Consistency and Rates for Clustering with DBSCAN

Bharath K. Sriperumbudur
Gatsby Computational Neuroscience Unit
University College London
bharath@gatsby.ucl.ac.uk

Ingo Steinwart
Institute for Stochastics and Applications
University of Stuttgart, Germany
ingo.steinwart@mathematik.uni-stuttgart.de

Abstract

We propose a simple and efficient modification of the popular DBSCAN clustering algorithm. This modification is able to detect the most interesting vertical threshold level in an automated, data-driven way. We establish both consistency and optimal learning rates for this modification.

1 Introduction

The algorithm DBSCAN, Ester et al. (1996), is among the clustering methods that are most popular for practitioners, and has been successfully used in a variety of different applications. Given thresholds $\rho > 0$ and $\delta > 0$ it first identifies the samples x_i from the dataset $D = (x_1, \ldots, x_n) \in X^n$ for which the balls $B(x_i, \delta)$ around x_i with radius δ contain at least ρn samples. The clusters returned by DBSCAN are then the δ-connected components of the set of identified samples. Although heuristics do exist for choosing the free parameters ρ and δ, a rigorous approach for their choice is an open problem. The goal of this work is to present a simple modification of DBSCAN, which eliminates spurious clusters and for which the choice of ρ is data driven and the choice of δ is deterministic. For this modification we show that it finds the first interesting level ρ^*. Moreover, we show both consistency and optimal learning rates for the corresponding clusters at this level.

The above description of DBSCAN shows that DBSCAN can be viewed as a modified single density level set approach based on a moving window estimate $\hat{h}$ for the density h of the data-generating distribution. This estimate is thresholded at the level ρ, so that roughly speaking an estimate of the set $\{h \geq \rho\}$ is considered when determining the connected components in the second step. If the set $\{\hat{h} \geq \rho\}$ would be used as an estimate of $\{h \geq \rho\}$, then the approach would be a so-called single-level density-based clustering method. This class of methods goes back to ideas of Carmichael et al. (1968), Hartigan (1975) and since then, it has been studied by several, see, e.g., Hartigan (1981), Cuevas and Fraiman (1997), Rigollet (2007), Maier et al. (2009), Rinaldo and Wasserman (2010) and the references therein. Unfortunately, however, determining the set $\{\hat{h} \geq \rho\}$ and its connected components algorithmically is extremely difficult, which motivates DBSCAN's strategy of considering the set $D \cap \{\hat{h} \geq \rho\}$ instead. While this approach removes the algorithmic difficulties, it becomes significantly more difficult to statistically analyze DBSCAN's output. To the best of our knowledge, the most advanced analysis in this respect can again be found in Rinaldo and Wasserman (2010), where it is shown that, under some regularity assumptions, the algorithm returns a certain approximation of the connected components of $\{h \geq \rho\}$. While this result guarantees that for *fixed* level ρ DBSCAN returns an estimate of the desired clusters, it does not answer the more interesting question of how to choose the level ρ.

One way of addressing this question, is hierarchical clustering, where the hierarchical tree structure of the connected components for different levels ρ is estimated. We refer to Hartigan (1975), Stuetzle (2003), Chaudhuri and Dasgupta (2010), Stuetzle and Nugent (2010) for definitions and methods. In particular, Chaudhuri and Dasgupta (2010) show that in a weak sense of Hartigan (1981), a modified single linkage algorithm converges to this tree under some assumptions on the density h. Recently, Kpotufe and von Luxburg (2011) further improve the analysis of Chaudhuri and Dasgupta (2010) and establish learning rates for Hölder-continuous densities h and provide a method that prunes nearest neighbor cluster trees.

A different approach has been recently taken in Stein-

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

wart (2011). Roughly speaking, the goal of that paper was to design and analyze an algorithm that automatically finds the smallest level ρ^* for which $\{h \geq \rho^*\}$ has more than one connected component. Here, we like to stress that the value of ρ^* is assumed to be *unknown*, an assumption we will adopt in this paper. Under a so-called thickness assumption on h, the author then shows that the algorithm identifies the corresponding components asymptotically, that is, he proves consistency. Clearly, by recursively applying this algorithm, we again get a consistent estimate of the cluster tree. In this sense, the question considered in Steinwart (2011) can be viewed as the basic building block for hierarchical clustering.

Unfortunately, however, the algorithm in Steinwart (2011) is based on an histogram estimate of h, and has thus little practical value. In addition, the focus is more put on showing consistency for a large class of densities than on obtain learning rates for a restricted class of distributions. In this paper, we address these issues by carrying over the analysis of Steinwart (2011) to a modification of DBSCAN. Roughly speaking, this modification iteratively considers DBSCAN's output for some increasing sequence $\rho_1, \rho_2, \ldots$ of levels and identifies and removes spurious clusters from this output by a simple strategy that can be efficiently implemented. The algorithm stops as soon as more than one cluster is identified. Note that the OPTICS algorithm (Ankerst et al., 1999) implements such an iterative inspection of different levels for DBSCAN, and hence our modification of DBSCAN inherits the algorithmic advantages of these algorithms, while it also possesses good statistical guarantees.

At first glance, the paper may appear to be rather similar to Steinwart (2011), since it heavily builds upon the techniques developed there. However, a closer look reveals that we had to develop quite a few ideas to the new situation. Let us illustrate this by possibly the biggest difference to Steinwart (2011): In Steinwart (2011), a plug-in approach $\{\hat{h} \geq \rho\}$ based on a simple histogram density estimator $\hat{h}$ is taken. Due to the simple structure of histogram estimates the connected components of $\{\hat{h} \geq \rho\}$ can then be easily and efficiently computed. Now, if DBSCAN computed the connected components of $\{\hat{h} \geq \rho\}$ based on a Parzen-window estimator, then our analysis and results would indeed be an unsurprising extension of Steinwart (2011). Unlike for level set estimation, however, kernel plug-in estimates including Parzen-window estimates are not algorithmically suitable for clustering. Indeed, due to overlapping balls around samples, there may be parts of $\{\hat{h} \geq \rho\}$ that do not contain samples from the training set D, which in turn makes it very difficult to identify the connected components of $\{\hat{h} \geq \rho\}$. DBSCANs approach to consider the algorithmically treatable set of balls around $D \cap \{\hat{h} \geq \rho\}$ addresses this algorithmic issue. On the downside, however, this approach is generally viewed as a heuristic with very little statistical justification. In fact, the only justification in view of clustering so far seems to be by the already mentioned Rinaldo and Wasserman (2010, Theorem 17), which however, needs stronger assumptions—see their equations (28) and (29)—to obtain a significant looser result in terms of *mollified* target clusters and *fixed* level ρ. Consequently, it seems fair to say that so far the analysis of DBSCANs heuristic has been an open problem, which we solved in this paper. However, we like to stress that without the recent techniques from Steinwart (2011), this task would have rendered impossible.

The paper is organized as follows. In Section 2, we present preliminaries about the notions of density level sets, connectivity, and clusters and collect related technical results from Steinwart (2011). In Section 3, we present our clustering algorithm which is a modification of DBSCAN and analyze it in Section 4 by proving its consistency and deriving the rates of convergence. The proofs of the results that are not provided in the main text are included in the supplementary material.

2 Preliminaries: Density level sets, connectivity, and clusters

In this section, we briefly recall all notions related to the definition and analysis of clusters from Steinwart (2011). Furthermore, we collect various technical yet important results from Steinwart (2011) needed throughout the paper.

In the following, $A \triangle B$ denotes the symmetric difference between two sets A and B. Unless specified otherwise, we assume throughout the paper that X is a compact subset of $\mathbb{R}^d$ with strictly positive volume. For $x \in X$, $B(x, \delta)$ denotes the closed δ-ball around x with respect to some metric d such as the Euclidean or the supremum metric. Given an $A \subset X$, we denote the closure and interior of A by $\overline{A}$ and $\mathring{A}$, respectively. Moreover, we write $d(x, A) := \inf_{x' \in A} d(x, x')$ for the distance between some $x \in X$ and A. For $\delta > 0$, we further define the δ-tube around A by

$$T_\delta(A) := \{x \in X : d(x, A) \leq \delta\} .$$

Note that $T_\delta(A)$ is closed. In the following, $\mathcal{B}(X)$ denotes the Borel σ-algebra on X and μ denotes a scaled version of the Lebesgue measure, where the scaling ensures $\mu(B(x, \delta)) = \delta^d$ for all $x \in \mathbb{R}^d$ and all $\delta > 0$. Furthermore, P is an *unknown* μ-absolutely continuous probability measure on $\mathcal{B}(X)$. In order to

avoid most of the technical difficulties arising in Steinwart (2011) from the treatment of general densities, we restrict our considerations to the case where P has Hölder-continuous μ-density h. This restriction makes it further possible to extend the analysis of Steinwart (2011) from consistency to learning rates.

Let us now recall the key concepts required to define clusters in the sense of Steinwart (2011):

Density level sets. To find a notion of density level sets that is topologically invariant against different choices of the density h, Steinwart (2011) defined a density level set at level $\rho \geq 0$ as

$$M_\rho := \operatorname{supp} \mu_\rho$$

where $\operatorname{supp} \mu_\rho$ denotes the support of μ_ρ and μ_ρ is the measure defined by

$$\mu_\rho(A) := \mu(A \cap \{h \geq \rho\}), \quad A \in \mathcal{B}(X).$$

By definition, the sets M_ρ are closed. Moreover, note that since we restrict our considerations to P that have a (necessary unique) Hölder continuous density h, the construction above could be replaced by the usual $\{h \geq \rho\}$ without changing our results. However, this would have required to basically modify most results from Steinwart (2011) without any new conceptual insight. We therefore decided to keep the definition M_ρ, but readers less interested in topological and measure theoretical details can safely replace M_ρ by $\{h \geq \rho\}$ to get the big picture. The following list recalls some basic yet important properties of the sets M_ρ, $\rho \geq 0$, from Steinwart (2011, Section 2.1):

(a_1) **Level Sets.** $\overline{\{h > \rho\}} \subset M_\rho \subset \{h \geq \rho\}$.

(a_2) **Monotonicity.** $M_{\rho_2} \subset M_{\rho_1}$ for all $\rho_1 \leq \rho_2$.

(a_3) **Regularity.** $\mu(M_\rho \triangle \{h \geq \rho\}) = 0$.

(a_4) **Normality.** $\bar{M}_\rho = \dot{M}_\rho$, where $\bar{M}_\rho := \bigcup_{\rho' > \rho} M_{\rho'}$ and $\dot{M}_\rho := \bigcup_{\rho' > \rho} \mathring{M}_{\rho'}$.

(a_5) **Open Level Sets.** $\bar{M}_\rho = \{h > \rho\}$.

Connectivity. Recall from topology that a closed non-empty $A \subset X$ is called connected, if, for every pair $A', A'' \subset A$ of closed disjoint subsets of A with $A' \cup A'' = A$, we have $A' = \emptyset$ or $A'' = \emptyset$. Moreover, the maximal connected subsets of A are called the connected components of A. These components are closed and form a partition of A. We denote the set of topologically connected components of A by $\mathcal{C}(A)$. Now, the key idea of Steinwart (2011) is to relate the connected components of different subsets of X. Let us recall the following two main ingredients of this approach from Steinwart (2011, Section 2.2):

(b_1) Given two closed subsets $A \subset B$ of X, there exists exactly one map $\zeta : \mathcal{C}(A) \to \mathcal{C}(B)$ such that

$$A' \subset \zeta(A'), \qquad A' \in \mathcal{C}(A).$$

We call ζ the topologically connected components relating map (top-CCRM) between A and B. Sometimes, we write $\zeta_{A,B} := \zeta$ to emphasize the involved pair (A, B).

(b_2) Given three closed subsets $A \subset B \subset C$ of X, the top-CCRMs of these sets satisfy

$$\zeta_{A,C} = \zeta_{B,C} \circ \zeta_{A,B}.$$

Clusters. With these preparations we can now recall the definition of clusters from Steinwart (2011). Note that we have cleaned this definition from some technical assumptions that are not necessary since we are only dealing with continuous densities.

Definition 2.1. *Let P be a μ-absolutely continuous probability measure on X with continuous μ-density h. Then we say that P can be topologically clustered between the critical levels $\rho^* \geq 0$ and $\rho^{**} > \rho^*$, if, for all $\rho \in [0, \rho^{**}]$, the following conditions hold:*

(c_1) *The set M_ρ has either one or two topologically connected components.*

(c_2) *If $|\mathcal{C}(M_\rho)| = 1$, then $\rho \leq \rho^*$.*

(c_3) *If $|\mathcal{C}(M_\rho)| = 2$, then $\rho \geq \rho^*$ and the top-CCRM $\zeta : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$ is bijective.*

Note that the definition above does not exclude $|\mathcal{C}(M_{\rho^*})| = 1$, and hence the connected components of M_{ρ^*} cannot be used to define clusters. However, for $\rho > \rho^*$, each $A \in \mathcal{C}(M_\rho)$ should be a subset of a cluster of P. This idea is used in the following definition, which defines the clusters of P by a limit for $\rho \searrow \rho^*$.

Definition 2.2. *Let P be a μ-absolutely continuous probability measure on X that can be topologically clustered between the critical levels ρ^* and ρ^{**}. For $\rho \in (\rho^*, \rho^{**}]$, we write $\zeta_\rho : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$ for the top-CCRM. Moreover, let A_1 and A_2 be the topologically connected components of $M_{\rho^{**}}$. Then the sets*

$$A_i^* := \bigcup_{\rho \in (\rho^*, \rho^{**}]} \zeta_\rho(A_i), \qquad i \in \{1, 2\},$$

are called the topological clusters of P.

3 The Clustering Algorithm

In this section, we present our clustering algorithm (Algorithm 1) that approximates the optimal level ρ^* and estimates the corresponding clusters. Since it is

based on a kernel density estimator, we begin by recalling the latter. To this end, let $\delta > 0$ and P be a probability measure with density h. Then the infinite-sample kernel density estimator for h is defined as

$$\bar{h}_{P,\delta}(x) := \delta^{-d} \int_{\mathbb{R}^d} K\left(\frac{x-y}{\delta}\right) dP(y), \quad x \in \mathbb{R}^d, \quad (1)$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a bounded measurable function with $\int_{\mathbb{R}^d} K(x)\, d\mu(x) = 1$. Let us now assume that we have a data set $D = (x_1, \ldots, x_n) \in X^n$. The empirical kernel density estimator of h is then defined by

$$\bar{h}_{D,\delta}(x) := \frac{1}{n\delta^d} \sum_{i=1}^{n} K\left(\frac{x-x_i}{\delta}\right), \quad x \in \mathbb{R}^d. \quad (2)$$

Throughout the paper, we solely consider kernels of the form $K := \mathbf{1}_{B(0,1)}$, which leads to

$$K\left(\frac{x-y}{\delta}\right) = \mathbf{1}_{B(x,\delta)}(y), \quad x, y \in \mathbb{R}^d. \quad (3)$$

In addition, we always assume that the collection $\mathcal{B} := \{B(x,1) : x \in \mathbb{R}^d\}$ of unit balls has a finite VC-dimension. Examples of such norms include, but are not limited to, the Euclidean norm (Devroye and Lugosi, 2001, Corollary 4.2) and uniform norm (Devroye and Lugosi, 2001, Lemma 4.1). The following result shows that in this case the empirical kernel density estimator, $\bar{h}_{D,\delta}$ *uniformly* approximates the infinite-sample kernel density estimator $\bar{h}_{P,\delta}$.

Theorem 3.1. *Let P be a probability measure on X. Then, under the above assumptions, there exists a constant C, such that, for all $n \geq 1$, $\delta > 0$ and $\tau > 0$, we have*

$$P^n\left(\left\{D \in X^n : \|\bar{h}_{D,\delta} - \bar{h}_{P,\delta}\|_\infty < \frac{C}{n\delta^d}\log\frac{C}{\delta} + \sqrt{\frac{C}{n\delta^d}\log\frac{C}{\delta} + \frac{\tau C}{n\delta^d} + \frac{C\sqrt{\tau}}{\sqrt{n\delta^d}}}\right\}\right) \geq 1 - e^{-\tau}.$$

It is easy to check that, for $n \geq 1$, $\delta \in (0,1)$, and $\tau \geq 1$ satisfying $n\delta^d \geq \tau$ and $n\delta^d \geq |\log\delta|$, the estimate of Theorem 3.1 can be simplified to

$$P^n\left(\left\{D : \|\bar{h}_{D,\delta} - \bar{h}_{P,\delta}\|_\infty < C'\sqrt{\frac{\tau|\log\delta|}{n\delta^d}}\right\}\right) \geq 1 - e^{-\tau},$$

where C' is a new constant that is independent of n, δ, and τ.

Our Algorithm will rely on $\bar{h}_{D,\delta}$ in the sense that it estimates the density level set at level ρ by

$$\mathcal{M}_{\rho,\delta} := T_\delta(\{x \in D : \hat{f}_\rho(x) = 1\}),$$

where

$$\hat{f}_\rho := \operatorname{sign}(\bar{h}_{D,\delta} - \rho).$$

The following key result, which will make it possible to adapt the techniques from Steinwart (2011) to our setting, provides an upper and a lower bound on $\mathcal{M}_{\rho,\delta}$ in terms of some true density level sets $M_{\rho+\varepsilon+\eta}$ and $M_{\rho-\varepsilon-\eta}$.

Lemma 3.2. *Let $X \subset \mathbb{R}^d$ be compact and P be a μ-absolutely continuous probability measure on X with a continuous density h. For $\eta > 0$, we define*

$$\delta_\eta := \sup\big\{\delta > 0 : \forall x, x' \in X \text{ with } d(x,x') \leq 2\delta \text{ we have } |h(x) - h(x')| < \eta\big\}. \quad (4)$$

Moreover, let $\varepsilon > 0$, $\delta \in (0, \delta_\eta)$, and $D \in X^n$ be a data set with $\|\bar{h}_{D,\delta} - \bar{h}_{P,\delta}\|_\infty < \varepsilon$. Then, for all $\rho \geq 0$, we have

$$M_{\rho+\varepsilon+\eta} \subset \mathcal{M}_{\rho,\delta} \subset M_{\rho-\varepsilon-\eta}.$$

Note that since X is assumed to be compact, the density h in Lemma 3.2 is actually uniformly continuous, and hence we find a $\delta_\eta > 0$ for all $\eta > 0$. Moreover, the smoother h is, the larger we can pick δ_η. For example, if h is Hölder-continuous with exponent $\alpha \in (0,1]$ and constant $C_h > 0$, that is

$$|h(x) - h(x')| \leq C_h\, d^\alpha(x,x') \quad (5)$$

for all $x, x' \in X$, then an easy calculation shows that $\delta_\eta \geq 0.5\,(\eta/C_h)^{1/\alpha}$.

Proof. Let us begin by showing the first inclusion. To this end, we fix an $x \in M_{\rho+\varepsilon+\eta} \subset \{h \geq \rho + \varepsilon + \eta\}$, where the inclusion follows from (a_1). For all $x' \in B(x, 2\delta)$, we then have $h(x') > \rho + \varepsilon$, i.e., $B(x,2\delta) \subset \{h > \rho+\varepsilon\} \subset M_{\rho+\varepsilon}$, where the second inclusion follows again from (a_1). Let us now suppose that there exists a sample $x_i \in D$ such that $\hat{f}_\rho(x_i) = -1$ and $d(x,x_i) \leq \delta$. Then the definition of $\hat{f}_\rho$ yields $\hat{h}(x_i) < \rho$ and therefore we find

$$\delta^{-d} \int_{B(x_i,\delta)} h\, d\mu = \bar{h}_{P,\delta}(x_i) < \rho + \varepsilon\,.$$

On the other hand, $d(x,x_i) \leq \delta$ together with the already shown $B(x,2\delta) \subset M_{\rho+\varepsilon}$ implies $B(x_i,\delta) \subset M_{\rho+\varepsilon} \subset \{h \geq \rho + \varepsilon\}$ by a simple application of the triangle inequality and (a_1). Therefore, we obtain

$$\delta^{-d} \int_{B(x_i,\delta)} h\, d\mu \geq \rho + \varepsilon\,,$$

leading to a contradiction. In other words, for all samples $x_i \in D$, we have $\hat{f}_\rho(x_i) = 1$ or $d(x,x_i) > \delta$. Let us assume that we had $d(x,x_i) > \delta$ for all $x_i \in D$. Then we clearly find

$$h_{D,\delta}(x) = \frac{1}{n\delta^d} \sum_{i=1}^{n} \mathbf{1}_{B(x,\delta)}(x_i) = 0\,.$$

On the other hand, we have already seen that $B(x,\delta) \subset B(x,2\delta) \subset M_{\rho+\varepsilon} \subset \{h \geq \rho+\varepsilon\}$. This yields

$$\bar{h}_{P,\delta}(x) = \delta^{-d} \int_{B(x,\delta)} h \, d\mu \geq \rho + \varepsilon$$

and hence we obtain $\bar{h}_{D,\delta}(x) > \rho$. In other words, we have found a contradiction, and therefore there does exist a sample $x_i \in D$ with $d(x,x_i) \leq \delta$. Our previous consideration then shows that this sample must satisfy $\hat{f}_\rho(x_i) = 1$, and hence we finally obtain $x \in \mathcal{M}_{\rho,\delta}$.

To show the second inclusion, let us now fix an $x \in \mathcal{M}_{\rho,\delta}$. This means that there exists an $x_i \in D$ such that $\hat{h}(x_i) \geq \rho$ and $d(x,x_i) \leq \delta$. This implies $\bar{h}_{P,\delta}(x_i) > \rho - \varepsilon$, i.e.,

$$\delta^{-d} \int_{B(x_i,\delta)} h \, d\mu > \rho - \varepsilon \,.$$

This means, there exists $x' \in B(x_i,\delta)$ such that $h(x') > \rho - \varepsilon$. The triangle inequality yields $d(x,x') \leq 2\delta$, and hence we obtain $h(x) > \rho - \varepsilon - \eta$ by the definition of δ_η and $\delta < \delta_\eta$. Using the inclusion $\{h > \rho - \varepsilon - \eta\} \subset M_{\rho-\varepsilon-\eta}$ from (a_1), we then find $x \in M_{\rho-\varepsilon-\eta}$. □

Based on the above result, the following theorem relates the topological connected components of our estimate, $\mathcal{M}_{\rho,\delta}$ to the topological connected components of $M_{\rho+\varepsilon+\eta}$.

Theorem 3.3. *Let $X \subset \mathbb{R}^d$ be compact and P be a μ-absolutely continuous probability measure on X with a continuous density h that can be topologically clustered between the critical levels ρ^* and ρ^{**}. For $\eta > 0$, we define δ_η by (4). Moreover, let $\varepsilon > 0$, $\delta \in (0,\delta_\eta)$, and $D \in X^n$ be a data set with $\|\bar{h}_{D,\delta} - \bar{h}_{P,\delta}\|_\infty < \varepsilon$. Then, for all $\rho \in (0, \rho^{**} - 3\varepsilon - 3\eta]$, the following disjoint union holds*

$$\mathcal{C}(\mathcal{M}_{\rho,\delta}) = \big\{B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta}) : B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} = \emptyset\big\} \cup \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))\,,$$

where $\zeta : \mathcal{C}(M_{\rho+\varepsilon+\eta}) \to \mathcal{C}(\mathcal{M}_{\rho,\delta})$ is the top-*CCRM.*

Proof. Our first goal is to establish the following *disjoint* union:

$$\begin{aligned}\mathcal{C}(\mathcal{M}_{\rho,\delta}) = \big\{&B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta}) \setminus \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta})) \\ &: B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} \neq \emptyset\big\} \\ &\cup \big\{B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta}) : B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} = \emptyset\big\} \\ &\cup \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))\,. \qquad (6)\end{aligned}$$

We begin by showing the auxiliary result

$$V' \cap M_{\rho+3\varepsilon+3\eta} \neq \emptyset\,, \qquad V' \in \mathcal{C}(M_{\rho+\varepsilon+\eta}). \quad (7)$$

To this end, we observe that (c_3) yields $|\mathcal{C}(M_{\rho^{**}})| = 2$, which implies $M_{\rho^{**}} \neq \emptyset$. Let W' and W'' be the two topologically connected components of $M_{\rho^{**}}$. Let us first assume that $M_{\rho+\varepsilon+\eta}$ has exactly one connected component V', i.e. $V' = M_{\rho+\varepsilon+\eta}$. Then $\rho \leq \rho^{**} - 3\varepsilon - 3\eta$ implies

$$\begin{aligned}\emptyset \neq M_{\rho^{**}} \subset M_{\rho+3\varepsilon+3\eta} &= M_{\rho+\varepsilon+\eta} \cap M_{\rho+3\varepsilon+3\eta} \\ &= V' \cap M_{\rho+3\varepsilon+3\eta}\,,\end{aligned}$$

i.e. we have shown (7). Let us now assume that $M_{\rho+\varepsilon+\eta}$ has more than one topological connected component. Then it has exactly two such components V' and V'' by (c_1). By (c_3), we may then assume without loss of generality that we have $W' \subset V'$ and $W'' \subset V''$. Since $\rho \leq \rho^{**} - 3\varepsilon - 3\eta$ implies $M_{\rho^{**}} \subset M_{\rho+3\varepsilon+3\eta}$, these inclusions yield $\emptyset \neq W' = W' \cap M_{\rho^{**}} \subset V' \cap M_{\rho+3\varepsilon+3\eta}$ and $\emptyset \neq W'' = W'' \cap M_{\rho^{**}} \subset V'' \cap M_{\rho+3\varepsilon+3\eta}$. Consequently, we have proved (7) in this case, too.

Using the top-CCRM property $V' \subset \zeta(V')$ and the inclusion $M_{\rho+3\varepsilon+3\eta} \subset \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta}$ shown in Lemma 3.2, we now conclude that $B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} \neq \emptyset$ for all $B' \in \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))$. This yields

$$\begin{aligned}\big\{B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta}) \setminus \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta})) : B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} = \emptyset\big\} \\ = \big\{B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta}) : B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} = \emptyset\big\}\,,\end{aligned}$$

and the latter immediately implies (6).

It remains to show that $\big\{B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta}) \setminus \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta})) : B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} \neq \emptyset\big\} = \emptyset$. Let us assume the converse, that is, there exists some $B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta})$ with $B' \notin \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))$ and $B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} \neq \emptyset$. Since $\mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} \subset M_{\rho+\varepsilon+\eta}$ by Lemma 3.2, there then exists an $x \in B' \cap M_{\rho+\varepsilon+\eta}$. Let $B'' \in \mathcal{C}(M_{\rho+\varepsilon+\eta})$ be the unique component such that $x \in B''$. Then we have $x \in \zeta(B'')$ by the top-CCRM property, i.e. x is contained in a topologically connected component of $\mathcal{M}_{\rho,\delta}$ that belongs to the image of ζ. However, we assumed that $x \in B' \in \mathcal{C}(\mathcal{M}_{\rho,\delta}) \setminus \zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))$, and hence we have found a contradiction. □

The above result shows that eventually all connected components B' of our estimate $\mathcal{M}_{\rho,\delta}$ of M_ρ are either contained in $\zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))$ or satisfy $B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} = \emptyset$. The latter components are easy to identify and remove, and hence we can exactly identify the connected components that are contained in $\zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))$. Based on this result, we present Algorithm 1 that scans through the values of ρ, removes the connected components satisfying $B' \cap \mathcal{M}_{\rho+2\varepsilon+2\eta,\delta} = \emptyset$, and stops as soon as the remaining set of components $\zeta(\mathcal{C}(M_{\rho+\varepsilon+\eta}))$ contains more than one component.

The first step of our analysis shows that, for sufficiently small $\eta > 0$, $\delta > 0$, and $\varepsilon > 0$, Algorithm 1

identifies exactly the connected components B' that are contained in $\zeta(\mathcal{C}_\tau(M_{\rho+\varepsilon+\eta}))$. This is formalized by the following theorem, which shows that the level returned by Algorithm 1, $\rho^*(D)$, is close to ρ^*.

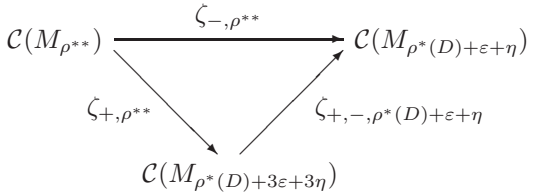
Theorem 3.4. *Let $X \subset \mathbb{R}^d$ be compact and P be a μ-absolutely continuous probability measure on X with a continuous density h that can be topologically clustered between the critical levels ρ^* and ρ^{**}. For $\eta > 0$, we define δ_η by (4), and we further fix an $\varepsilon^* > 0$ and an $\eta^* > 0$ that satisfy $\varepsilon^* < (\rho^{**} - \rho^*)/16$ and $\eta^* < (\rho^{**} - \rho^*)/16$. Then, for all $\varepsilon \in (0, \varepsilon^*]$, $\eta \in (0, \eta^*]$, $\delta \in (0, \delta_\eta)$, $n \geq 1$, and all data sets $D \in X^n$ satisfying $\|\bar{h}_{D,\delta} - \bar{h}_{P,\delta}\|_\infty < \varepsilon$, the following statements are true:*

i) $\rho^*(D) \in [\rho^* - \varepsilon - \eta, \rho^* + \varepsilon^* + \eta^* + 2\varepsilon + 2\eta]$.

ii) $|\mathcal{C}(M_{\rho^*(D)+3\varepsilon+3\eta})| = 2$ *and the top-CCRM* $\zeta : \mathcal{C}(M_{\rho^*(D)+3\varepsilon+3\eta}) \to \mathcal{C}(\mathcal{M}_{\rho^*(D)+2\varepsilon+2\eta,\delta})$ *is injective.*

iii) Algorithm 1 returns the two connected components of $\zeta(C(M_{\rho^*(D)+3\varepsilon+3\eta}))$.

iv) All top-CCRMs in the following commutative diagram are bijective:



4 Consistency and Rates

In this section, we prove a consistency result and, under additional assumptions, some optimal rates of convergence for Algorithm 1. Before proving the consistency result, we motivate it by first analyzing Theorem 3.4 assuming that we are in the situation of this theorem. To this end, let A_1 and A_2 be the topologically connected components of $M_{\rho^{**}}$, V_1' and V_2' be the topologically connected components of $M_{\rho^*(D)+3\varepsilon+3\eta}$, and $B_1(D)$ and $B_2(D)$ be the components returned by Algorithm 1. By Theorem 3.4, we may then assume without loss of generality that $A_i \subset V_i' \subset B_i(D)$ for $i = 1, 2$. This yields $A_i \subset B_i(D) \cap A_i^*$. Consequently, the returned components $B_i(D)$ contain a chunk of the desired clusters A_i^*, $i = 1, 2$. Now, the goal of our consistency result is to show that $B_i(D) \triangle A_i^*$ actually becomes arbitrarily small for arbitrarily large n. To this end, we assume in the following without loss of generality that Algorithm 1 always returns two components, denoted by $B_1(D)$ and $B_2(D)$. With these preparation we can now formulate the consistency result for Algorithm 1.

Algorithm 1 Estimate clusters of a uniformly continuous density using a kernel density estimator

Require: Some $\delta > 0$, $\eta > 0$, and $\varepsilon > 0$ with $\delta < \delta_\eta$. A dataset $D \in X^n$.

Ensure: An estimate of the topological clusters A_1^* and A_2^*.

1: Compute the kernel density estimator $\bar{h}_{D,\delta}$.
2: $\rho \leftarrow -\varepsilon - \eta$
3: **repeat**
4: $\quad \rho \leftarrow \rho + \varepsilon + \eta$
5: $\quad$ Compute $\hat{f}_\rho(x) = \operatorname{sign}(\bar{h}_{D,\delta}(x) - \rho)$ for all $x \in D$.
6: $\quad$ Identify the connected components $B_1', \ldots, B_M'$ of $T_\delta(\{x_i \in D : \hat{f}_\rho(x_i) = 1\})$ satisfying

$$B_j' \cap T_\delta(\{x_i \in D : \hat{f}_{\rho+2\varepsilon+2\eta}(x_i) = 1\}) \neq \emptyset.$$

7: **until** $M \neq 1$
8: Compute $\hat{f}_{\rho+2\varepsilon+2\eta}(x_i) = \operatorname{sign}(\bar{h}_{D,\delta}(x) - \rho - 2\varepsilon - 2\eta)$ for all $x_i \in D$.
9: Identify the connected components $B_1', \ldots, B_M'$ of $T_\delta(\{x_i \in D : \hat{f}_{\rho+2\varepsilon+2\eta}(x_i) = 1\})$ satisfying

$$B_j' \cap T_\delta(\{x_i \in D : \hat{f}_{\rho+4\varepsilon+4\eta}(x_i) = 1\}) \neq \emptyset.$$

10: **return** ρ and $B_1', \ldots, B_M'$.

Theorem 4.1. *Let $X \subset \mathbb{R}^d$ be compact and P be a μ-absolutely continuous probability measure on X with a continuous density h that can be topologically clustered between the critical levels ρ^* and ρ^{**}. For $\eta > 0$, we define δ_η by (4), and we further fix strictly positive sequences (ε_n), (η_n), and (δ_n) converging to zero such that $\frac{n\delta_n^d \varepsilon_n^2}{|\log \delta_n|} \to \infty$ and $\delta_n < \delta_{\eta_n}$ for all $n \geq 1$. For $n \geq 1$, consider Algorithm 1 using the parameters ε_n, η_n, and δ_n. Then, for all $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P^n(\Delta_\epsilon) = 1\,,$$

where $\Delta_\epsilon := \big\{D \in X^n : \mu(B_1(D) \triangle A_1^*) + \mu(B_2(D) \triangle A_2^*) \leq \epsilon\big\}$.

Proof. Let us write $A_{\rho^{**},i}$, $i = 1, 2$, for the two topologically connected components of $M_{\rho^{**}}$. Moreover, for $\rho \in (\rho^*, \rho^{**}]$, we define $A_{\rho,i} := \zeta_\rho(A_{\rho^{**},i})$, where $\zeta_\rho : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$ is the top-CCRM. In addition, we write $A_{\rho,i} := \emptyset$ for $\rho > \rho^{**}$ and $A_{\rho,i} := X$ for $\rho \leq \rho^*$. Our first goal is to show that

$$\mu(\bar{A}_{\rho^*,i} \setminus \dot{A}_{\rho^*,i}) = 0 \tag{8}$$

for $i = 1, 2$, where $\bar{A}_{\rho^*,i}$ and $\dot{A}_{\rho^*,i}$ are defined in (a_4). First note that since P has a continuous μ-density h, (a_4) shows that $\bar{M}_{\rho^*} = \dot{M}_{\rho^*}$. Note that

$$\bar{M}_{\rho^*} = \bigcup_{\rho > \rho^*} M_\rho = \bigcup_{\rho > \rho^*} (A_{\rho,1} \cup A_{\rho,2})$$

$$= \bigcup_{\rho>\rho^*} A_{\rho,1} \cup \bigcup_{\rho>\rho^*} A_{\rho,2} = \bar{A}_{\rho^*,1} \cup \bar{A}_{\rho^*,2}\,.$$

Similarly, it can be shown that $\dot{M}_{\rho^*} = \dot{A}_{\rho^*,1} \cup \dot{A}_{\rho^*,2}$, and thus $\bar{A}_{\rho^*,1} \cup \bar{A}_{\rho^*,2} = \dot{A}_{\rho^*,1} \cup \dot{A}_{\rho^*,2}$. Since $\bar{A}_{\rho^*,1}$ and $\bar{A}_{\rho^*,2}$ are disjoint and $\dot{A}_{\rho^*,i} \subset \bar{A}_{\rho^*,i}$ for $i = 1, 2$, we have $\bar{A}_{\rho^*,i} = \dot{A}_{\rho^*,i}$ for $i = 1, 2$ and therefore (8) follows.

Note that since for any $\rho \geq \rho' > \rho^*$, we have $A_{\rho,i} \subset A_{\rho',i}$ for $i = 1, 2$, it can be easily shown that

$$\begin{aligned}\bar{A}_{\rho^*,i} &= \bigcup_{\rho>\rho^*} \bigcap_{\varepsilon>0} \bigcap_{\eta>0} A_{\rho-\varepsilon-\eta,i} \\ &= \bigcup_{\rho>\rho^*} \bigcup_{\varepsilon>0} \bigcup_{\eta>0} A_{\rho+\varepsilon+\eta,i}. \end{aligned} \tag{9}$$

Let us now fix an $\epsilon > 0$. By (8) and (9) there then exist $\eta_\epsilon > 0$, $\varepsilon_\epsilon > 0$, and $\rho_\epsilon > \rho^*$ such that, for all $\varepsilon \in (0, \varepsilon_\epsilon]$, $\eta \in (0, \eta_\epsilon]$, $\rho \in (\rho^*, \rho_\epsilon]$, and $i = 1, 2$ we have

$$\mu(\bar{A}_{\rho^*,i} \setminus A_{\rho+\varepsilon+\eta,i}) \leq \epsilon\,. \tag{10}$$

Moreover, (9) shows that, for all $\rho > \rho^*$, we have

$$\bigcap_{\varepsilon>0} \bigcap_{\eta>0} M_{\rho-\varepsilon-\eta} \subset \bar{M}_{\rho^*}\,.$$

Clearly, this implies $\bigcap_{\varepsilon>0} \bigcap_{\eta>0} M_{\rho-\varepsilon-\eta} \setminus \bar{M}_{\rho^*} = \emptyset$. Consequently, we have

$$\mu(M_{\rho-\varepsilon-\eta} \setminus \bar{M}_{\rho^*}) \leq \epsilon \tag{11}$$

for all $\rho > \rho^*$ and all sufficiently small $\varepsilon > 0$, $\eta > 0$. Without loss of generality, we may thus assume that (11) holds for all $\varepsilon \in (0, \varepsilon_\epsilon]$, $\eta \in (0, \eta_\epsilon]$ and all $\rho > \rho^*$. We now define $\varepsilon^* := \min\{\frac{\rho_\epsilon - \rho^*}{16}, \frac{\rho^{**} - \rho^*}{16}\}$, $\eta^* := \min\{\frac{\rho_\epsilon - \rho^*}{16}, \frac{\rho^{**} - \rho^*}{16}\}$, $\varepsilon^\star := \min\{\varepsilon^*, \varepsilon_\epsilon\}$, $\eta^\star := \min\{\eta^*, \eta_\epsilon\}$. Then, for all sufficiently large n, we have $\varepsilon_n \in (0, \varepsilon^\star]$, $\eta_n \in (0, \eta^\star]$, and by considering $\tau_n := |\log \delta_n|$ in Theorem 3.1 we further see that the probability P^n of $\|\bar{h}_{D,\delta_n} - \bar{h}_{P,\delta_n}\|_\infty < \varepsilon_n$ converges to 1 for $n \to \infty$. Let us therefore only consider such data sets D and parameters satisfying $\varepsilon_n \in (0, \varepsilon^\star]$ and $\eta_n \in (0, \eta^\star]$. Then our construction ensures that we can apply Theorem 3.4. In particular, we have

$$\begin{aligned}\rho^* < \rho^*(D) + 2\varepsilon_n + 2\eta_n &\leq \rho^* + \varepsilon^* + \eta^* + 4\varepsilon_n + 4\eta_n \\ &\leq \rho^* + 5\varepsilon^* + 5\eta^* \leq \rho_\epsilon\,,\end{aligned}$$

and hence (10) and (11) hold for $\rho := \rho^*(D) + 2\varepsilon_n + 2\eta_n$, i.e.,

$$\mu(\bar{A}_{\rho^*,i} \setminus A_{\rho^*(D)+3\varepsilon_n+3\eta_n,i}) \leq \epsilon, \tag{12a}$$

and

$$\mu(M_{\rho^*(D)+\varepsilon_n+\eta_n} \setminus \bar{M}_{\rho^*}) \leq \epsilon. \tag{12b}$$

Using $A_i^* = \bar{A}_{\rho^*,i}$ and $A_{\rho^*(D)+3\varepsilon_n+3\eta_n,i} \subset B_i(D)$, we now obtain

$$\begin{aligned}\mu\big(A_i^* \setminus B_i(D)\big) &= \mu\big(\bar{A}_{\rho^*,i} \setminus B_i(D)\big) \\ &\leq \mu(\bar{A}_{\rho^*,i} \setminus A_{\rho^*(D)+3\varepsilon_n+3\eta_n,i}) \leq \epsilon\,. \end{aligned} \tag{13}$$

Conversely, using $\mu(B \setminus A) = \mu(B) - \mu(A \cap B)$, we obtain

$$\begin{aligned}&\mu\big(B_1(D) \setminus (A_1^* \cup A_2^*)\big) \\ &\quad= \mu(B_1(D)) - \mu\big(B_1(D) \cap (A_1^* \cup A_2^*)\big) \\ &\quad\geq \mu(B_1(D)) - \mu(B_1(D) \cap A_1^*) - \mu(B_1(D) \cap A_2^*) \\ &\quad= \mu(B_1(D) \setminus A_1^*) - \mu(B_1(D) \cap A_2^*)\,.\end{aligned}$$

Since $B_1(D) \cap B_2(D) = \emptyset$ implies $B_1(D) \cap A_2^* \subset A_2^* \setminus B_2(D)$ and Theorem 3.3 shows $B_1(D) \subset M_{\rho^*(D)+\varepsilon_n+\eta_n}$, we can thus conclude that

$$\begin{aligned}&\mu(B_1(D) \setminus A_1^*) \\ &\quad\leq \mu\big(B_1(D) \setminus (A_1^* \cup A_2^*)\big) + \mu(A_2^* \setminus B_2(D)) \\ &\quad\leq \mu\big(M_{\rho^*(D)+\varepsilon_n+\eta_n} \setminus (A_1^* \cup A_2^*)\big) + \mu(A_2^* \setminus B_2(D)) \\ &\quad\leq 2\epsilon\,,\end{aligned}$$

where in the last step we used (12) and (13). Clearly, we can establish $\mu(B_2(D) \setminus A_2^*) \leq 2\epsilon$ analogously, and hence we finally obtain $\mu(B_i(D) \triangle A_i^*) \leq 3\epsilon$ for $i = 1, 2$. □

Note that the consistency result requires us to know a minimal smoothness of h in order to ensure that $\delta_n < \delta_{\eta_n}$ for all $n \geq 1$. For example, we have already seen around (5) that for Hölder-continuous densities we know δ_η, and hence ensuring the above relation is straightforward. Moreover, for such densities, we can actually obtain finite samples guarantees, if we additionally the flat density assumption of Polonik (1995):

Theorem 4.2. *Let $X \subset \mathbb{R}^d$ be compact and P be a μ-absolutely continuous probability measure on X with a continuous density h that can be topologically clustered between the critical levels ρ^* and ρ^{**}. Furthermore, assume that h is Hölder-continuous, that is, it satisfies (5). In addition, assume that there exist constants $\theta > 0$ and $c \geq 0$ such that, for all $s > 0$, we have*

$$\mu(\{|h - \rho^*| \leq s\}) \leq c\, s^\theta. \tag{14}$$

For some fixed $a > 0$, $\tau \geq 1$, and $n \geq 3$, we define

$$\varepsilon_n := \eta_n := a\sqrt{\tau}\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}} \log(\log n)$$

and

$$\delta_n := \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$$

Suppose that we further have $n\delta_n^d \geq \max\{\tau, |\log \delta_n|\}$, $a \log(\log n) \geq \max\{C', 2^\alpha C_h \tau^{-1/2}\}$, and $16\varepsilon_n \leq \rho^{**} -$

ρ^*, where C' is the constant appearing in the simplified estimate following Theorem 3.1. Then, Algorithm 1 using the parameters ε_n, η_n, and δ_n satisfies

$$P^n\left(\Delta_{6\cdot 2^\theta c\varepsilon_n^\theta}\right) \geq 1 - e^{-\tau}\,,$$

where $\Delta_{6\cdot 2^\theta c\varepsilon_n^\theta}$ is the set defined in Theorem 4.1.

Recall that Assumption (14) goes back to Polonik (1995) and is basically used in every paper dealing with learning rates for density level set estimation. Interestingly, we do not need this assumption for a range of values ρ around ρ^*, as one might conjecture considering the fact that we do estimate ρ^* as well, but only for the (unknown) level ρ^*. Furthermore, note that our algorithm does not need to know the constants ρ^*, c and θ in Assumption (14). Moreover, the $\log(\log(n))$ term in the definition of ε_n and η_n makes it possible that the algorithm only needs to know the Hölder-exponent α but not the Hölder-constant C_h.

Under the assumptions of Theorem 4.2, Algorithm 1 learns, modulo logarithmic factors, the clusters with rate

$$n^{-\frac{\alpha\theta}{2\alpha+d}}\,.$$

In particular, if h is C^1, then it is known that we can pick $\theta = 1$, and hence we obtain the rate $n^{-\frac{\alpha}{2\alpha+d}}$. Moreover, modulo logarithmic factors our rates equal the optimal rates $n^{-\frac{\alpha\theta}{2\alpha+d}}$ for plain level set estimation, obtained by e.g. Rigollet and Vert (2009), and hence it is obvious that our rates are optimal modulo a logarithmic factor. The fact that we essentially obtain the same rates as for plain level set estimation is somewhat remarkable, since *a)* for clustering the estimation must be stronger in order to guarantee that the topological properties are correctly recognized, and *b)* estimating the correct level ρ^* *simultaneously* may add another layer of difficulty compared to level set estimation for fixed level ρ.

Proof. Throughout the proof we use the notations from the proof of Theorem 4.1. As discussed in (a_5), we have $\bar{M}_{\rho^*} = \{h > \rho^*\}$, and using (a_1) we thus we find, for $\rho > \rho^*$ and $\varepsilon, \eta > 0$,

$$\begin{aligned}\mu(M_{\rho-\varepsilon-\eta} \setminus \bar{M}_{\rho^*}) &= \mu(M_{\rho-\varepsilon-\eta} \setminus \{h > \rho^*\}) \\ &\leq \mu(\{h \geq \rho - \varepsilon - \eta\} \setminus \{h > \rho^*\}) \\ &= \mu(\{\rho - \varepsilon - \eta \leq h \leq \rho^*\}) \\ &\leq c(\varepsilon+\eta)^\theta\end{aligned}$$

and

$$\begin{aligned}\mu(\bar{M}_{\rho^*} \setminus M_{\rho+\varepsilon+\eta}) &\leq \mu(\{h > \rho^*\} \setminus \{h > \rho + \varepsilon + \eta\}) \\ &= \mu(\{\rho^* < h \leq \rho + \varepsilon + \eta\}) \\ &\leq c(\varepsilon+\eta)^\theta\,.\end{aligned}$$

Moreover, we have $\bar{A}_{\rho^*,i} \cap A_{\rho,j} = \emptyset$ for all $\rho > \rho^*$ and all $i \neq j$, and hence we obtain $\bar{A}_{\rho^*,i} \setminus A_{\rho+\varepsilon+\eta,i} = \bar{M}_{\rho^*} \setminus A_{\rho+\varepsilon+\eta,i}$. With the help of our previous estimate, this yields

$$\mu(\bar{A}_{\rho^*,i} \setminus A_{\rho+\varepsilon+\eta,i}) \leq c(\varepsilon+\eta)^\theta\,,$$

and by the monotonicity of the components $A_{\rho+\varepsilon+\eta,i}$ in $\varepsilon+\eta$, we hence have established (10) and (11) for $\varepsilon_\epsilon := \eta_\epsilon := \frac{1}{2}(\epsilon/c)^{1/\theta}$, $\rho_\epsilon := \rho^* + 16\varepsilon_\epsilon$, and all $\varepsilon \in (0, \varepsilon_\epsilon]$, $\eta \in (0, \eta_\epsilon]$, and $\rho \in (\rho^*, \rho_\epsilon]$. Note that, if $16\varepsilon_\epsilon \leq \rho^{**} - \rho^*$, then the quantities in the proof of Theorem 4.1 become $\varepsilon^* = \varepsilon_\epsilon$ and $\varepsilon^\star = \varepsilon_\epsilon$, and analogously, $\eta^\star = \varepsilon_\epsilon$. Let us consider $\epsilon := 2^\theta c\varepsilon_n^\theta$. Then we have $\varepsilon_n = \varepsilon^\star$ and $\eta_n = \eta^\star$, and our assumption $16\varepsilon_n \leq \rho^{**} - \rho^*$ thus guarantees $16\varepsilon_\epsilon \leq \rho^{**} - \rho^*$. Moreover, our assumption $a\log(\log n) \geq 2^\alpha C_h \tau^{-1/2}$ ensures $\delta_n < \delta_{\eta_n}$. Then, the proof of Theorem 4.1 shows that, for all data sets $D \in X^n$ with $\|\bar{h}_{D,\delta_n} - \bar{h}_{P,\delta_n}\|_\infty < \varepsilon_n$, we have

$$\mu(B_1(D) \,\triangle\, A_1^*) + \mu(B_2(D) \,\triangle\, A_2^*) \leq 6\epsilon = 6\cdot 2^\theta c\varepsilon_n^\theta\,.$$

It therefore remains to determine the probability of such data sets D. To this end, we recall that Theorem 3.1 can be simplified to

$$\begin{aligned}P^n\left(\left\{D \in X^n : \|\bar{h}_{D,\delta_n} - \bar{h}_{P,\delta_n}\|_\infty < C'\sqrt{\frac{\tau|\log\delta_n|}{n\delta_n^d}}\right\}\right) \\ \geq 1 - e^{-\tau}.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}C'\sqrt{\frac{\tau|\log\delta_n|}{n\delta_n^d}} &= C' n^{-\frac{\alpha}{2\alpha+d}} \sqrt{\frac{\tau|\log\delta_n|}{(\log n)^{\frac{d}{2\alpha+d}}}} \\ &\leq C' n^{-\frac{\alpha}{2\alpha+d}} \sqrt{\frac{\tau\log n}{(\log n)^{\frac{d}{2\alpha+d}}}} \\ &= C'\sqrt{\tau}\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}} \\ &\leq \varepsilon_n\,,\end{aligned}$$

where in the last step we used the assumption $a\log(\log n) \geq C'$. Combining all estimates we now obtain the assertion. □

Acknowledgments

Part of the work was done while B. K. S. was visiting I. S. at the Institute for Stochastics and Applications. B. K. S. wishes to acknowledge support from the Gatsby Charitable Foundation and the University of Stuttgart. I. S. thanks Michael Eisermann for fruitful discussions on connectivity.

References

Ankerst, M., Breunig, M., Kriegel, H.-P., and Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. In *ACM SIGMOD international conference on Management of data*, pages 49–60. ACM Press.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334:495–500.

Carmichael, J., George, G., and Julius, R. (1968). Finding natural clusters. *Systematic Zoology*, 17:144–150.

Chaudhuri, K. and Dasgupta, S. (2010). Rates of convergence for the cluster tree. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 343–351.

Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *Ann. Statist.*, 25:2300–2312.

Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer, New York.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press.

Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: Rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist.*, 37:503–522.

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

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, New York.

Kpotufe, S. and von Luxburg, U. (2011). Pruning nearest neighbor cluster trees. In Getoor, L. and Scheffer, T., editors, *Proceedings of the* 28^{th} *International Conference on Machine Learning*, pages 225–232. ACM, New York, NY, USA.

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

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass aproach. *Ann. Statist.*, 23:855–881.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392.

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

Rinaldo, A. and Wasserman, L. (2010). Generalized density clustering. *Ann. Statist.*, 38:2678–2722.

Steinwart, I. (2011). Adaptive density level set clustering. In *Conference on Learning Theory*. To appear.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer, New York.

Stuetzle, W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20:25–47.

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.