \mathcal{X} : f(x) \geq \lambda\}$ is connected at level $\lambda - \delta$, and for all μ any two disjoint clusters B, B' merging at μ are separated at level $\mu + \delta$. Assume $n \geq N$, and consider any $x, x' \in X_n$. W.L.O.G., assume $f(x') \geq f(x)$. We will show that $|m_{\hat{\mathsf{C}}_{f,n}}(x,x') - m_{\mathsf{C}_f}(x,x')| \leq \delta$.

Let A be the connected component of $\{f \geq f(x)\}$ containing x, and let A' be the connected component of $\{f \geq f(x')\}$ containing x'. There are two cases: either $A' \subseteq A$, or $A \cap A' = \emptyset$.

Case I: $A' \subseteq A$. Then $m_{\mathsf{C}_f}(x,x') = f(x)$. Minimality implies that $A \cap X_n$ is connected at level $f(x) - \delta$, and therefore $m_{\hat{\mathsf{C}}_{f,n}}(x,x') \geq f(x) - \delta$. On the other hand, clearly $m_{\hat{\mathsf{C}}_{f,n}}(x,x') \leq f(x)$. Hence $|m_{\hat{\mathsf{C}}_{f,n}}(x,x') - m_{\mathsf{C}_f}(x,x')| \leq \delta$.

Case II: $A \cap A' = \emptyset$. Let $\mu := m_{\mathsf{C}_f}(x, x')$ be the merge height of x and x' in the density cluster tree of f, and suppose that M is the connected component of $\{f \geq \mu\}$ containing x and x'. Then separation implies that x and x' are separated at level $\mu + \delta$, such that $m_{\hat{\mathsf{C}}_{f,n}}(x, x') < \mu + \delta$. On the other hand, minimality implies that $M \cap X_n$ is connected at level $\mu - \delta$, so that $m_{\hat{\mathsf{C}}_{f,n}}(x, x') \geq \mu - \delta$. Therefore $|m_{\hat{\mathsf{C}}_{f,n}}(x, x') - m_{\mathsf{C}_f}(x, x')| \leq \delta$. ∎

Stability. An important property to study for a distance measure is its stability; namely, to quantify how much cluster tree varies as input is perturbed. We provide two such results.

The first result says that the density cluster tree induced by a density function is stable under our merge-distortion metric with respect to L_∞-perturbation of the density function. The second result states that given a fixed hierarchical clustering, the cluster tree is stable w.r.t. small changes of the height function it is equipped with. The proofs of these results are in Appendix A.

Theorem 17 (L_∞-stability of true cluster tree) *Given a density function $f : \mathcal{X} \to \mathbb{R}$ supported on $\mathcal{X} \subset \mathbb{R}^d$, and a perturbation $\tilde{f} : \mathcal{X} \to \mathbb{R}$ of f, let $\mathcal{C}_f$ and $\mathcal{C}_{\tilde{f}}$ be the resulting density cluster tree as defined in Definition 1, and let $\mathsf{C}_f := (\mathcal{X}, \mathcal{C}_f, f)$ and $\mathsf{C}_{\tilde{f}} := (\mathcal{X}, \mathcal{C}_{\tilde{f}}, \tilde{f})$ denote the cluster tree equipped with height functions. We have $d_\gamma(\mathsf{C}_f, \mathsf{C}_{\tilde{f}}) \leq \|f - \tilde{f}\|_\infty$, where $\gamma \subset \mathcal{X} \times \mathcal{X}$ is the natural correspondence induced by identity $\gamma = \{(x, x) \mid x \in \mathcal{X}\}$.*

Theorem 18 (L_∞-stability w.r.t. f) *Given a cluster tree $(X, \mathcal{C})$, let $\mathsf{C}_1 = (X, \mathcal{C}, f_1)$ and $(C)_2 = (X, \mathcal{C}, f_2)$ be the hierarchical clusterings equipped with two height function f_1 and f_2, respectively. Let $\gamma : X \times X$ be the natural correspondence induced by identity on X; that is, $\gamma = \{(x, x) \mid x \in X\}$. We then have $d_\gamma(\mathsf{C}_1, \mathsf{C}_2) \leq 2\|f_1 - f_2\|_\infty$.*

Theorem 18 in particular leads to the following: Given a density $f : \mathcal{X} \to \mathbb{R}$ supported on $\mathcal{X} \subset \mathbb{R}^d$, suppose we have a hierarchical clustering $\hat{\mathcal{C}}_n$ constructed from a sample $X_n \subset \mathcal{X}$. However, we do not know the true density function f. Instead, suppose we have a density estimator producing an empirical density function $\tilde{f}_n : X_n \to \mathbb{R}$. Set $\hat{\mathsf{C}}_{f,n} = (X_n, \hat{\mathcal{C}}_n, f)$ as before, and $\tilde{\mathsf{C}}_{\tilde{f},n} = (X_n, \hat{\mathcal{C}}_n, \tilde{f}_n)$. Theorem 18 implies that $d(\hat{\mathsf{C}}_{f,n}, \tilde{\mathsf{C}}_{\tilde{f},n}) \leq \|f - \tilde{f}_n\|_\infty$. By the triangle inequality, this further bounds

$$d(\mathcal{C}_f, \tilde{\mathsf{C}}_{\tilde{f},n}) \leq d(\mathcal{C}_f, \hat{\mathsf{C}}_{f,n}) + \|f - \tilde{f}_n\|_\infty. \tag{1}$$

Assuming that the density estimator is consistent, we note that the cluster tree $\tilde{\mathsf{C}}_{\tilde{f},n}$ also converges to $\mathcal{C}_f$ if $\hat{\mathsf{C}}_{f,n}$ converges to $\mathcal{C}_f$. This has an important implication from a practical point of view. Imagine that we are given a sequence of more and more samples $X_{n_1}, X_{n_2}, \ldots$, and we construct a sequence of hierarchical clusterings $\hat{\mathcal{C}}_{n_1}, \hat{\mathcal{C}}_{n_2}, \ldots$. In practice, in order to test whether the current hierarchical clustering converges or not, one may wish to compare two consecutive clusterings $\hat{\mathcal{C}}_{n_i}$ and $\hat{\mathcal{C}}_{n_{i+1}}$ and measure their distance. However, since the true density is not available, one cannot compute the cluster tree distance $d_{\gamma_{n_i}}(\hat{\mathsf{C}}_{f,n_i}, \hat{\mathsf{C}}_{f,n_{i+1}})$, where the correspondence is induced by the natural inclusion from $X_{n_i} \subseteq X_{n_{i+1}}$, that is, $\gamma_{n_i} = \{(x, x) \mid x \in X_{n_i}\}$. Eqn. (1) justifies the use of a consistent empirical density estimator and computing $d_{\gamma_{n_i}}(\tilde{\mathsf{C}}_{\tilde{f},n_i}, \tilde{\mathsf{C}}_{\tilde{f},n_{i+1}})$ instead.

6. Convergence of robust single linkage

We now analyze the robust single linkage algorithm of Chaudhuri and Dasgupta (2010) in the context of our formalism. Chaudhuri and Dasgupta (2010) and Chaudhuri et al. (2014) previously studied the sense in which robust single linkage ensures that clusters are separated and connected at the appropriate levels of the empirical tree. Our analysis translates their results to our definitions of minimality and separation, thereby reinterpreting the convergence of robust single linkage in terms of our merge distortion metric.

A simple description of the algorithm is given in Appendix B. Essentially, the method produces a sequence of graphs G_r as r ranges from 0 to ∞. The sequence has a nesting property: if $r \le r'$, then $V_r \subset V_{r'}$ and $E_r \subset E'_{r'}$. We interpret this sequence of graphs as a cluster tree by taking each connected component in any graph G_r as a cluster. We equip this cluster tree with the true density f as a height function, and refer to it as $\hat{\mathsf{C}}_{f,n}$ in conformity with the preceding sections of this paper.

In what follows, assume that f is: 1) c-Lipschitz; 2) compactly supported (and hence bounded from above); and 3) such that $\{f \ge \lambda\}$ has finitely-many connected components for any λ. We will prove that the algorithm ensures minimality and separation. This, together with the assumptions on f and Theorem 16, will imply convergence in the merge distortion distance.

Suppose we run the robust single linkage algorithm on a sample of size n. Denote by v_d the volume of the d-dimensional unit hypersphere, and let $B(x,r)$ the closed ball of radius r around x in $\mathbb{R}^d$. We will write $f(B(x,r))$ to denote the probability of $B(x,r)$ under f. Define $r(\lambda)$ to be the value of r such that $v_d r^d \lambda = \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}$. Here, k is a parameter of the algorithm which we will constrain, and C_δ is the constant appearing in the Lemma IV.1 of Chaudhuri et al. (2014). First, we must show that in the limit, $G_{r(\lambda)}$ contains no points of density less than $\lambda - \epsilon$, for arbitrary ϵ.

Lemma 19 *Fix $\epsilon > 0$ and $\lambda \ge 0$. Then if $\alpha \ge \sqrt{2}$ and $k \ge (8C_\delta\lambda/\epsilon)^2\, d\log n$, there exists an N such that for all $n \ge N$, if $x \in G_{r(\lambda)}$, then $f(x) > \lambda - \epsilon$.*

Proof Define $\tilde{r} = r(\lambda - \epsilon/2)$. There exists an N such that for any $n \ge N$, $\tilde{r}\mathsf{c} \le \epsilon/4$. Consider any point $x \in G_{\tilde{r}}$. By virtue of x's membership in the graph, X_n contains k points within $B(x,\tilde{r})$. Lemma IV.1 in (Chaudhuri et al., 2014) implies that $f(B(x,\tilde{r})) > \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}$. From our smoothness assumption, we have $v_d\tilde{r}^d(f(x) + \tilde{r}\mathsf{c}) \ge f(B(x,\tilde{r})) > \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}$. Multiplying both sides by $\lambda - \epsilon/2$ and substituting gives: $v_d\tilde{r}^d(\lambda-\epsilon/2)(f(x)+\tilde{r}\mathsf{c}) = \left(\frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}\right)(f(x)+\tilde{r}\mathsf{c}) > (\lambda-\epsilon/2)(\frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n})$ so that

$$\begin{aligned} f(x) > (\lambda - \epsilon/2)\left\{\frac{k - C_\delta\sqrt{kd\log n}}{k + C_\delta\sqrt{kd\log n}}\right\} - \tilde{r}\mathsf{c} &\ge \left(1 - 2\frac{C_\delta\sqrt{d\log n}}{\sqrt{k}}\right)(\lambda - \epsilon/2) - \epsilon/4 \\ &\ge \left(1 - \frac{\epsilon}{4\lambda}\right)(\lambda - \epsilon/2) - \epsilon/4 \ge \lambda - \epsilon \end{aligned}$$

Hence for any point $x \in G_{\tilde{r}}$, $f(x) > \lambda - \epsilon$. Note that $\tilde{r} > r(\lambda)$, implying that any point in $G_{r(\lambda)}$ is also in $G_{\tilde{r}}$. Therefore if $x \in G_{r(\lambda)}$, $f(x) > \lambda - \epsilon$. ■

We now make our claim. We will use the following fact without proof: For any $A \in \{f \ge \lambda\}$ and $\delta > 0$, there exists an N such that for all $n \ge N$, if $A \cap X_n \ne \emptyset$, there is at least one point $x \in A \cap X_n$ with $f(x) < \lambda + \delta$. This follows immediately from the continuity of f and the inequalities in the Lemma IV.1 of Chaudhuri et al. (2014).

Theorem 20 *Robust single linkage converges in probability to the density cluster tree C_f in the merge distortion distance.*

Proof It is sufficient to prove minimality and separation, as then Theorem 16 will imply convergence. Fix any $\varepsilon > 0$, and let A be a connected component of $\{f \ge \lambda\}$. Define $\sigma = \varepsilon/(2\mathsf{c})$, and let A_σ be the set A thickened by closed balls of radius σ. Define $\lambda' := \inf_{x\in A_\sigma} f(x) \ge \lambda - \varepsilon/2$. Theorem IV.7 in (Chaudhuri et al., 2014) implies that there exists an N_1 such that for all $n \ge N_1$, $A \cap X_n$ is connected in $G_{r(\lambda')}$. Take $\epsilon = \varepsilon/2$ in our Lemma 19; there exists an N_2 above which each point x in $G_{r(\lambda')}$ has density $f(x) > \lambda' - \epsilon \ge (\lambda - \varepsilon/2) - \varepsilon/2 = \lambda - \varepsilon$. Then for all $n \ge \max\{N_1, N_2\}$, $A \cap X_n$ is connected in $G_{r(\lambda')}$ at level no less than $\lambda - \varepsilon$. This proves minimality.

Again fix $\varepsilon > 0$ and let A and A' be connected components of $\{f \geq \lambda\}$ merging at some height $\mu = m_{\mathsf{C}_f}(A \cup A')$. Let $\tilde{A}$ and $\tilde{A}'$ be the connected components of $\{f \geq \mu + \varepsilon/2\}$ containing A and A', respectively. Define $\sigma = \varepsilon/(4\mathsf{c})$, and let $\tilde{A}_\sigma$ (resp. $\tilde{A}'_\sigma$) be the set $\tilde{A}$ (resp. $\tilde{A}'$) thickened by closed balls of radius σ. Define $\mu' := \inf_{x \in \tilde{A}_\sigma \cup \tilde{A}'_\sigma} f(x) \geq \mu + \varepsilon/4$. Then Lemma IV.3 in (Chaudhuri et al., 2014) implies[2] that there exists some N_1 such that for all $n \geq N_1$, $\tilde{A} \cap X_n$ and $\tilde{A}' \cap X_n$, are disconnected in $G_r(\mu')$ and individually connected. Let N_2 be large enough that there exists a point $x_1 \in \tilde{A} \cap X_n$ with $f(x_1) < \mu + \varepsilon$. Then for all $n \geq \max\{N_1, N_2\}$, $A \cap X_n$ and $A' \cap X_n$ are separated at level $\mu + \varepsilon$. This proves separation. ■

7. Split-tree based hierarchical clustering

We also consider a different approach to estimate the cluster tree, using ideas from the field of computational topology. The method is based on the clustering algorithm proposed by Chazal et al. (2013), while that work focuses on analyzing flat clustering. We briefly describe our method and state our main result. A detailed description and the proof are relegated to Appendix C.

Our algorithm takes as input a set of points P_n sampled iid from a density f supported on an unknown Riemannian manifold, an empirically-estimated density function $\tilde{f}_n$, and a parameter $r > 0$, and outputs a hierarchical clustering tree T_n^{r} on P_n. Let K_n be the proximity graph on P_n, in which every point in P_n is connected to every other point that is within distance r. We then track the connected components of the subgraph of K_n spanned by all points $P_n^\lambda = \{p \in P_n : \tilde{f}_n(p) \geq \lambda\}$ as we sweep λ from high to low. The set of clusters (connected components in the subgraphs) produced this way and their natural nesting relations give rise to a hierarchical clustering that we refer to as the *split-cluster* tree T_n^{r}.

Comparing this with the definition of high-density cluster tree in Definition 2, we note that intuitively, the split-cluster tree T_n^{r} is a discrete approximation of the high-density cluster tree $\mathcal{C}_f$ for the true density function $f : \mathcal{M} \to \mathbb{R}$ where (i) the density f is approximated by the empirical density $\tilde{f}_n$; and (ii) the connectivity of the domain $\mathcal{M}$ is approximated by the proximity graph K_n.

It turns out that the constructed tree T_n^{r} is related to the so-called *split tree* studied in the computational geometry and topology literature as a variant of the *contour tree*; see e.g, (Carr et al., 2003; Wang et al., 2014). Due to this relation, the split-cluster tree can be constructed efficiently in $O(n\alpha(n))$ time using a union-find data structure once nodes in P_n are sorted (Carr et al., 2003).

Our main result is that under mild conditions, the split-cluster tree converges to the true high-density cluster tree C_f of $f : \mathcal{M} \to \mathbb{R}$ in merge distortion distance. See Appendix C for full details.

Theorem 21 *Let $\mathcal{M}$ be a compact m-dimensional Riemannian manifold embedded in $\mathbb{R}^d$ with bounded absolute curvature and positive strong convexity radius. Let $f : \mathcal{M} \to \mathbb{R}$ be a c-Lipschitz probability density function supported on $\mathcal{M}$. Let P_n be a set of n points sampled i.i.d. according to f. Assume that we are given a density estimator such that $\|f - \tilde{f}_n\|_\infty$ converges to 0 as $n \to \infty$. For any fixed $\varepsilon > 0$, we have, with probability 1 as $n \to \infty$, that $d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) \leq (4\mathsf{c} + 1)\varepsilon$, where the parameter r in computing the split-cluster tree T_n^{r} is set to be 2ε, and $\hat{\mathsf{C}}_{f,n} = (P_n, T_n^{\mathsf{r}}, f)$ is the hierarchical clustering tree T_n^{r} equipped with the height function f.*

Acknowledgements. The authors thank anonymous reviewers for insightful comments. This work is in part supported by the National Science Foundation (NSF) under grants CCF-1319406, RI-1117707, and CCF-1422830.

2. More precisely, Lemma IV.3 requires A and A' to be so-called (σ, ϵ)-separated, for some σ and ϵ. It follows from the Lipschitz-continuity of f that there is some ϵ so that A and A' are (σ, ϵ)-*separated* for this choice of σ.

References

Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Cluster Trees on Manifolds. In *Advances in Neural Information Processing Systems*, pages 2679–2687, 2013.

Mickaël Buchet, Frédéric Chazal, Tamal K. Dey, Fengtao Fan, Steve Y. Oudot, and Yusu Wang. Topological analysis of scalar fields with outliers. *CoRR*, arXiv:1412.1680, 2014.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society, 2001.

Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *The Journal of Machine Learning Research*, 11:1425–1470, 2010.

Hamish Carr, Jack Snoeyink, and Ulrike Axen. Computing contour trees in all dimensions. *Comput. Geom*, 24(2):75–94, 2003.

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, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6):41, 2013.

T. K. Dey, J. Sun, and Y. Wang. Approximating cycles in a shortest basis of the first homology group from point data. *Inverse Problems*, 27(12):124004, 2011.

J. A. Hartigan. Consistency of Single Linkage for High-Density Clusters. *Journal of the American Statistical Association*, 76(374):388–394, June 1981. ISSN 0162-1459. doi: 10.1080/01621459.1981.10477658.

John A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975. ISBN 047135645X.

Anil K Jain and Richard C Dubes. *Algorithms for clustering data*, volume 6. Prentice hall Englewood Cliffs, 1988.

Jon Kleinberg. An impossibility theorem for clustering. *Advances in neural information processing systems*, pages 463–470, 2003.

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, New York, NY, USA, 2011. ACM.

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):397–418, January 2010. ISSN 1061-8600. doi: 10.1198/jcgs.2009.07049.

Suyi Wang, Yusu Wang, and Rephael Wenger. The js-graphs of join and split trees. In *Proc. 29th Annu. ACM Sympos. Comput. Geom.*, page 539, 2014.

David Wishart. Mode analysis: A generalization of nearest neighbor which reduces chaining effects. *Numerical taxonomy*, 76(282-311):17, 1969.

Appendix A. Proofs

A.1. Proof of Theorem 12

Theorem 22 *Let f be a density supported on $\mathcal{X}$, and let $\{\hat{\mathsf{C}}_{f,n}\}$ be a sequence of cluster trees computed from finite samples $X_n \subset \mathcal{X}$. Suppose $f \leq M$ for some $M \in \mathbb{R}$, and that for any λ, $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ contains finitely many connected components. Then*

1. *If $\{\hat{\mathsf{C}}_{f,n}\}$ ensures minimality for f, it ensures uniform minimality.*
2. *If $\{\hat{\mathsf{C}}_{f,n}\}$ ensures separation for f, it ensures uniform separation.*

Proof We will prove the first case, in which $\hat{\mathsf{C}}_{f,n}$ ensures minimality. The proof of uniform separation follows closely, and is therefore omitted.

Pick $\delta > 0$. Let $\mathsf{C}_f(\lambda)$ denote the (finite) set of connected components of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$. Consider the collection of connected components of superlevel sets spaced $\delta/2$ apart:

$$\mathcal{D} = \bigcup_{n=0}^{\lfloor 2M/\delta \rfloor} \mathsf{C}_f(n\delta/2)$$

The fact that $\hat{\mathsf{C}}_{f,n}$ ensures minimality implies that for each $C \in \mathcal{D}$ there exists an $N(C)$ such that for all $n \geq N(C)$, $C \cap X_n$ is connected at level $h(C) - \delta/2$. Let $N = \max_{C \in \mathcal{D}} N(C)$. This is well-defined, as $\mathcal{D}$ is a finite set.

Let A be a connected component of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ for an arbitrary λ. Let $\lambda' = \lfloor 2\lambda/\delta \rfloor \frac{\delta}{2}$, i.e., λ' is the largest multiple of $\delta/2$ such that $\lambda' \leq \lambda$. Then A is a subset of some connected component A' of $\{x \in \mathcal{X} : f(x) \geq \lambda'\}$. Note that $A' \in \mathcal{D}$, so that $A' \cap X_n$ is connected at level $\lambda' - \delta/2$. Therefore $A \cap X_n$ is connected at level $\lambda' - \delta/2 > (\lambda - \delta/2) - \delta/2 = \lambda - \delta$. Since A was arbitrary, and the choice of N depended only upon δ, it follows that $\hat{\mathsf{C}}_{f,n}$ ensures uniform minimality. ■

A.2. Proof of Theorem 17

Proof Set $\delta = \|f - \tilde{f}\|_\infty$. Let x, x' be two arbitrary points from X. We need to show that $|d_{\mathsf{C}_f}(x, x') - d_{\mathsf{C}_{\tilde{f}}}(x, x')| \leq 4\delta$, which will then implies the theorem. In what follows, we prove that $d_{\mathsf{C}_f}(x, x') \leq d_{\mathsf{C}_{\tilde{f}}}(x, x') + 4\delta$.

Let $m = m_{\mathsf{C}_f}(x, x')$ denote the merge height of x and x' w.r.t. C_f. This means that there exists a connected component $C \in \{y \in \mathcal{X} \mid f(y) \geq m\}$ such that $x, x' \in C$. Since $\|f - \tilde{f}\|_\infty = \delta$, we have that for any point $y \in C$, $|\tilde{f}(y) - f(y)| \leq \delta$ and thus $\tilde{f}(y) \geq m - \delta$. Hence all points in C must belong to the same connected component, call it $\tilde{C}(\supseteq C) \in \{y \in \mathcal{X} \mid \tilde{f}(y) \geq m - \delta\}$ with respect to the clustering $\mathcal{C}_{\tilde{f}}$. It then follows that the merge height $m_{\mathsf{C}_{\tilde{f}}}(x, x') \geq m - \delta$. Combining this with that $\|f - \tilde{f}\|_\infty = \delta$, we have:

$$\begin{aligned} d_{\mathsf{C}_{\tilde{f}}}(x, x') &= \tilde{f}(x) + \tilde{f}(x') - 2m_{\mathsf{C}_{\tilde{f}}}(x, x') \\ &\leq f(x) + \delta + f(x') + \delta - 2m + 2\delta = d_{\mathsf{C}_f}(x, x') + 4\delta. \end{aligned}$$

The proof for $d_{\mathsf{C}_f}(x, x') \leq d_{\mathsf{C}_{\tilde{f}}}(x, x') + \delta$ is symmetric. The theorem then follows. ■

A.3. Proof of Theorem 18

Proof Set $\delta := \|f_1 - f_2\|_\infty$. Let x, x' be two arbitrary points from X. We need to show that $|d_{\mathsf{C}_1}(x, x') - d_{\mathsf{C}_2}(x, x')| \leq 4\delta$, which will then implies the theorem. In what follows, we prove that $d_{\mathsf{C}_2}(x, x') \leq d_{\mathsf{C}_1}(x, x') + 4\delta$.

Let $m_1 = m_{\mathsf{C}_1}(x, x')$ denote the merge height of x and x' w.r.t. C_1. This means that there exists a cluster $C \in \mathcal{C}$ such that $x, x' \in C$ and $f_1(C) = m_1$. Since $f_i(C) = \min_{y \in C} f_i(y)$, for $i = 1, 2$, we thus have that $f_2(C) \in [m_1 - \delta, m_1 + \delta]$. It then follows that $m_{\mathsf{C}_2}(x, x') \geq f_2(C) \geq m_1 - \delta$. Combining with that $\|f_1 - f_2\|_\infty = \delta$, we have:

$$\begin{aligned} d_{\mathsf{C}_2}(x, x') &= f_2(x) + f_2(x') - 2m_{\mathsf{C}_2}(x, x') \\ &\leq f_1(x) + \delta + f_1(x') + \delta - 2m_1 + 2\delta = d_{\mathsf{C}_1}(x, x') + 4\delta. \end{aligned}$$

The proof for $d_{\mathsf{C}_1}(x, x') \leq d_{\mathsf{C}_2}(x, x') + \delta$ is symmetric. The theorem then follows. ■

Appendix B. Robust single linkage

We briefly describe the robust single linkage algorithm, and refer readers to the work of Chaudhuri and Dasgupta (2010) and Chaudhuri et al. (2014) for details. In what follows, let $B(x, r)$ denote the closed ball of radius r around x.

The algorithm operates as follows: Given a sample X_n of n points drawn from a density f supported on $\mathcal{X}$, and parameters α and k, perform the following steps:

1. For each $x_i \in X_n$, set $r_k(x_i) = \min\{r : B(x_i, r) \text{ contains } k \text{ points}\}$.
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.

The algorithm produces a series of graphs as r ranges from 0 to ∞. Each connected component in G_r for any r is considered a cluster. The clusters exhibit hierarchical structure, and can be interpreted as a cluster tree. We may therefore discuss the sense in which this discrete tree converges to the ideal density cluster tree.

Appendix C. Split-tree based hierarchical clustering

In this section we inspect a different approach, based on the one proposed and studied by Chazal et al. (2013), to obtain a hierarchical clustering for points sampled from a density function supported on a Riemannian manifold, using tools from the emerging field of computational topology.

In particular, we focus on the following setting: Let $\mathcal{M} \subseteq \mathbb{R}^d$ be a smooth m-dimensional Riemannian manifold $\mathcal{M}$ embedded in the ambient space $\mathbb{R}^d$, and $f : \mathcal{M} \to \mathbb{R}$ a c-Lipschitz probability density function supported on $\mathcal{M}$. Let P_n denote a set of n points sampled i.i.d. according to f. We further assume that we have a density estimator $\tilde{f}_n : P_n \to \mathbb{R}$ which estimates the true density f with the guarantee that $\|f - \tilde{f}_n\|_\infty \leq \mathcal{E}(n)$ for an error function $\mathcal{E}(n)$ which tends to zero as $n \to +\infty$.

C.1. Split-cluster tree construction.

We now describe an algorithm which takes as input P_n and the empirical density function $\tilde{f}_n : P_n \to \mathbb{R}$, and outputs a hierarchical clustering tree T_n^{r} on P_n. The algorithm uses a parameter $\mathsf{r} > 0$, which intuitively should go to zero as n tends to infinity.

Let $K_n = (P_n, E)$ denote the 1-dimensional simplicial complex, where $E := \{(p, p') \mid \|p - p'\| \leq \mathsf{r}\}$. In other words, K_n is the proximity graph on P_n where every point in P_n is connected to all other points from P_n within r distance to it. We now define the following hierarchical clustering (cluster tree) T_n^{r}:

Given any value λ, let $P_n^\lambda := \{p \in P_n \mid \tilde{f}_n(p) \geq \lambda\}$ be the set of vertices with estimated density at least λ, and let K_n^λ be the subgraph of K_n induced by P_n^λ. The subgraph K_n^λ may have multiple connected components, and the vertex set of each connected component gives rise to a cluster. The collection of such clusters for all $\lambda \in \mathbb{R}$ is T_n^{r}, which we call the *split-cluster tree* of P_n w.r.t. $\tilde{f}_n$. We put the parameter r in T_n^{r} to emphasize the dependency of this cluster tree on r.

In particular, note that the function $\tilde{f}_n : P_n \to \mathbb{R}$ induces a piecewise-linear (PL) function on the underlying space $|K_n|$ of K_n, which we denote as $\tilde{f} : |K_n| \to \mathbb{R}$. It turns out that the tree representation of this cluster tree T_n^{r} is exactly the so-called *split tree* of this PL function $\tilde{f}$ as studied in the literature of computational geometry and topology, as a variant of the contour tree; see e.g, (Carr et al., 2003; Wang et al., 2014). This is why we refer to T_n^{r} as split-cluster tree of P_n. The split-cluster tree T_n^{r} can be easily computed in $O(n\alpha(n))$ time using the union-find data structure, once the vertices in P_n are already sorted (Carr et al., 2003).

We note that Chazal et al. (2013) proposed a clustering algorithm based on this idea, and provided various nice theoretical studies of *flat clusterings* resulted from such a construction. We instead focus on the hierarchical clustering tree constructed using this split tree idea.

Finally, recall that $f : \mathcal{M} \to \mathbb{R}$ is the true density function. Given T_n^{r}, let $\hat{\mathsf{C}}_{f,n} = (P_n, T_n^{\mathsf{r}}, f)$ be the corresponding cluster tree equipped with height function $f : P_n \to \mathbb{R}$ (which is the restriction of f to P_n). As before, we still use $\mathcal{C}_f$ to denote the high-density cluster tree w.r.t. the true density function f, and $\mathsf{C}_f = (\mathcal{M}, \mathcal{C}_f, f)$ be the corresponding cluster tree equipped with height function $f : \mathcal{M} \to \mathbb{R}$.

In what follows, we will study the convergence of the distance $d_\gamma(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n})$, where $\gamma : M \times P_n$ is the natural correspondence induced by identity in P_n, that is, $\gamma = \{(p,p) \mid p \in P_n$ (thus $p \in \mathcal{M})\}$. For simplicity of presentation, we will omit the reference of this natural correspondence γ in the remainder of this section.

C.2. Convergence of split-cluster tree

First, we introduce some notation. Let $d(x,y)$ denote the Euclidean distance between any two points $x, y \in \mathbb{R}^d$, while $d_M(x,y)$ denotes the geodesic distance between points $x, y \in \mathcal{M}$ on the manifold $\mathcal{M}$. Given a smooth manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$, the *medial axis* of $\mathcal{M}$, denoted by $\mathcal{A}_M$, is the set of points in $\mathbb{R}^d$ which has more than one nearest neighbor in $\mathcal{M}$. The *reach of* $\mathcal{M}$, denoted by $\rho(\mathcal{M})$, is the infimum of the closest distance from any point in $\mathcal{M}$ to the medial axis, that is, $\rho(\mathcal{M}) = \inf_{x\in\mathcal{M}} d(x, \mathcal{A}_M)$.

Following the notations of Chazal et al. (2013), we further define:

Definition 23 ((Geodesic) ε-sample) *Given a subset $Y \subseteq \mathcal{M}$ and a parameter $\varepsilon > 0$, a set of points $Q \subset Y$ is a* (geodesic) ε-sample of Y *if every point of Y is within ε geodesic distance to some point in Q; that is, $\forall x \in Y, \min_{q\in Q} d_{\mathcal{M}}(x,q) \leq \varepsilon$.*

In what follows, let $\mathcal{M}^\lambda = \{x \in \mathcal{M} \mid f(x) \geq \lambda\}$ be the super-level set of $f : \mathcal{M} \to \mathbb{R}$ w.r.t. λ.

Lemma 24 *We are given an m-dimensional smooth manifold $\mathcal{M} \subset \mathbb{R}^d$ with a c-Lipschitz density function $f : \mathcal{M} \to \mathbb{R}$ on $\mathcal{M}$. Let $\rho(\mathcal{M})$ be the reach of $\mathcal{M}$. Let P_n be an ε-sample of $\mathcal{M}^\lambda$. Assume that $\|f - \tilde{f}_n\|_\infty \leq \eta$ for $\tilde{f}_n : P_n \to \mathbb{R}$, and that the parameter we use to construct T_n^{r} satisfies $\mathsf{r} \geq 2\varepsilon$ and $\mathsf{r} < \rho(\mathcal{M})/2$. Then $d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) \leq \max\{\mathsf{c}\mathsf{r} + 2\eta, \lambda\}$.*

Proof Consider any two points $p, p' \in P_n$. Let m and $\hat{m}$ denote the merge height of p and p' in C_f and in $\hat{\mathsf{C}}_{f,n}$, respectively. By definition, we have that

$$d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) = \max_{p,p' \in P_n} |m_{\mathsf{C}_f}(p,p') - m_{\hat{\mathsf{C}}_{f,n}}(p,p')|. \tag{2}$$

We now distinguish two cases:

Case 1: $m \geq \lambda$.

In this case, by definition of the merge height m of p and p', we know that: (1) $f(p), f(p') \geq m$, and thus both p and p' are from $P_n \cap \mathcal{M}^\lambda$, and (2) p and p' are connected in $\mathcal{M}^m$, thus there is a path $\Gamma \subset \mathcal{M}^\lambda$ connecting p and p' such that for any $x \in \Gamma$, $f(x) \geq m$. We now show that the merge height of p and p' in $\hat{\mathsf{C}}_{f,n}$ satisfies $\hat{m} \in [m - \mathsf{c}\mathsf{r} - 2\eta, m + \mathsf{c}\mathsf{r} + 2\eta]$.

Indeed, let $\pi : \mathcal{M}^\lambda \to P_n$ be the projection map that sends any $x \in \mathcal{M}^\lambda$ to its nearest neighbor in P_n. Since P_n is an ε-sample for $\mathcal{M}^\lambda$, we have that $d_M(x, \pi(x)) \leq \varepsilon$ and thus $|f(x) - f(\pi(x))| \leq \mathsf{c}\varepsilon$ (as f is c-Lipschitz), for any $x \in \mathcal{M}^\lambda$. Consider any two sufficiently close points $x, x' \in \Gamma$ (i.e, $\|x - x'\| < \mathsf{r} - 2\varepsilon$), we have that (1) either $\pi(x) = \pi(x')$, (2) or $\pi(x) \neq \pi(x')$ but

$$\|\pi(x) - \pi(x')\| \leq \|\pi(x) - x\| + \|x - x'\| + \|x' - \pi(x')\| < \mathsf{r}.$$

In other words, $\pi(\Gamma)$ consists of a sequence of vertices $p = q_1, q_2, \ldots, q_s = p'$ in K_n such that there is an edge in K_n connecting any two consecutive q_i, q_{i+1}, $i \in [1, s-1]$. The concatenation of these edges forms a path $\Gamma' = \langle p = q_1, q_2, \ldots, q_s = p' \rangle$ in K_n connecting p to p'. Since for each $i \in [1, s-1]$, $q_i = \pi(x)$ for some $x \in \Gamma$, we have that

$$f(q_i) = f(\pi(x)) \geq f(x) - \mathsf{c}\varepsilon \geq m - \mathsf{c}\varepsilon,$$

where the two inequalities follow from that $|f(x) - f(\pi(x))| \leq \mathsf{c}\varepsilon$ and $f(x) \geq m$ for any $x \in \Gamma$. Since $\|f - \tilde{f}_n\| \leq \eta$, it then follows that $\tilde{f}(q_i) \geq m - \mathsf{c}\varepsilon - \eta$ for any $q_i \in \Gamma'$.

Recall that K_n^α denotes the subgraph of K_n induced by the set of points $P_n^\alpha = \{p \in P_n \mid \tilde{f}_n(p) \geq \alpha\}$ whose function value w.r.t. the empirical density function $\tilde{f}_n$ is at least α. It then follows that p and p' should be connected in K_n^α for $\alpha = m - \mathsf{c}\varepsilon - \eta$. It then follows that the merge height

$$\hat{m} = m_{\hat{\mathsf{C}}_{f,n}}(p,p') \geq \min_{q \in K_n^\alpha} f(q) \geq \min_{q \in K_n^\alpha} \tilde{f}_n(q) - \eta \geq \alpha - \eta = m - \mathsf{c}\varepsilon - 2\eta.$$

We now show the other direction, namely $m \geq \hat{m} - \mathsf{c}\mathsf{r} - 2\eta$. Indeed, by definition of $\hat{m}$, there is a path $\tilde{\Gamma} = \langle \tilde{q}_1 = p, \tilde{q}_2, \ldots, \tilde{q}_t = \tilde{p}' \rangle$ in K_n connecting p and p' such that for any $\tilde{q}_i \in \tilde{\Gamma}$, $f(\tilde{q}_i) \geq \hat{m}$. Now let $\ell_\mathcal{M}(q, q')$ denote a minimizing geodesic between two points $q, q' \in \mathcal{M}$. We then have that there is a path

$$\tilde{\Gamma}' := \ell_\mathcal{M}(\tilde{q}_1, \tilde{q}_2) \circ \ell_\mathcal{M}(\tilde{q}_2, \tilde{q}_3) \circ \cdots \circ \ell_\mathcal{M}(\tilde{q}_{t-1}, \tilde{q}_t)$$

in $\mathcal{M}$ connecting $p = \tilde{q}_1$ to $p' = \tilde{q}_t$.

At the same time, note that for any two consecutive nodes $\tilde{q}_i$ and $\tilde{q}_{i+1}$ from $\tilde{\Gamma}$, we know $\|\tilde{q}_i - \tilde{q}_{i+1}\| \leq \mathsf{r}$ as $(\tilde{q}_i, \tilde{q}_{i+1})$ is an edge in K_n. For $\mathsf{r} < \rho(\mathcal{M})/2$, where $\rho(\mathcal{M})$ is the reach of the manifold

$\mathcal{M}$, by Proposition 1.2 of Dey et al. (2011), we have that the geodesic distance $d_{\mathcal{M}}(\tilde{q}_i, \tilde{q}_{i+1})$ is at most $\frac{4}{3}\|\tilde{q}_i - \tilde{q}_{i+1}\|$. Thus $d_{\mathcal{M}}(\tilde{q}_i, \tilde{q}_{i+1}) \leq \frac{4}{3}\mathsf{r}$. In particular, for any point $x \in \ell_{\mathcal{M}}(\tilde{q}_i, \tilde{q}_{i+1})$, it is within $\frac{2}{3}\mathsf{r}$ distance to either $\tilde{q}_i$ or $\tilde{q}_{i+1}$. Hence we have that

$$f(x) \geq \min\{f(\tilde{q}_i), f(\tilde{q}_{i+1})\} - \frac{2}{3}\mathsf{cr} > \hat{m} - \mathsf{cr} \quad \Rightarrow \quad m = m_{\mathsf{C}_f}(p, p') \geq \min_{x \in \tilde{\Gamma}'} f(x) \geq \hat{m} - \mathsf{cr}.$$

Putting everything together, we have that $|m - \hat{m}| \leq \mathsf{cr} + 2\eta$ for the case $m \geq \lambda$.

Case 2: $m < \lambda$.

First, note that the proof of $m \geq \hat{m} - \mathsf{cr}$ holds regardless of the value of m. Hence we have $\hat{m} - m \leq \mathsf{cr}$ for the case $m < \lambda$ as well. On the other hand, since $m < \lambda$, $m - \hat{m} < \lambda$. Thus $|\hat{m} - m| \leq \max\{\mathsf{cr}, \lambda\}$.

The lemma follows from combining these two cases with Eqn. (2). ■

Remark: The bound in the above result can be large if the value λ is large. We can obtain a stronger result for points in $P_n \cap \mathcal{M}^\lambda$ which is independent of λ. However, the above result is cleaner to present and it suffices to prove our main convergence result in Theorem 27.

To obtain a convergence result, we need to incur the following results from Chazal et al. (2013).

Definition 25 (Chazal et al. (2013)) *Let $\mathcal{M}$ be an m-dimensional Riemannian manifold with intrinsic metric $d_{\mathcal{M}}$. Given a subset $A \subseteq \mathcal{M}$ and a parameter $r > 0$, define $\mathcal{V}_r(A)$ to be the infimum of the Hausdorff measures achieved by geodesic balls of radius r centered in A; that is:*

$$\mathcal{V}_r(A) = \inf_{x \in A} \mathcal{H}^m(B_{\mathcal{M}}(x, r)), \;\; \textit{where} \;\; B_{\mathcal{M}}(x, r) := \{y \in \mathcal{M} \mid d_{\mathcal{M}}(x, y) \leq r\}. \tag{3}$$

*We also define the r-*covering number* of A, denoted by $\mathcal{N}_r(A)$ to be the minimum number of closed geodesic balls of radius r needed to cover A (the balls do not have to be centered in A).*

Theorem 26 (Theorem 7.2 of Chazal et al. (2013)) *Let $\mathcal{M}$ be an m-dimensional Riemannian manifold and $f : \mathcal{M} \to \mathbb{R}$ a c-Lipschitz probability density function. Consider a set P sampled according to f in i.i.d. fashion. Then, for any parameter $\varepsilon > 0$ and $\alpha > \mathsf{c}\varepsilon$, we are guaranteed that P forms an ε-sample of $\mathcal{M}^\alpha$ with probability at least $1 - \mathcal{N}_{\varepsilon/2}(\mathcal{M}^\alpha) e^{-n(\alpha - \mathsf{c}\varepsilon)\mathcal{V}_{\varepsilon/2}(\mathcal{M}^\alpha)}$.*

Remarks. For simplicity, we now focus on the case where $\mathcal{M}$ is a compact smooth embedded manifold with bounded absolute sectional curvature and positive strong convexity radius $\rho_c(\mathcal{M})$. It follows from the Günther-Bishop Theorem that (see e.g, Appendix B of Buchet et al. (2014)) in this case, there exists a constant μ depending only on the intrinsic property of $\mathcal{M}$ such that $\mathcal{V}_r(\mathcal{M}^\alpha) \geq \mathcal{V}_r(\mathcal{M}) \geq \mu r^m$ for sufficiently small r. Due to the compactness of $\mathcal{M}$, this further gives an upper bound on $\mathcal{N}_r(\mathcal{M})$ (and thus for $\mathcal{N}_r(\mathcal{M}^\alpha) \leq \mathcal{N}_r(\mathcal{M})$). Thus for fixed ε and α, P_n forms an ε-sample for $\mathcal{M}^\alpha$ with probability 1 as $n \to +\infty$.

We remark that Lemma 7.3 of Chazal et al. (2013) also states that $\mathcal{N}_{\varepsilon/2}(\mathcal{M}^\alpha) < +\infty$ (i.e, it is finite) and $\mathcal{V}_{\varepsilon/2}(\mathcal{M}^\alpha) > 0$ for the more general case where $\mathcal{M}$ is a complete Riemannian manifold with bounded absolute sectional curvature, for any $\varepsilon < 2\rho_c(\mathcal{M})$. Hence again for fixed ε and α, P_n forms an ε-sample for $\mathcal{M}^\alpha$ with probability 1 as $n \to +\infty$.

Putting Theorem 24 and 26 together, we obtain the following:

Theorem 27 *Let $\mathcal{M}$ be a compact m-dimensional Riemannian manifold embedded in $\mathbb{R}^d$ with positive strong convexity radius. Let $f : \mathcal{M} \to \mathbb{R}$ be a c-Lipschitz probability density function supported on $\mathcal{M}$. Let P_n be a set of n points sampled i.i.d. according to f. Assume that we are given a density estimator such that $\|f - \tilde{f}_n\|_\infty$ converges to 0 as $n \to \infty$. For any fixed $\varepsilon > 0$, we have, with probability 1 as $n \to \infty$, that $d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) \leq (4\mathsf{c} + 1)\varepsilon$, where the parameter r in computing the split-cluster tree T_n^{r} is set to be 2ε.*

Proof Set λ in Lemma 24 to be $2\mathsf{c}\varepsilon$. We then have that $d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) \leq 2\mathsf{c}\varepsilon + 2\mathsf{c}\varepsilon + 2\|f - \tilde{f}_n\|_\infty$ if P_n is an ε-sample of $\mathcal{M}^\lambda$. Since $\|f - \tilde{f}_n\|_\infty$ converges to 0 as n tends to ∞, there exists N_ε such that $\|f - \tilde{f}_n\|_\infty \leq \varepsilon$ for any $n > N_\varepsilon$. Hence $d(\mathsf{C}_f, \hat{\mathsf{C}}_{f,n}) \leq (4\mathsf{c} + 1)\varepsilon$ if P_n is an ε-sample of $\mathcal{M}^\lambda$ and for $n > N_\varepsilon$. The theorem follows from this and Theorem 26 above. ■