Adaptive Clustering Using Kernel Density Estimators

Ingo Steinwart*
Institute for Stochastics and Applications
Faculty 8: Mathematics and Physics
University of Stuttgart
D-70569 Stuttgart Germany
ingo.steinwart@mathematik.uni-stuttgart.de

Bharath K. Sriperumbudur
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA
bks18@psu.edu

Philipp Thomann
D ONE Solutions AG
Sihlfeldstrasse 58
8003 Zürich
philipp.thomann@d1-solutions.com

Abstract

We derive and analyze a generic, recursive algorithm for estimating all splits in a finite cluster tree as well as the corresponding clusters. We further investigate statistical properties of this generic clustering algorithm when it receives level set estimates from a kernel density estimator. In particular, we derive finite sample guarantees, consistency, rates of convergence, and an adaptive data-driven strategy for choosing the kernel bandwidth. For these results we do not need continuity assumptions on the density such as Hölder continuity, but only require intuitive geometric assumptions of non-parametric nature.

MSC2010 Subject Classification: Primary: 62H30; **Secondary:** 62H12, 62G05
Keywords: cluster analysis, kernel density estimation, consistency, rates, adaptivity

1 Introduction

A widely acknowledged problem in cluster analysis is the definition of a learning goal that describes a conceptually and mathematically convincing definition of clusters. One such definition, which goes back to Hartigan [7] and is known as *density-based clustering*, assumes i.i.d. data $D = (x_1, \dots, x_n)$ generated by some unknown distribution P. Given some $\rho \geq 0$, the clusters of P are then defined to be the connected components of the level set $\{h \geq \rho\}$, where h is the density associated with P w.r.t. the Lebesgue measure. This *single level approach* has been studied,

*Corresponding author

for example in [7, 3, 12, 10, 13]. However, one of the conceptual drawbacks of the single level approach is that different values of ρ may lead to different (numbers of) clusters, and there is also no general rule for choosing ρ, either. To address this conceptual shortcoming, one often considers the so-called *cluster tree approach* instead, which tries to consider all levels and the corresponding connected components simultaneously.

If the focus lies on the identification of the *hierarchical tree structure* of the connected components, then one can find a variety of articles investigate properties of the cluster tree approach, see e.g. [7, 20, 2, 21, 9, 23] for details. For example, [2] shows, under some assumptions on h, that a modified single linkage algorithm recovers this tree in the sense of [8], and [9] obtains similar results for an underlying k-NN density estimator. In addition, [9] proposes a simple pruning strategy, that removes connected components that artificially occur because of finite sample variability. However, the notion of recovery taken from [8] only focuses on the correct estimation of the cluster tree structure and not on the estimation of the clusters itself, cf. the discussion in [15]. Finally, the most recent paper [23] establishes guarantees including rates of convergence for each fixed level set, provided that a kernel-density estimator is used to produce level set estimates and the density has a certain behavior such as α-Hölder continuity.

A third approach taken in [15, 14, 16] tries to estimate both the *first split ρ^* in the cluster tree*, and the corresponding clusters. As in the previously discussed papers, finite sample bounds are derived, which in [16] are extended to learning rates, which can also be obtained by an adaptive, fully data-driven hyper-parameter selection strategy. Unfortunately, however, [15, 16] only consider the simplest possible density estimator, namely a histogram approach, and [14] restricts its considerations to compactly supported moving window density estimates for α-Hölder-continuous densities. In addition, the method in [14] requires to know α, and in particular, it is not data-driven. Finally, all three papers completely ignore the behavior of the considered algorithm for *single cluster* distributions, i.e. for distributions that do not have a split in the cluster tree. As a consequence, it is unclear whether and how a suitably modified version of this algorithm can be used to estimate the *split-tree*, i.e. all levels at which a split in the cluster tree occurs, as well as the resulting clusters at these splits

The goal of this paper is to address the discussed issues of [15, 14, 16]. To be more precise, compared to these articles, we establish the following new results:

i) For single cluster distributions, we propose a new set of regularity assumptions for levels ρ at which the level set $\{h \geq \rho\}$ is small. For example, for bounded densities, these assumptions roughly speaking guarantee, that the level sets do not frazzle for levels ρ close to the maximum $\|h\|_\infty$ of the density h. Such assumptions were missing in [15, 14, 16].

ii) We present a simple modification of the output behavior of the generic cluster algorithm of [16] to deal with distributions that do not have a split in the cluster tree. Based on our new regularity assumptions in *i)* and the ones from [16], we then show that this new cluster algorithm is able to: *a)* estimate the first split ρ^* in the cluster tree whenever there is one; and *b)* correctly detect distributions for which there is no such split ρ^*. Note that from a technical side, *a)* directly follows from [16], since our modification of the generic cluster algorithm scans through candidate levels ρ in exactly the same way as the original algorithm of [16] does. Therefore, the surprising, and compared to [16] new part of our finite sample guarantee is the fact that this scanning procedure does not need to be changed for correctly detecting single cluster distributions in *b)*. Note that a highly beneficial side-effect of this fact is that our analysis in *b)* as well as in *iii)* and *iv)* below, can rely on the extensive set of tools developed in [16].

iii) We then show how the results of *ii)* can be used to estimate the entire *split-tree* by recursively applying the new generic cluster algorithm. While from a higher perspective this result does not seem to be too surprising, it turns out that there are still a couple of serious technical difficulties involved. In a nutshell, these difficulties relate to the fact, that the generic algorithm may return an estimate ρ_{out} for ρ^* for which the connected components of $\{h \geq \rho_{\text{out}}\}$ are not yet sufficiently apart from each other. While such an estimate ρ_{out} for ρ^* is desirable, it also prohibits a direct recursive application of the results of *ii)*. To address this issue, we analyze the behavior of the generic cluster algorithm above the returned level ρ_{out}. In this analysis, which also goes beyond [15, 14, 16], it turns out, that the algorithm behaves correctly until it reaches a level ρ for which the connected components of $\{h \geq \rho\}$ are sufficiently apart from each other. Above this level ρ, the results of *ii)* can then be recursively applied, leading to guarantees for the entire split-tree.

iv) We show that the new generic cluster algorithm does not only work with an underlying histogram density estimator (HDEs) as in [15, 16], but also for a variety of kernel density estimators (KDEs). Here it turns out that the results of [16], including those for the adaptive, fully data-driven hyper-parameter selection strategy, remain valid for the resulting new clustering algorithm, provided that the kernel has a bounded support. Moreover, if the kernel has an exponential tail behavior, then the results remain true modulo an extra logarithmic term, while in the case of even heavier tails, we show that the rates become worse by a polynomial factor. Note that compared to [16], all the results for KDEs are new. Moreover, the results for KDEs substantially extend the results of [14], since there only moving window kernels were treated, and since [14] only considered α-Hölder continuous densities with known α. In contrast, our new results do not even require continuous densities, and for this reason, we also obtain significantly more general results than the currently best results for KDE-based clustering achieved in [23]. The latter improvement is partially made possible, because we can rely on the tools of [16]. However, compared to the HDEs in [16] considering KDEs still requires significant technical efforts such as finite sample bound for the $\|\cdot\|_\infty$-distance between a KDE and its population version.

The rest of this paper is organized as follows: In Section 2 we briefly recall the key concepts of [16]. In Section 3 we first introduce the new regularity assumptions mentioned in *i)*. We then introduce and analyze the new generic cluster algorithm as described in *ii)*. Moreover, the recursive approach described in *iii)* is analyzed in detail. Section 4 then presents key uncertain guarantees for level sets generated by KDEs, and Section 5 contains the material mentioned in *iv)*, namely finite sample bounds as well as consistency results, rates of convergence, and an adaptive data-driven strategy for choosing the kernel bandwidth. All proofs can be found in Section 6.

2 Preliminaries

In this section we recall from [16] the setup for defining density-based clusters in a general context. To this end let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then we denote the closed unit ball of this norm by $B_{\|\cdot\|}$ and write $B_{\|\cdot\|}(x,\delta) := x + \delta B_{\|\cdot\|}$ for the closed ball with center $x \in \mathbb{R}^d$ and radius $\delta > 0$. If the norm is known from the context, we usually write $B(x,\delta)$ instead. Moreover, the Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|_2$ and for the Lebesgue volume of its unit ball we write vol_d. Finally, $\|\cdot\|_\infty$ denotes the supremum norm for functions.

Let us now assume that we have some $A \subset X \subset \mathbb{R}^d$ as well as some norm $\|\cdot\|$ on $\mathbb{R}^d$. Then, for $\delta > 0$ we define

$$A^{+\delta} := \{x \in X : d(x, A) \le \delta\}\,, \qquad \text{and} \qquad A^{-\delta} := X \setminus (X \setminus A)^{+\delta}\,,$$

where $d(x, A) := \inf_{x' \in A} \|x - x'\|$. We further write $\mathring{A}$ for the interior of A and $\overline{A}$ for the closure of A. Moreover, $\partial A := \overline{A} \setminus \mathring{A}$ denotes the boundary of A. Finally, $\mathbf{1}_A$ denotes the indicator function of A, and $A \vartriangle B$ the symmetric difference of two sets A and B.

Let us now assume that P is a non-zero finite measure[1] on a closed $X \subset \mathbb{R}^d$ that is absolutely continuous with respect to the Lebesgue measure λ^d. Then P has a λ^d-density h and as explained in the introduction, one could define the clusters of P to be the connected components of the level set $\{h \ge \rho\}$, where $\rho \ge 0$ is some user-defined threshold. Unfortunately, however, this notion leads to serious issues if there is no canonical choice of h such as a continuous version, see the illustrations in [16, Section 2.1]. To address this issue, [16] considered, for $\rho \ge 0$, the measures

$$\mu_\rho(A) := \lambda^d(A \cap \{h \ge \rho\})\,, \qquad A \in \mathcal{B}(\mathbb{R}^d)\,.$$

Since μ_ρ is independent of the choice of $h := \mathrm{d}P/\,\mathrm{d}\lambda^d$, the set

$$M_\rho := \operatorname{supp} \mu_\rho\,, \tag{1}$$

where $\operatorname{supp} \mu_\rho$ denotes the support of the measure μ_ρ, is independent of this choice, too. For any λ^d-density h of P, the definition immediately gives

$$\lambda^d\big(\{h \ge \rho\} \setminus M_\rho\big) = \lambda^d\big(\{h \ge \rho\} \cap (\mathbb{R}^d \setminus M_\rho)\big) = \mu_\rho(\mathbb{R}^d \setminus M_\rho) = 0\,, \tag{2}$$

i.e. modulo λ^d-zero sets, the level sets $\{h \ge \rho\}$ are not larger than M_ρ. In fact, M_ρ turns out to be the smallest closed set satisfying (2) and it is shown in [17, Lemma A.1.2], that

$$\{h \mathring{\ge} \rho\} \subset M_\rho \subset \overline{\{h \ge \rho\}} \qquad \text{and} \qquad M_\rho \vartriangle \{h \ge \rho\} \subset \partial\{h \ge \rho\}\,. \tag{3}$$

In order to ensure inclusions that are "inverse" to (2), [16] said that P is normal at level ρ if there exist two λ^d-densities h_1 and h_2 of P such that

$$\lambda^d(M_\rho \setminus \{h_1 \ge \rho\}) = \lambda^d(\{h_2 > \rho\} \setminus \mathring{M}_\rho) = 0\,.$$

It is shown in [17, Lemma A.1.3][2] that P is normal at every level, if it has both an upper semi-continuous λ^d-density h_1 and a lower semi-continuous λ^d-density h_2. Moreover, if P has a λ^d-density h such that $\lambda^d(\partial\{h \ge \rho\}) = 0$, then the same lemma shows that P is normal at level ρ. Finally, note that if the conditions of normality at level ρ are satisfied for some λ^d-densities h_1 and h_2 of P, then they are actually satisfied for all λ^d-densities h of P and we have $\lambda^d(M_\rho \vartriangle \{h \ge \rho\}) = 0$. The next assumption collects the concepts introduced so far.

Assumption P. We have a compact and connected $X \subset \mathbb{R}^d$, μ denotes the Lebesgue measure on X, and P is a μ-absolutely continuous, finite, non-zero measure that is normal at every level.

Let us now recall the definition of clusters from [16]. We begin with the following definition that compares different partitions.

[1] In [16] only probability measures were considered, but for later use it is more convenient to consider finite measures instead. It it easy to check that all results mentioned in the following remain true for finite measures by a simple reparametrization of the levels ρ.

[2] In this lemma, the term "upper normal at level ρ" means that $\lambda^d(M_\rho \setminus \{h_1 \ge \rho\}) = 0$ for some density $h_1 := \mathrm{d}P/\,\mathrm{d}\lambda^d$ while "lower normal at level ρ" means $\lambda^d(\{h_2 > \rho\} \setminus \mathring{M}_\rho) = 0$ for some density $h_2 := \mathrm{d}P/\,\mathrm{d}\lambda^d$.

Definition 2.1. *Let $A \subset B$ be non-empty sets, and $\mathcal{P}(A)$ and $\mathcal{P}(B)$ be partitions of A and B, respectively. Then $\mathcal{P}(A)$ is comparable to $\mathcal{P}(B)$, write $\mathcal{P}(A) \sqsubset \mathcal{P}(B)$, if, for all $A' \in \mathcal{P}(A)$, there is a $B' \in \mathcal{P}(B)$ with $A' \subset B'$.*

Informally speaking, $\mathcal{P}(A)$ is comparable to $\mathcal{P}(B)$, if no cell $A' \in \mathcal{P}(A)$ is broken into pieces in $\mathcal{P}(B)$. In particular, if $\mathcal{P}_1$ and $\mathcal{P}_2$ are two partitions of A, then $\mathcal{P}_1 \sqsubset \mathcal{P}_2$ if and only if $\mathcal{P}_1$ is finer than $\mathcal{P}_2$. Let us now assume that we have have two partitions $\mathcal{P}(A)$ and $\mathcal{P}(B)$ with $\mathcal{P}(A) \sqsubset \mathcal{P}(B)$. Then [17, Lemma A.2.1] shows that there exists a unique map $\zeta : \mathcal{P}(A) \to \mathcal{P}(B)$ such that, for all $A' \in \mathcal{P}(A)$, we have

$$A' \subset \zeta(A') .$$

Following [15, 16], we call ζ the cell relating map (CRM) between A and B.

The first example of comparable partitions come from connected components. To be more precise, let $A \subset \mathbb{R}^d$ be a closed subset and $\mathcal{C}(A)$ be the collection of its connected components. By definition, $\mathcal{C}(A)$ forms a partition of A, and if $B \subset \mathbb{R}^d$ is another closed subset with $A \subset B$ and $|\mathcal{C}(B)| < \infty$ then we have $\mathcal{C}(A) \sqsubset \mathcal{C}(B)$, see [17, Lemma A.2.3].

Following [16], another class of partitions arise from a discrete notion of path-connectivity. To recall the latter, we fix a $\tau > 0$, an $A \subset \mathbb{R}^d$, and a norm $\|\cdot\|$ on $\mathbb{R}^d$. Then $x, x' \in A$ are τ-connected in A, if there exist $x_1, \dots, x_n \in A$ such that $x_1 = x$, $x_n = x'$ and $\|x_i - x_{i+1}\| < \tau$ for all $i = 1, \dots, n-1$. Clearly, being τ-connected gives an equivalence relation on A. We write $\mathcal{C}_\tau(A)$ for the resulting partition and call its cells the *τ-connected components of A*. It has been shown in [17, Lemma A.2.7], that $\mathcal{C}_\tau(A) \sqsubset \mathcal{C}_\tau(B)$ for all $A \subset B$ and $\tau > 0$. Moreover, if $|\mathcal{C}(A)| < \infty$ then $\mathcal{C}(A) = \mathcal{C}_\tau(A)$ for all sufficiently small $\tau > 0$, see [16, Section 2.2] for details.

We can now describe finite measures that can be clustered.

Definition 2.2. *Let* Assumption P *be satisfied. Then P can be clustered between $\rho^* \geq 0$ and $\rho^{**} > \rho^*$, if for all $\rho \in [0, \rho^{**}]$, the following three conditions are satisfied:*

i) We either have $|\mathcal{C}(M_\rho)| = 1$ or $|\mathcal{C}(M_\rho)| = 2$.

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

iii) If we have $|\mathcal{C}(M_\rho)| = 2$, then $\rho \geq \rho^*$ and $\mathcal{C}(M_{\rho^{**}}) \sqsubseteq \mathcal{C}(M_\rho)$.

Using the CRMs $\zeta_\rho : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$, we then define the clusters of P by

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

where A_1 and A_2 are the topologically connected components of $M_{\rho^{**}}$. Finally, we define

$$\tau^*(\varepsilon) := \frac{1}{3} \cdot d\big(\zeta_{\rho^*+\varepsilon}(A_1), \zeta_{\rho^*+\varepsilon}(A_2)\big) , \qquad \varepsilon \in (0, \rho^{**} - \rho^*]. \tag{4}$$

Definition 2.2 ensures that the level sets below ρ^* are connected, while for a certain range above ρ^* the level sets have exactly two components, which, in addition, are assumed to be comparable. Consequently, the topological structure between ρ^* and ρ^{**} is already determined by that of $M_{\rho^{**}}$, and we can use the connected components of $M_{\rho^{**}}$ to number the connected components of M_ρ for $\rho \in (\rho^*, \rho^{**})$. This is done in the definition of the clusters A_i^* as well as in the definition of the function τ^*, which essentially measures the distance between the two connected components at level $\rho^* + \varepsilon$.

The major goal of [15, 16] was to design an algorithm that is able to asymptotically estimate both the correct value of ρ^* and the clusters A_1^* and A_2^*. Moreover, [16] established rates of convergence for both estimation problems, and these rates did depend on the behavior of the function τ^*. However, this algorithm required that the level sets do not have bridges or cusps that are too thin. To make this precise, let us recall that for a closed $A \subset \mathbb{R}^d$, [15, 16] considered the function $\psi_A^* : (0,\infty) \to [0,\infty]$ defined by

$$\psi_A^*(\delta) := \sup_{x\in A} d(x, A^{-\delta}) , \qquad\qquad \delta > 0.$$

Roughly speaking, $\psi_A^*(\delta)$ describes the smallest diameter ε needed to "recover" A from $A^{-\delta}$ in the sense of $A \subset (A^{-\delta})^{+\varepsilon}$, see [17, Section A.5] for this and various other results on ψ_A^*. In particular, we have $\psi_A^*(\delta) \geq \delta$ for all $\delta > 0$ and $\psi_A^*(\delta) = \infty$ if $A^{-\delta} = \emptyset$. Moreover, ψ_A^* behaves linearly, if bridges and cusps of A are not too thin, and even thinner cusps and bridges can be included by considering sets with $\psi_A^*(\delta) \leq c\delta^\gamma$ for some constant c and all sufficiently small $\delta > 0$. Finally, for our later results we need to recall from [17, Lemma A.4.3] that for all $\delta > 0$ with $A^{-\delta} \neq \emptyset$ and all $\tau > 2\psi_A^*(\delta)$ we have

$$|\mathcal{C}_\tau(A^{-\delta})| \leq |\mathcal{C}(A)| \tag{5}$$

whenever A is contained in some compact $X \subset \mathbb{R}^d$ and $|\mathcal{C}(A)| < \infty$.

With the help of these preparations we can now recall the following definition taken from [16], which categorizes the behavior of $\psi_{M_\rho}^*$.

Definition 2.3. *Let* Assumption P *be satisfied. Then we say that P has thick level sets of order $\gamma \in (0,1]$ up to the level $\rho^{**} > 0$, if there exist constants $c_{\text{thick}} \geq 1$ and $\delta_{\text{thick}} \in (0,1]$ such that, for all $\delta \in (0, \delta_{\text{thick}}]$ and $\rho \in [0, \rho^{**}]$, we have*

$$\psi_{M_\rho}^*(\delta) \leq c_{\text{thick}}\, \delta^\gamma . \tag{6}$$

In this case, we call $\psi(\delta) := 3c_{\text{thick}}\delta^\gamma$ the thickness function of P.

The following assumption, which collects all concepts introduced so far, describes the finite measures we wish to cluster.

Assumption M. The finite measure P can be clustered between ρ^* and ρ^{**}. In addition, P has thick level sets of order $\gamma \in (0,1]$ up to the level ρ^{**}. We denote the corresponding thickness function by ψ and write τ^* for the function defined in (4).

Recall that the theory developed in [15, 14, 16] focused on the question, whether it is possible to estimate ρ^* and the resulting clusters for distributions that can be clustered. To this end, a generic cluster algorithm was developed, which receives some level set estimates $L_{D,\rho}$ of M_ρ satisfying

$$M_{\rho+\varepsilon}^{-\delta} \subset L_{D,\rho} \subset M_{\rho-\varepsilon}^{+\delta} \tag{7}$$

for all $\rho \in [0, \rho^{**}]$ and some $\varepsilon, \delta > 0$. The key result [16, Theorem 2.9] then specified in terms of ε and δ how well this algorithm estimates both ρ^* and the clusters A_1^* and A_2^*. What is missing in this analysis, however, is an investigation of the behavior of the generic cluster algorithm in situations in which P cannot be clustered because all level sets are connected.

Now observe that the reason for this gap was the notion of thickness: Indeed, if P is a *single-cluster finite measure*, i.e. $|\mathcal{C}(M_\rho)| \leq 1$ for all $\rho \geq 0$, and P has thick level sets of the order γ up to the level $\rho^{**} := \sup\{\rho : \rho \geq 0 \text{ and } |\mathcal{C}(M_\rho)| = 1\}$, then the proof of [16, Theorem 2.9] can be easily extended to show that at each visited level ρ the algorithm correctly detects exactly one connected component. Unfortunately, however, the assumption of having thick levels up to the height ρ^{**} of the peak of h is too unrealistic, as it requires $M_\rho^{-\delta} \neq \emptyset$ for all $\rho \in [0, \rho^{**}]$ and $\delta \in (0, \delta_{\text{thick}}]$, that is, *the peak needs to be a plateau that contains a ball of radius δ_{thick}*.

3 A Generic Algorithm for Estimating the Split-Tree

The overall goal of this section is to present a generic algorithm for estimating the entire split-tree. To this end, we first introduce a new set of assumptions for single-cluster distributions that rule out irregular behavior of the level sets in the vicinity of the peak of the density. Unlike the naïve approach for extending the results of [16] to single-cluster distributions, which we have discussed at the end of Section 2, this new set of assumptions includes a variety of realistic behaviors. In the second step we then present a generic cluster algorithm, whose only difference to the one in [16] is its output behavior in situations in which no split has been detected. We then show that this new cluster algorithm, like its predecessor in [16], correctly identifies a split in the cluster tree. Moreover, we demonstrate that, unlike the one in [16], the new cluster algorithm also correctly identifies single-cluster distributions. Finally, we combine these insights to develop a new generic algorithm for estimating the entire cluster tree.

Let us begin by introducing the already mentioned new assumption for dealing with single-cluster distributions.

Assumption S. Assumption P is satisfied and there are $\rho_* \geq 0$, $\gamma \in (0,1]$, $c_{\text{thick}} \geq 1$ and $\delta_{\text{thick}} \in (0,1]$ such that for all $\rho \geq \rho_*$ and $\delta \in (0, \delta_{\text{thick}}]$, we have:

i) $|\mathcal{C}(M_\rho)| \leq 1$.

ii) If $M_\rho^{-\delta} \neq \emptyset$ then $\psi^*_{M_\rho}(\delta) \leq c_{\text{thick}}\delta^\gamma$.

iii) If $M_\rho^{-\delta} = \emptyset$, then for all $\emptyset \neq A \subset M_\rho^{+\delta}$ and $\tau > 2c_{\text{thick}}\delta^\gamma$ we have $|\mathcal{C}_\tau(A)| = 1$.

iv) For each $\delta \in (0, \delta_{\text{thick}}]$ there exists a $\rho \geq \rho_*$ with $M_\rho^{-\delta} = \emptyset$.

Note that condition *i)* simply means that the level sets of P above ρ_* are either empty or connected. If they are empty, there is nothing more to assume and in the other case, we can either have $M_\rho^{-\delta} \neq \emptyset$ or $M_\rho^{-\delta} = \emptyset$. If $M_\rho^{-\delta} \neq \emptyset$, then condition *ii)* ensures that the level set M_ρ is still thick in the sense of Definition 2.3, while in the other case $M_\rho^{-\delta} = \emptyset$, condition *iii)* guarantees that the larger sets $M_\rho^{+\delta}$ cannot have multiple τ-connected components as long as we choose τ in a way that is required in the case of multiple clusters, too. Finally, *iv)* is satisfied, for example, if P has a bounded density, since this even guarantees $M_\rho = \emptyset$ for all $\rho > \|h\|_\infty$.

Clearly, if all M_ρ above ρ_* are balls, then it is easy to check that Assumption S is satisfied, and this remains true if one slightly perturbed the shape. Moreover, using the examples discussed in [18], it is easy to construct further examples of realistic distributions satisfying Assumption S. On the other hand, Assumption S excludes, for example, distributions whose level sets have a "bone shape" and whose peak is a horizontal ridge.

The next task is to formulate a generic algorithm that is able to estimate ρ^* and the resulting clusters if P can be clustered in the sense of Assumption M and that is able to detect distributions that cannot be clustered in the sense of Assumption S. We will see in the following that Algorithm 1 is such an algorithm. Before we present the corresponding results we first note that the only difference of Algorithm 1 to the algorithm considered in [16] is the more flexible start level ρ_0, compared to $\rho_0 = 0$ in [16], and the modified output in Lines 8-12. Indeed, the algorithm in [16] always produces the return values of Line 9. In contrast, Algorithm 1 distinguishes between the cases $M > 1$ and $M = 0$. While for $M > 1$ the output of both algorithms exactly coincide, the new Algorithm 1 now returns ρ_0 and L_{ρ_0} in the case of $M = 0$. We will see in Theorem 3.2 that the latter case typically occurs for distributions satisfying Assumption S. In this respect recall

Algorithm 1 Clustering with the help of a generic level set estimator

Require: Some $\tau > 0$ and $\varepsilon > 0$ and a start level $\rho_0 \geq 0$.
A decreasing family $(L_\rho)_{\rho\geq 0}$ of subsets of X.
Ensure: An estimate of ρ_* or ρ^* and the corresponding clusters.

$\rho \leftarrow \rho_0$
repeat
  Identify the τ-connected components $B_1, \dots, B_M$ of L_ρ satisfying

$$B_i \cap L_{\rho+2\varepsilon} \neq \emptyset.$$

  $\rho \leftarrow \rho + \varepsilon$
until $M \neq 1$
$\rho \leftarrow \rho + 2\varepsilon$
Identify the τ-connected components $B_1, \dots, B_M$ of L_ρ satisfying

$$B_i \cap L_{\rho+2\varepsilon} \neq \emptyset.$$

if $M > 1$ **then**
  return $\rho_{\text{out}} = \rho$ and the sets B_i for $i = 1, \dots, M$.
else
  return $\rho_{\text{out}} = \rho_0$ and the set L_{ρ_0}.
end if

that L_{ρ_0} can be viewed as an estimate of M_{ρ_0} and therefore returning L_{ρ_0} makes sense for such distributions.

With these preparations we can now formulate the following adaptation of [16, Theorem 2.9] to the new Algorithm 1. Since the proof of [16, Theorem 2.9] can be easily adapted to arbitrary start levels $\rho_0 \geq 0$ and this proof also shows that the case $M \leq 1$ is not occurring under the assumptions of this theorem, we omit the proof of Theorem 3.1.

Theorem 3.1. *Let* Assumption M *be satisfied. Furthermore, let* $\varepsilon^* \leq (\rho^{**} - \rho^*)/9$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi(\delta), \tau^*(\varepsilon^*)]$, *and* $\varepsilon \in (0, \varepsilon^*]$, *and* $\rho_0 \leq \rho^*$. *In addition, let* $(L_\rho)_{\rho\geq 0}$ *be a decreasing family satisfying* (7) *for all* $\rho \geq \rho_0$. *Then we have:*

i) The level ρ_{out} *returned by Algorithm 1 satisfies* $\rho_{\text{out}} \in [\rho^* + 2\varepsilon, \rho^* + \varepsilon^* + 5\varepsilon]$ *and*

$$\tau - \psi(\delta) < 3\tau^*\big(\rho_{\text{out}} - \rho^* + \varepsilon\big)\,. \tag{8}$$

ii) Algorithm 1 returns two sets B_1 *and* B_2 *and these can be ordered such that we have*

$$\sum_{i=1}^{2} \mu\big(B_i \vartriangle A_i^*\big) \leq 2\sum_{i=1}^{2} \mu\big(A_i^* \setminus (A^i_{\rho_{\text{out}}+\varepsilon})^{-\delta}\big) + \mu\big(M^{+\delta}_{\rho_{\text{out}}-\varepsilon} \setminus \{h > \rho^*\}\big)\,. \tag{9}$$

Here, $A^i_{\rho_{\text{out}}+\varepsilon} \in \mathcal{C}(M_{\rho_{\text{out}}+\varepsilon})$ *are ordered in the sense of* $A^i_{\rho_{\text{out}}+\varepsilon} \subset A_i^*$.

Theorem 3.1 shows that the modified Algorithm 1 is still able to estimate ρ^* and the corresponding clusters if the distribution can be clustered in the sense of Assumption M. The main motivation for this section was, however, to have an algorithm that also behaves correctly for distributions that cannot be clustered in the sense of Assumption S. The next theorem shows that Algorithm 1 does indeed have such a behavior.

Algorithm 2 Estimating the split-tree with the help of a generic level set estimator

Require: Some $\tau > 0$, $\varepsilon > 0$ and a start level $\rho_0 \geq 0$.
decreasing family $(L_\rho)_{\rho \geq 0}$ of subsets of X.
Ensure: Estimates of all splits of the cluster tree and the corresponding clusters.

```
Call Algorithm 1 with ρ_0 and (L_ρ)_{ρ≥0} and store its return values in the split-tree
while ρ_out > ρ_0 do
   Call Algorithm 2 with ρ_out + ε and (L_{1,ρ})_{ρ≥0} and store its return values in the split-tree
   Call Algorithm 2 with ρ_out + ε and (L_{2,ρ})_{ρ≥0} and store its return values in the split-tree
end while
return split-tree
```

Theorem 3.2. *Let* Assumption S *be satisfied and* $(L_\rho)_{\rho \geq 0}$ *be a decreasing family of sets* $L_\rho \subset X$ *such that* (7) *holds for some fixed* $\varepsilon, \delta > 0$ *and all* $\rho \geq \rho_0$*. If* $\rho_0 \geq \rho_*$*,* $\delta \in (0, \delta_{\mathrm{thick}}]$*, and* $\tau > 2c_{\mathrm{thick}}\delta^\gamma$*, then Algorithm 1 returns* ρ_0 *and* L_0*.*

Note that Theorem 3.1 requires $\tau > \psi(d) = 3c_{\mathrm{thick}}\delta^\gamma$, while Theorem 3.2 even holds under the milder assumption $\tau > 2c_{\mathrm{thick}}\delta^\gamma$. Consequently, if we choose a τ with $\tau > 3c_{\mathrm{thick}}\delta^\gamma$, then the corresponding assumptions of both theorems are satisfied. Moreover, the additional assumption $\tau < \tau^*(\varepsilon^*)$ in Theorem 3.1 is actually more an assumption on ε^* than on τ as we will see later when we apply Theorems 3.1 and 3.2.

Roughly speaking, Theorem 3.1 shows that Algorithm 1 correctly detects the next split ρ^* above the start level ρ_0 whenever there is such a split, while Theorem 3.2 shows that Algorithm 1 also correctly identifies situations, in which there is no split above the start level ρ_0.

Now assume that the assumptions of Theorem 3.1 are satisfied and that Algorithm 1 returned ρ_{out} and the cluster estimates B_1, B_2. Our goal is to apply Algorithm 1 on the two detected clusters B_1 and B_2 separately in a recursive fashion, see Algorithm 2. To this end, we define the new level set estimates

$$L_{i,\rho} := L_\rho \cap B_i\,, \qquad\qquad i = 1, 2,\ \rho \geq \rho_{\mathrm{out}},$$

and let the Algorithm 1 run on both families of level set estimates separately. Of course, we want to use our insights into Algorithm 1 as much as possible. For this reason, we need to replace (7) by a suitable new horizontal and vertical control.

To find such a new control, let us assume that *Assumption M* is satisfied and that we have fixed a $\rho^\dagger \in (\rho^*, \rho^{**}]$. Moreover, let $A_{1,\rho^\dagger}$, and $A_{2,\rho^\dagger}$ be the two connected components of $M_{\rho^\dagger}$, i.e. $\mathcal{C}(M_{\rho^\dagger}) = \{A_{1,\rho^\dagger}, A_{2,\rho^\dagger}\}$. For $i = 1, 2$ we then define two new probability measures P_1 and P_2 by

$$P_i(B) := \frac{P(B \cap A_{i,\rho^\dagger})}{P(A_{i,\rho^\dagger})} \tag{10}$$

for all measurable sets $B \subset X$. Moreover, for $\rho \geq 0$ we denote the level sets of P_1 and P_2 by $M_{1,\rho}$ and $M_{2,\rho}$, respectively. With the help of these notations we can now introduce distributions having a finite split tree.

Definition 3.3. *Let* P *be a distribution satisfying* Assumption P *and* $|\mathcal{C}(M_\rho)| < \infty$ *for all* $\rho \geq 0$*. Moreover, assume that there is a* $\rho_{\max} > 0$ *such that* $M_\rho = \emptyset$ *for all* $\rho \geq \rho_{\max}$*. Then we say that* P *has a finite split-tree with minimal step size* $\epsilon > 0$*, if one of the following two conditions are satisfied:*

i) P satisfies Assumption S.

ii) P satisfies Assumption M *with* $\rho^{**} - \rho^* \geq \epsilon$*, and for* $\rho^\dagger := (\rho^{**} + \rho^*)$ *the two measures* P_1 *and* P_2 *defined by* (10) *have a finite split-tree with minimal step size* $\epsilon > 0$.

Our next goal is to show that Algorithm 2 can be used to estimate the split-tree for distributions having a finite split-tree with some unknown minimal step size $\epsilon > 0$. To this end, we need Theorem 3.4 below, which in its formulation requires the sets

$$\widehat{\mathcal{C}}_\tau(L_\rho) := \left\{B \in \mathcal{C}_\tau(L_\rho) : B \cap L_{\rho+2\varepsilon} \neq \emptyset\right\}, \qquad \rho \geq 0.$$

Note that this set consists of exactly those τ-connected components of L_ρ that are identified in Lines 4 and 7 of Algorithm 1.

Theorem 3.4. *Let* Assumption M *be satisfied. Furthermore, let* $\varepsilon^* \leq (\rho^{**} - \rho^*)/16$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi(\delta), \tau^*(\varepsilon^*)]$*, and* $\varepsilon \in (0, \varepsilon^*]$*, and* $\rho_0 \leq \rho^*$*. In addition, let* $(L_\rho)_{\rho\geq 0}$ *be a decreasing family satisfying* (7) *for all* $\rho \geq \rho_0$*. Finally, let* ρ_{out} *be the estimate of* ρ^* *and* B_1, B_2 *be the cluster estimates returned by Algorithm 1. Then the following statements are true:*

i) We have $|\mathcal{C}_\tau(M^{-\delta}_{\rho^{**}})| = 2$ *and the sets* $V_1 := A^{-\delta}_{1,\rho^{**}}$ *and* $V_2 := A^{-\delta}_{2,\rho^{**}}$ *are the two* τ*-connected components of* $M^{-\delta}_{\rho^{**}}$.

ii) For all $\rho \in [\rho_{\text{out}}, \rho^{**} - 3\varepsilon]$ *we have* $|\widehat{\mathcal{C}}_\tau(L_\rho)| = 2$*. Moreover, we can order the two elements* B_1^ρ *and* B_2^ρ *of* $\widehat{\mathcal{C}}_\tau(L_\rho)$ *such that*

$$V_i \subset B_i^\rho \subset B_i\,, \qquad i = 1, 2\,. \tag{11}$$

iii) If $\rho^\dagger \in [\rho^* + \varepsilon^* + 6\varepsilon, \rho^{**} - 5\varepsilon]$*, then for all* $\rho \geq \rho^\dagger + 4\varepsilon$ *we have* $L_{i,\rho} \subset B_i^{\rho^\dagger+2\varepsilon}$ *and*

$$M^{-\delta}_{i,\rho+\varepsilon} \subset L_{i,\rho} \subset M^{+\delta}_{i,\rho-\varepsilon}. \tag{12}$$

To illustrate Theorem 3.4, we now define $\rho^\dagger := (\rho^{**} + \rho^*)/2$ and assume $\varepsilon^* \leq (\rho^{**} - \rho^*)/16$. For $\varepsilon \in (0, \varepsilon^*]$ we then find $\rho^\dagger \in [\rho^* + \varepsilon^* + 6\varepsilon, \rho^{**} - 5\varepsilon]$ and

$$\rho^\dagger + 4\varepsilon \leq \frac{\rho^{**} + \rho^*}{2} + \frac{\rho^{**} - \rho^*}{4} = \frac{3\rho^{**}}{4} + \frac{\rho^{**}}{4} =: \rho^{\dagger\dagger}\,.$$

Part *iii)* of Theorem 3.4 thus shows (12) for all $\rho \geq \rho^{\dagger\dagger}$. Consequently, Theorems 3.1 and 3.2 can be applied to Algorithm 1 when working with the level sets $(L_{i,\rho})_{\rho\geq\rho^{\dagger\dagger}}$ for the distribution P_i. In addition, $\rho^{**} - 3\varepsilon > \rho^{\dagger\dagger}$ together with part *ii)* shows that for all $\rho \in [\rho_{\text{out}}, \rho^{\dagger\dagger}]$ we have (11), and therefore Algorithm 1 when working with the level sets $(L_{i,\rho})_{\rho\in[\rho_{\text{out}},\rho^{\dagger\dagger}]}$ does only identify one connected component in its Line 3. In other words, the loop between its Lines 2 and 5 is not left for such ρ. Together, these considerations show that Algorithm 2 can be recursively analyzed with the help of Theorems 3.1 and 3.2 to show that Algorithm 2 indeed estimates the split-tree for all distributions P having a finite split-tree with some unknown minimal step size $\epsilon > 0$.

4 Uncertainty Control for Kernel Density Estimators

The results of Section 3 provide guarantees for Algorithm 1 as soon as the input level sets satisfy (7). In [16] it has been shown that guarantees of the form (7) can be established for the level sets of histogram-based density estimators. The goal of this section is to show that (7) can also be established for a variety of kernel density estimators.

Our first definition introduces the kernels we are considering in the following.

Definition 4.1. *A bounded, measurable function $K : \mathbb{R}^d \to [0, \infty)$ is called symmetric kernel, if $K(x) > 0$ in some neighborhood of 0, $K(x) = K(-x)$ for all $x \in \mathbb{R}^d$, and*

$$\int_{\mathbb{R}^d} K(x)\,\mathrm{d}\lambda^d(x) = 1\,. \tag{13}$$

For $\delta > 0$ we write $K_\delta := \delta^{-d} K(\delta^{-1}\,\cdot)$, and for $r > 0$ and a norm $\|\cdot\|$ on $\mathbb{R}^d$ we define

$$\kappa_1(r) := \int_{\mathbb{R}^d \setminus B(0,r)} K(x)\ \mathrm{d}\lambda^d(x)\,, \qquad\qquad \kappa_\infty(r) := \sup_{x \in \mathbb{R}^d \setminus B(0,r)} K(x)\,.$$

Moreover, $\kappa_1(\cdot)$ and $\kappa_\infty(\cdot)$ are called tail functions. Finally, we say that K has:

i) a bounded support if $\operatorname{supp} K \subset B_{\|\cdot\|}$.

ii) an exponential tail behavior, if there exists a constant $c > 0$ such that

$$K(x) \le c \exp\bigl(-\|x\|_2\bigr)\,, \qquad\qquad x \in \mathbb{R}^d. \tag{14}$$

Note that kernels of the form $K(x) = k(\|x\|)$ are always symmetric and if the representing function $k : [0, \infty) \to [0, \infty)$ is bounded and measurable, so is K. Moreover, if $k(r) > 0$ for all $r \in [0, \epsilon)$, where $\epsilon > 0$ is some constant, then we further have $K(x) > 0$ in some neighborhood of 0. The integrability condition (13) is standard for kernel density estimators, and for kernels of the form $K(x) = k(\|x\|)$ it can be translated into a condition on k. In particular, for $k = c\mathbf{1}_{[0,1]}$ we obtain the "rectangular window kernel", which is a symmetric kernel with bounded support in the sense of Definition 4.1 and if k is of the form $k(r) = c\exp(-r^2)$ or $k(r) = c\exp(-r)$, then we obtain a symmetric kernel with exponential tail behavior. Examples of the latter are Gaussian kernels, while the triangular, the Epanechnikov, the quartic, the triweight, and the tricube kernels are further examples of symmetric kernels with bounded support. Finally note that each symmetric kernel with bounded support also has exponential tail behavior, since we always assume that K is bounded.

Before we proceed with our main goal of establishing (7) let us briefly discuss a couple of simple properties of symmetric kernels K in the sense of Definition 4.1. To this end, we first note that the properties of the Lebesgue measure λ^d ensure that

$$\int_{\mathbb{R}^d} K_\delta(x-y)\,\mathrm{d}\lambda^d(y) = \int_{\mathbb{R}^d} K(x-y)\,\mathrm{d}\lambda^d(y) = \int_{\mathbb{R}^d} K(y-x)\,\mathrm{d}\lambda^d(y) = 1 \tag{15}$$

for all $x \in \mathbb{R}^d$, $\delta > 0$, and then by an analogous calculation we obtain

$$\int_{\mathbb{R}^d \setminus B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) = \int_{\mathbb{R}^d \setminus B(0,\sigma/\delta)} K(y)\ \mathrm{d}\lambda^d(y) = \kappa_1(\tfrac{\sigma}{\delta})\,. \tag{16}$$

In addition, we always have $\kappa_1(r) \to 0$ for $r \to \infty$ and if K has bounded support, then the tail functions with respect to this norm satisfy

$$\kappa_1(r) = \kappa_\infty(r) = 0\,, \qquad\qquad r \geq 1\,. \tag{17}$$

The following lemma considers the behavior of the tail functions for kernels with exponential tail behavior.

Lemma 4.2. *Let $K : \mathbb{R}^d \to [0, \infty)$ be a symmetric, kernel with exponential tail behavior* (14)*. Then, for all $r \geq 0$, the corresponding tail functions satisfy*

$$\begin{aligned}\kappa_1(r) &\leq cd^2 \operatorname{vol_d} e^{-r} r^{d-1}\\ \kappa_\infty(r) &\leq ce^{-r}\,.\end{aligned}$$

Now, let K be a symmetric kernel on $\mathbb{R}^d$ and P be a distribution on $\mathbb{R}^d$. For $\delta > 0$ we then define the infinite-sample kernel density estimator $h_{P,\delta} : \mathbb{R}^d \to [0, \infty)$ by

$$h_{P,\delta}(x) := \delta^{-d} \int_{\mathbb{R}^d} K\Big(\frac{x-y}{\delta}\Big)\,\mathrm{d}P(y) \qquad\qquad x \in \mathbb{R}^d.$$

It is easy to see that $h_{P,\delta}$ is a bounded measurable function with $\|h_{P,\delta}\|_\infty \leq \delta^{-d}\|K\|_\infty$. Moreover, a quick application of Tonelli's theorem together with (15) yields

$$\int_{\mathbb{R}^d} h_{P,\delta}\,\mathrm{d}\lambda^d = \int_{\mathbb{R}^d} \delta^{-d} \int_{\mathbb{R}^d} K\Big(\frac{x-y}{\delta}\Big)\,\mathrm{d}\lambda^d(x)\,\mathrm{d}P(y) = \int_{\mathbb{R}^d} \mathbf{1}_{\mathbb{R}^d}(y)\,\mathrm{d}P(y) = 1\,,$$

and hence $h_{P,\delta}$ is a Lebesgue probability density. Moreover, if P has a Lebesgue density h, then it is well-known, see e.g. [4, Theorem 9.1], that $\|h_{P,\delta} - h\|_{L_1(\lambda^d)} \to 0$ for $\delta \to 0$. In addition, if this density is bounded, then (15) yields

$$\begin{aligned}\|h_{P,\delta}\|_\infty &= \sup_{x\in\mathbb{R}^d} \delta^{-d} \int_{\mathbb{R}^d} K\Big(\frac{x-y}{\delta}\Big) h(y)\,\mathrm{d}\lambda^d(y)\\ &\leq \|h\|_\infty \sup_{x\in\mathbb{R}^d} \int_{\mathbb{R}^d} K_\delta(x-y)\,\mathrm{d}\lambda^d(y) = \|h\|_\infty\,.\end{aligned} \tag{18}$$

Clearly, if $D = (x_1, \ldots, x_n) \in X^n$ is a data set, we can consider the corresponding empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, where δ_x denotes the Dirac measure at x. In a slight abuse of notation we also denote this empirical measure by D. The resulting function $h_{D,\delta} : \mathbb{R}^d \to \mathbb{R}$, called kernel density estimator (KDE), can then be computed by

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

Now, one way to define level set estimates with the help of $h_{D,\delta}$ is a simple plug-in approach, that is

$$L_{D,\rho} := \{h_{D,\delta} \geq \rho\}\,. \tag{19}$$

One can show that from a theoretical perspective, this level set estimator is perfectly fine. Unfortunately, however, it is computationally intractable. For example, if $h_{D,\delta}$ is a moving window estimator, that is $K(x) = c\mathbf{1}_{[0,1]}(\|x\|)$ for $x \in \mathbb{R}^d$, then the up to 2^n different level sets (19) are generated by intersection of balls around the samples, and the structure of these intersections may be too complicated to compute τ-connected components in Algorithm 1. For this reason, we consider level set estimates of the form

$$L_{D,\rho} := \{x \in D : h_{D,\delta}(x) \geq \rho\}^{+\sigma}, \tag{20}$$

where $\sigma > 0$. Note that computing connected components of (20) is indeed feasible, since it amounts to computing the connected components of the neighborhood graph, in which two vertices x_i and x_j with $i \neq j$ have an edge if $\|x_i - x_j\| \leq \sigma + \tau$. In particular, DBSCAN can be viewed as such a strategy for the moving window kernel.

With these preparations we can now present our first result that establishes a sort of uncertainty control (7) for level set estimates of the form (20).

Theorem 4.3. *Let $\|\cdot\|$ be some norm on $\mathbb{R}^d$, $K : \mathbb{R}^d \to [0,\infty)$ be a symmetric kernel, and $\kappa_1(\cdot)$ and $\kappa_\infty(\cdot)$ be its associated tail functions. Moreover, let P be a distribution for which* Assumption P *is satisfied, and D be a data set such that the corresponding KDE satisfies $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon$ for some $\varepsilon > 0$ and $\delta > 0$. For $\rho \geq 0$ and $\sigma > 0$ we define*

$$L_{D,\rho} := \{x \in D : h_{D,\delta}(x) \geq \rho\}^{+\sigma}$$

and $\epsilon := \max\{\rho\kappa_1(\frac{\sigma}{\delta}), \delta^{-d}\kappa_\infty(\frac{\sigma}{\delta})\}$. Then, for all $\rho \geq \delta^{-d}\kappa_\infty(\frac{\sigma}{\delta})$, we have

$$M^{-2\sigma}_{\rho+\varepsilon+\epsilon} \subset L_{D,\rho} \subset M^{+2\sigma}_{\rho-\varepsilon-\epsilon}. \tag{21}$$

Moreover, if P has a bounded density h, then (21) also holds for $\epsilon = \|h\|_\infty \kappa_1(\frac{\sigma}{\delta})$.

Note that for kernels K having bounded support for the norm considered in Theorem 4.3, Equation (17) shows that (21) actually holds for $\epsilon = 0$ and all $\rho \geq 0$ and all $\sigma \geq \delta$. Therefore, we have indeed (7) for δ replaced by 2σ. In general, however, we have an additional horizontal uncertainty ϵ that of course affects the guarantees of Theorem 3.1. To control this influence, our strategy will be to ensure that $\epsilon \leq \varepsilon$, which in view of $\epsilon = \|h\|_\infty \kappa_1(\frac{\sigma}{\delta})$ means that we need to have an upper bound on $\kappa_1(\cdot)$ and σ.

Theorem 4.3 tells us that the uncertainty control (21) is satisfied as soon as we have a data set D with $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon$. Therefore, our next goal is to establish such an estimate with high probability. Before we begin we like to mention that rates for $\|h_{D,\delta} - h_{P,\delta}\|_\infty \to 0$ have already been proven in [6]. However, those rates only hold for $n \geq n_0$, where n_0, although it almost surely exists, may actually depend on the data set D. In addition one is required to choose a sequence (δ_n) of bandwidths a-priori. To apply the theory developed in [16] including the adaptivity, however, we need bounds of the form $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon(\delta, n, \varsigma)$ that hold with probability not smaller than $1 - e^{-\varsigma}$. For these reasons, the results of [6] are not suitable for our purposes.

To establish the bounds described above, we need to recall some notions first.

Definition 4.4. *Let E be a Banach space and $A \subset E$ be a bounded subset. Then, for all $\varepsilon > 0$, the covering numbers of A are defined by*

$$\mathcal{N}(A, \|\cdot\|_E, \varepsilon) := \inf\left\{n \geq 1 : \exists\, x_1, \ldots, x_n \in E \text{ such that } A \subset \bigcup_{i=1}^{n} (x_i + \varepsilon B_{\|\cdot\|})\right\},$$

where $\inf \emptyset := \infty$. *Furthermore, we use the notation $\mathcal{N}(A, E, \varepsilon) := \mathcal{N}(A, \|\cdot\|_E, \varepsilon)$.*

We now introduce the kind of covering number bound we will use in our analysis.

Definition 4.5. *Let (Z, P) be a probability space and $\mathcal{G}$ be a set of bounded measurable functions from Z to $\mathbb{R}$ for which there exists a $B > 0$ such that $\|g\|_\infty \le B$ for all $g \in \mathcal{G}$. Then $\mathcal{G}$ is called a uniformly bounded VC-class, if there exist $A > 0$ and $\nu > 0$ such that, for every probability measure P on Z and every $0 < \epsilon \le B$, the covering numbers satisfy*

$$\mathcal{N}(\mathcal{G}, L_2(P), \epsilon) \le \left(\frac{AB}{\epsilon}\right)^\nu . \tag{22}$$

Before we proceed, let us briefly look at two important sufficient criteria for ensuring that the set of functions

$$\mathcal{K}_\delta := \big\{K_\delta(x - \cdot) : x \in X\big\} \tag{23}$$

is a uniformly bounded VC-class. The first result in this direction considers kernels used in moving window estimates.

Lemma 4.6. *Consider the kernel $K = c\mathbf{1}_{B_{\|\cdot\|}}$, where $\|\cdot\|$ is either the Euclidean- or the supremum norm. Then for all $\delta > 0$ the set $\mathcal{K}_\delta$ defined by (23) is a uniformly bounded VC-class with $B := \delta^{-d}\|K\|_\infty = \delta^{-d}c$ and A and ν being independent of δ.*

The next lemma shows that Hölder continuous kernels also induce a uniformly bounded VC-class $\mathcal{K}_\delta$, provided that the input space X in (23) is compact. For its formulation we need to recall that for every norm $\|\cdot\|$ on $\mathbb{R}^d$ and every compact subset $X \subset \mathbb{R}^d$ there exists a finite constant $C_{\|\cdot\|}(X) > 0$ such that for all $0 < \varepsilon \le \operatorname{diam}_{\|\cdot\|}(X)$ we have

$$\mathcal{N}(X, \|\cdot\|, \epsilon) \le C_{\|\cdot\|}(X)\epsilon^{-d} . \tag{24}$$

We can now formulate the announced result for Hölder-continuous kernels.

Lemma 4.7. *Let $K : \mathbb{R}^d \to [0, \infty)$ be a symmetric kernel that is α-Hölder continuous. For some arbitrary but fixed norm $\|\cdot\|$ on $\mathbb{R}^d$ we write $|K|_\alpha$ for the corresponding α-Hölder constant. Moreover, let $X \subset \mathbb{R}^d$ be a compact subset and $\mathcal{K}_\delta$ defined by (23). Then for all $\delta > 0$ with $\delta \le \big(\frac{|K|_\alpha}{\|K\|_\infty}\big)^{1/\alpha} \operatorname{diam}_{\|\cdot\|}(X)$, all $0 < \epsilon \le \delta^{-d}\|K\|_\infty$, and all distributions P on $\mathbb{R}^d$ we have*

$$\mathcal{N}\big(\mathcal{K}_\delta, L_2(P), \epsilon\big) \le C_{\|\cdot\|}(X)\left(\frac{|K|_\alpha}{\delta^{\alpha+d}\epsilon}\right)^{d/\alpha} . \tag{25}$$

In particular, $\mathcal{K}_\delta$ is a uniformly bounded VC-class with $\nu := d/\alpha$, $A := (C_{\|\cdot\|}(X))^{\alpha/d}|K|_\alpha\|K\|_\infty^{-1}\delta^{-\alpha}$ and $B := \delta^{-d}\|K\|_\infty$.

Now that we have collected sufficiently many examples of kernels for which $\mathcal{K}_\delta$ is a uniformly bounded VC-class, we can now present the second main result of this section that establishes a finite sample bound $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon(\delta, n, \varsigma)$.

Theorem 4.8. *Let $X \subset \mathbb{R}^d$ and P be distribution on X that has a Lebesgue density $h \in L_1(\mathbb{R}^d) \cap L_p(\mathbb{R}^d)$ for some $p \in (1, \infty]$. We write $\frac{1}{p} + \frac{1}{p'} = 1$. Moreover, let $K : \mathbb{R}^d \to [0, \infty)$ be a symmetric kernel for which there is a $\delta_0 \in (0, 1]$ such that for all $\delta \in (0, \delta_0]$ the set $\mathcal{K}_\delta$ defined in (23) is a uniformly bounded VC-class with constants of the form $B_\delta = \delta^{-d}\|K\|_\infty$, $A_\delta = A_0\delta^{-a}$, and $A_0 > 0, a \ge 0, \nu \ge 1$ being independent of δ, that is,*

$$\mathcal{N}\big(\mathcal{K}_\delta, L_2(Q), \epsilon\big) \le \left(\frac{A_0\|K\|_\infty\delta^{-(d+a)}}{\epsilon}\right)^\nu \tag{26}$$

holds for all $\delta \in (0, \delta_0]$, all $\epsilon \in (0, B_\delta]$, and all distributions Q on $\mathbb{R}^d$. Then, there exists a $C > 0$ only depending on d, p, and K such that, for all $n \geq 1$, $\delta > 0$, and $\varsigma \geq 1$ satisfying

$$\delta \leq \min\left\{\delta_0, \frac{4^{p'}\|K\|_\infty}{\|h\|_p^{p'}}, \frac{\|h\|_p^{\frac{1}{2a+d/p'}}}{C}\right\} \qquad \text{and} \qquad \frac{|\log \delta|}{n\delta^{d/p'}} \leq \frac{\|h\|_p}{C\varsigma} \tag{27}$$

we have

$$P^n\left(\left\{D : \|h_{D,\delta} - h_{P,\delta}\|_{\ell_\infty(X)} < C\sqrt{\frac{\|h\|_p\,|\log \delta|\,\varsigma}{n\delta^{d(1+1/p)}}}\right\}\right) \geq 1 - e^{-\varsigma}. \tag{28}$$

Theorem 4.8 recovers the same rates as [6], but not in almost sure asymptotic form but in form of a finite sample bound. Moreover, unlike [6], Theorem 4.8 also yields rates for unbounded densities.

5 Statistical Analysis of KDE-based Clustering

In this section we combine the generic results of Section 3 with the uncertainty control for level set estimates obtained from kernel density estimates we obtained in Section 4. As a result we will present finite sample guarantees, consistency results, and rates for estimating ρ^* and the clusters.

Our first result presents finite sample bounds for estimating both ρ^* and the single or multiple clusters with the help of Algorithm 1. To treat kernels with bounded and unbounded support simultaneously, we restrict ourselves to the case of bounded densities, but at least for kernels with bounded support an adaption to p-integrable densities is straightforward as we discuss below.

Theorem 5.1. *Let P be a distribution for which* Assumption P *is satisfied and whose Lebesgue density is bounded. Moreover, consider a symmetric kernel $K : \mathbb{R}^d \to [0, \infty)$ with exponential tail behavior, for which the assumptions of Theorem 4.8 hold. For fixed $\delta \in (0, \mathrm{e}^{-1}]$ and $\tau > 0$, we choose a $\sigma > 0$ with*

$$\sigma \geq \begin{cases} \delta & \text{if } \operatorname{supp} K \subset B_{\|\cdot\|}\,, \\ \delta\,|\log \delta|^2 & \text{otherwise.} \end{cases} \tag{29}$$

and assume this σ further satisfies both $\sigma \leq \delta_{\mathrm{thick}}/2$ and $\tau > \psi(2\sigma)$. Moreover, for fixed $\varsigma \geq 1$, $n \geq 1$ satisfying the assumptions (27), we pick an $\varepsilon > 0$ satisfying the bound

$$\varepsilon \geq \frac{C}{2}\sqrt{\frac{\|h\|_\infty\,|\log \delta|\,\varsigma}{n\delta^d}}\,, \tag{30}$$

and if K does not have bounded support, also

$$\varepsilon \geq \max\{1, 2d^2 \operatorname{vol_d}\} \cdot c \cdot \delta^{|\log \delta| - d}\,. \tag{31}$$

Now assume that for each data set $D \in X^n$ sampled from P^n, we feed Algorithm 1 with the level set estimators $(L_{D,\rho})_{\rho \geq 0}$ given by (20), the parameters τ and ε, and a start level $\rho_0 \geq \varepsilon$. Then the following statements are true:

i) If P satisfies Assumption S *and $\rho_0 \geq \rho_*$, then with probability P^n not less than $1 - e^{-\varsigma}$ Algorithm 1 returns ρ_0 and L_0 and we have*

$$\mu\big(L_{\rho_0} \triangle \hat{M}_{\rho_*}\big) \leq \mu\big(M^{+2\sigma}_{\rho_0-\varepsilon} \setminus \hat{M}_{\rho_*}\big) + \mu\big(\hat{M}_{\rho_*} \setminus M^{-2\sigma}_{\rho_0+\varepsilon}\big)\,, \tag{32}$$

where $\hat{M}_{\rho_*} := \bigcup_{\rho > \rho_*} M_\rho$.

ii) If P satisfies Assumption M *and we have an*

$$\varepsilon^* \geq \varepsilon + \inf\bigl\{\varepsilon' \in (0, \rho^{**} - \rho^*] : \tau^*(\varepsilon') \geq \tau\bigr\}\,. \tag{33}$$

with $9\varepsilon^* \leq \rho^{**} - \rho^*$, then with probability P^n not less than $1 - e^{-\varsigma}$, we have a $D \in X^n$ such that the following statements are true for Algorithm 1:

(a) The returned level $\rho_{D,\mathrm{out}}$ satisfies both $\rho_{D,\mathrm{out}} \in [\rho^* + 2\varepsilon, \rho^* + \varepsilon^* + 5\varepsilon]$ and

$$\tau - \psi(2\sigma) < 3\tau^*\bigl(\rho_{D,\mathrm{out}} - \rho^* + \varepsilon\bigr)\,.$$

(b) Two sets $B_1(D)$ and $B_2(D)$ are returned and these can be ordered such that for $A^i_{\rho_{D,\mathrm{out}}+\varepsilon} \in \mathcal{C}(M_{\rho_{D,\mathrm{out}}+\varepsilon})$ ordered in the sense of $A^i_{\rho_{D,\mathrm{out}}+\varepsilon} \subset A^*_i$ we have

$$\sum_{i=1}^{2} \mu\bigl(B_i(D) \vartriangle A^*_i\bigr) \leq 2\sum_{i=1}^{2} \mu\bigl(A^*_i \backslash (A^i_{\rho_{D,\mathrm{out}}+\varepsilon})^{-2\sigma}\bigr) + \mu\bigl(M^{+2\sigma}_{\rho_{D,\mathrm{out}}-\varepsilon} \backslash \{h > \rho^*\}\bigr)\,. \tag{34}$$

For our subsequent asymptotic analysis we note that the assumptions $\delta \in (0, \mathrm{e}^{-1}]$ and $\varsigma \geq 1$ of Theorem 5.1 show that (31) is satisfied if

$$\max\bigl\{1, 2d^2 \,\mathrm{vol_d}\bigr\} \cdot c \cdot \delta^{|\log\delta| + d/2} \leq \frac{C}{2}\sqrt{\frac{\|h\|_\infty}{n}}\,, \tag{35}$$

and if we choose δ in terms of n, i.e., $\delta = \delta_n$, then (35) is satisfied for large n if $\delta_n \in O(n^{-a})$ for some small $a > 0$. We shall see below, that such rates for δ_n are typical.

If we have a kernel with bounded support, then a variant of Theorem 5.1 also holds for unbounded densities. Indeed, if we have a density $h \in L_1(\mathbb{R}^d) \cap L_p(\mathbb{R}^d)$ for some $p \in (1, \infty]$, then all conclusions of Theorem 5.1 remain valid, if we replace (30) by

$$\varepsilon \geq \frac{C}{2}\sqrt{\frac{\|h\|_p\, |\log\delta|\, \varsigma}{n\delta^{d(1+1/p)}}}\,.$$

Note that for such kernels the additional assumption (31) is not necessary.

While (29) only provides a lower bound on possible values for σ, Theorem 5.1 actually indicates that σ should not be chosen significantly larger than these lower bounds, either. Indeed, the choice of σ also implies a minimal value for τ by the condition $\tau > \psi(2\sigma)$, which in turn influences ε^* by (33). Namely, larger values of σ lead to larger τ-values and therefore to larger values for ε^*. As a result, the guarantees in *(a)* become weaker, and in addition, larger values of σ also lead to weaker guarantees in *(b)*. For a similar reason we do not consider kernels K with heavier tails than (14). Indeed, if K only has a polynomial upper bound for its tail, i.e., there are constants c and $\alpha > d$ with

$$K(x) \leq c \cdot \|x\|_2^{-\alpha}\,, \qquad\qquad x \in \mathbb{R}^d\,,$$

then $\kappa_1(r) \preceq r^{-\alpha+d}$ and $\kappa_\infty(r) \preceq r^{-\alpha}$. Now, if we picked $\sigma = \delta|\log\delta|^b$ for some $b > 0$, then we would need to replace (31) by a bound of the form $\tilde{c}\delta^{-d}|\log\delta|^{-\alpha b} \leq \varepsilon$, and this would rule out $\varepsilon \to 0$ for $\delta \to 0$. As a result, no rates would be possible. Now, one could address this by choosing $\sigma := \delta^b$ for some $b \in (0,1)$, which in turn would require a bound of the form $\tilde{c} \cdot \delta^{\alpha(1-b)-d} \leq \varepsilon$, instead of (31). Arguing as around (35) this is guaranteed if

$$\tilde{c}\delta^{\alpha(1-b)-d/2} \leq \frac{C}{2}\sqrt{\frac{\|h\|_\infty}{n}}\,,$$

and if $\delta \to 0$ the latter would actually require $b < 1 - \frac{d}{2\alpha}$. In particular, b would be strictly bounded away from 1. However, such a choice for σ would significantly weaken the guarantees given in *(a)* and *(b)* as explained above, and as a consequence, the rates obtained below would be worse. Note that from a high-level perspective this phenomenon is not surprising: indeed, heavier tails smooth out the infinite sample density estimator $h_{P,\delta}$ and as consequence, the uncertainty guarantees (21) become worse in the horizontal direction, that is, we get more blurry estimates $L_{D,\rho}$ of M_ρ. However, for the detection of connected components at a level ρ, less blurry estimates are preferable.

In the remainder of this section, we illustrate how the finite sample guarantee of Theorem 5.1 can be used to derive both consistency and rates. To deal with kernels with unbounded support we restrict our considerations to the case of bounded densities, but it is straightforward to obtain results for unbounded densities if one restricts considerations to kernels with bounded support as already indicated above.

Corollary 5.2. *Let P be a distribution satisfying* Assumption P *and whose Lebesgue density is bounded. Moreover, consider a symmetric kernel $K : \mathbb{R}^d \to [0,\infty)$ with exponential tail behavior, for which the assumptions of Theorem 4.8 hold. Let (δ_n) be a positive sequence with $\delta_n \preceq n^{-a}$ for some $a > 0$ and pick a sequence (σ_n) converging to zero and satisfying* (29) *for all sufficiently large n. Moreover, let (ε_n) and (τ_n) be positive sequences converging to zero such that $\psi(2\sigma_n) < \tau_n$ for all sufficiently large n, and*

$$\lim_{n\to\infty} \frac{\log \delta_n^{-1}}{n\varepsilon_n^2\delta_n^d} = 0\,. \tag{36}$$

Now assume that for each data set $D \in X^n$ sampled from P^n, we feed Algorithm 1 with the level set estimators $(L_{D,\rho})_{\rho\geq 0}$ given by (20)*, the parameters τ_n and ε_n, and the start level $\rho_0 := \varepsilon_n$. Then the following statements are true:*

i) If P satisfies Assumption S *with $\rho_* = 0$, then for all $\epsilon > 0$ we have*

$$\lim_{n\to\infty} P^n\Big(\big\{D \in X^n : 0 < \rho_{D,\mathrm{out}} \leq \epsilon\big\}\Big) = 1\,,$$

and if $\mu(\overline{\{h > 0\}} \setminus \{h > 0\}) = 0$ we also have

$$\lim_{n\to\infty} P^n\Big(\big\{D \in X^n : \mu(L_{D,\rho_{D,\mathrm{out}}} \triangle \{h > 0\}) \leq \epsilon\big\}\Big) = 1\,,$$

ii) If P satisfies Assumption M*, then, for all $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P^n\Big(\big\{D \in X^n : 0 < \rho_D^* - \rho^* \leq \epsilon\big\}\Big) = 1\,,$$

and, if $\mu(\overline{A_i^* \cup A_2^*} \setminus (A_1^* \cup A_2^*)) = 0$, we also have, for $B_1(D)$, $B_2(D)$ as in (34)*:*

$$\lim_{n\to\infty} P^n\Big(\big\{D \in X^n : \mu(B_1(D) \triangle A_1^*) + \mu(B_2(D) \triangle A_2^*) \leq \epsilon\big\}\Big) = 1\,.$$

Our next goal is to establish rates of convergence for estimating ρ^* and the clusters. We begin with a result providing a rate of $\rho_D^* \to \rho^*$. To this end we need to recall the following definition from [16] that describes how well the clusters are separated above ρ^*.

Definition 5.3. *Let* Assumption M *be satisfied. Then the clusters of P have a separation exponent $\kappa \in (0,\infty]$, if there is a constant $\underline{c}_{\mathrm{sep}} > 0$ such that*

$$\tau^*(\varepsilon) \geq \underline{c}_{\mathrm{sep}}\, \varepsilon^{1/\kappa}$$

for all $\varepsilon \in (0, \rho^{**} - \rho^*]$. Moreover, the separation exponent κ is exact, if there exists another constant $\overline{c}_{\mathrm{sep}} > 0$ such that, for all $\varepsilon \in (0, \rho^{**} - \rho^*]$, we have

$$\tau^*(\varepsilon) \leq \overline{c}_{\mathrm{sep}}\, \varepsilon^{1/\kappa}\,.$$

The separation exponent describes how fast the connected components of the M_ρ approach each other for $\rho \searrow \rho^*$. Note that a distribution having separation exponent κ also has separation exponent κ' for all $\kappa' < \kappa$. In particular, the "best" separation exponent is $\kappa = \infty$ and this exponent describes distributions, for which we have $d(A_1^*, A_2^*) \geq \underline{c}_{\mathrm{sep}}$, i.e. the clusters A_1^* and A_2^* do not touch each other.

The separation exponent makes it possible to find a good value for ε^* in Theorem 5.1. Indeed, the proof of Theorem [16, Theorem 4.3] shows that for given ε and τ, the value

$$\varepsilon^* := \varepsilon + (\tau/\underline{c}_{\mathrm{sep}})^\kappa$$

satisfies (33) as soon as we have $9\varepsilon^* \leq \rho^{**} - \rho^*$. Consequently, the bound in part *ii) (a)* of Theorem 5.1 becomes

$$2\varepsilon \leq \rho_{D,\mathrm{out}} - \rho^* \leq 6\varepsilon + \Big(\frac{\tau}{\underline{c}_{\mathrm{sep}}}\Big)^\kappa \tag{37}$$

if we have a separation exponent $\kappa \in (0,\infty]$. Moreover, if the separation exponent $\kappa \in (0,\infty)$ is exact and we choose $\tau \geq 2\psi(2\sigma)$, then (37) can be improved to

$$\varepsilon + \frac{1}{4}\Big(\frac{\tau}{6\overline{c}_{\mathrm{sep}}}\Big)^\kappa \leq \rho_{D,\mathrm{out}} - \rho^* \leq 6\varepsilon + \Big(\frac{\tau}{\underline{c}_{\mathrm{sep}}}\Big)^\kappa \tag{38}$$

as the proof of Theorem [16, Theorem 4.3] shows. In order to establish rates, it therefore suffices to find null sequences (ε_n), (δ_n), (σ_n), and (τ_n) that satisfy (29) and (30), and additionally $\delta_n \in \mathcal{O}(n^{-a})$ for some $a > 0$, if K does not have bounded support. The following corollary presents resulting rates of this approach that are, modulo logarithmic terms, the best ones we can obtain from this approach.

Corollary 5.4. *Let P be a distribution for which* Assumption M *is satisfied and whose Lebesgue density is bounded. Moreover, consider a symmetric kernel $K : \mathbb{R}^d \to [0,\infty)$ with exponential tail behavior, for which the assumptions of Theorem 4.8 hold. In addition, assume that the clusters of P have separation exponent $\kappa \in (0,\infty)$. Furthermore, let (ε_n), (δ_n), (σ_n), and (τ_n) be sequences with*

$$\varepsilon_n \sim \Big(\frac{(\log n)^3 \cdot \log\log n}{n}\Big)^{\frac{\gamma\kappa}{2\gamma\kappa+d}}, \qquad \delta_n \sim \Big(\frac{\log n}{n}\Big)^{\frac{1}{2\gamma\kappa+d}},$$

$$\sigma_n \sim \Big(\frac{(\log n)^3}{n}\Big)^{\frac{1}{2\gamma\kappa+d}}, \qquad \tau_n \sim \Big(\frac{(\log n)^3 \cdot \log\log n}{n}\Big)^{\frac{\gamma}{2\gamma\kappa+d}},$$

and assume that, for $n \geq 1$ $D \in X^n$ sampled from P^n, we feed Algorithm 1 with the level set estimators $(L_{D,\rho})_{\rho\geq 0}$ given by (20), the parameters τ_n and ε_n, and the start level $\rho_0 := \varepsilon_n$. Then there exists a $\overline{K} \geq 1$ such that for all sufficiently large n we have

$$P^n\Big(\Big\{D \in X^n : \rho_D^* - \rho^* \leq \overline{K}\varepsilon_n\Big\}\Big) \geq 1 - \frac{1}{n}\,. \tag{39}$$

Moreover, if the separation exponent κ is exact, there exists another constant $\underline{K} \geq 1$ such that for all sufficiently large n we have

$$P^n\Big(\in X^n : \underline{K}\varepsilon_n \leq \rho^*_D - \rho^* \leq \overline{K}\varepsilon_n\Big) \geq 1 - \frac{1}{n}\,. \tag{40}$$

Finally, if $\kappa = \infty$ and $\operatorname{supp} K \subset B_{\|\cdot\|}$, *then* (40) *holds for all sufficiently large n, if $\sigma_n = \delta_n$ and*

$$\varepsilon_n \sim \Big(\frac{\log n \cdot \log\log n}{n}\Big)^{\frac{1}{2}}\,,\ \delta_n \sim \big(\log\log n\big)^{-\frac{1}{2d}}\,,\ \textit{and}\ \tau_n \sim \big(\log\log n\big)^{-\frac{\gamma}{3d}}\,.$$

Note that the rates obtained in Corollary 5.4 only differ by the factor $(\log n)^2$ from the rates in [16, Corollary 4.4]. Moreover, if K has a bounded support, then an easy modification of the above corollary yields exactly the same rates as in [16, Corollary 4.4].

Our next goal is to establish rates for $\mu(B_i(D) \vartriangle A^*_i) \to 0$. Since this is a modified level set estimation problem, let us recall some assumptions on P, which have been used in this context. The first assumption in this direction is one-sided variant of a well-known condition introduced by Polonik [11].

Definition 5.5. *Let μ be a finite measure on X and P be a distribution on X that has a μ-density h. For a given level $\rho \geq 0$, we say that P has flatness exponent $\vartheta \in (0, \infty]$, if there exists a constant $c_{\text{flat}} > 0$ such that*

$$\mu\big(\{0 < h - \rho < s\}\big) \leq (c_{\text{flat}}s)^{\vartheta}\,, \qquad\qquad s > 0. \tag{41}$$

The larger ϑ is, the steeper h must approach ρ from above. In particular, for $\vartheta = \infty$, the density h is allowed to take the value ρ but is otherwise bounded away from ρ.

Next, we describe the roughness of the boundary of the clusters.

Definition 5.6. *Let* Assumption M *be satisfied. Given some $\alpha \in (0, 1]$, the clusters have an α-smooth boundary, if there exists a constant $c_{\text{bound}} > 0$ such that, for all $\rho \in (\rho^*, \rho^{**}]$, $\delta \in (0, \delta_{\text{thick}}]$, and $i = 1, 2$, we have*

$$\mu\big((A^i_\rho)^{+\delta} \setminus (A^i_\rho)^{-\delta}\big) \leq c_{\text{bound}}\delta^{\alpha}\,, \tag{42}$$

where A^1_ρ and A^2_ρ denote the two connected components of the level set M_ρ.

In $\mathbb{R}^d$, considering $\alpha > 1$ does not make sense, and for an $A \subset \mathbb{R}^d$ with rectifiable boundary we always have $\alpha = 1$, see [17, Lemma A.10.4].

Assumption R. Assumption M is satisfied and P has a bounded Lebesgue density h. Moreover, P has flatness exponent $\vartheta \in (0, \infty]$ at level ρ^*, its clusters have an α-smooth boundary for some $\alpha \in (0, 1]$, and its clusters have separation exponent $\kappa \in (0, \infty]$.

Corollary 5.7. *Let* Assumption R *be satisfied and K be as in Corollary 5.4. and write $\varrho := \min\{\alpha, \vartheta\gamma\kappa\}$. Furthermore, let (ε_n), (δ_n), and (τ_n) be sequences with*

$$\varepsilon_n \sim \Big(\frac{\log n}{n}\Big)^{\frac{\varrho}{2\varrho+\vartheta d}}(\log\log n)^{-\frac{\vartheta d}{8\varrho+4\vartheta d}}\,, \qquad \delta_n \sim \Big(\frac{\log n \cdot \log\log n}{n}\Big)^{\frac{\vartheta}{2\varrho+\vartheta d}}\,,$$

$$\sigma_n \sim \Big(\frac{(\log n)^3 \cdot \log\log n}{n}\Big)^{\frac{\vartheta}{2\varrho+\vartheta d}}\,, \qquad \tau_n \sim \Big(\frac{(\log n)^3 \cdot (\log\log n)^2}{n}\Big)^{\frac{\vartheta\gamma}{2\varrho+\vartheta d}}.$$

Assume that, for $n \geq 1$, we feed Algorithm 1 as in Corollary 5.4. Then there is a constant $\overline{K} \geq 1$ such that, for all $n \geq 1$ and the ordering as in (34)*, we have*

$$P^n\Big(D : \sum_{i=1}^{2}\mu\big(B_i(D) \vartriangle A^*_i\big) \leq K\Big(\frac{(\log n)^3 \cdot (\log\log n)^2}{n}\Big)^{\frac{\vartheta\varrho}{2\varrho+\vartheta d}}\Big) \geq 1 - \frac{1}{n}\,.$$

Again, the rates obtained in Corollary 5.7 only differ by the factor $(\log n)^2$ from the rates in [16, Corollary 4.8]. Moreover, if K has a bounded support, then an easy modification of Corollary 5.7 again yields exactly the same rates as in [16, Corollary 4.8].

Our final goal is to modify the adaptive parameter selection strategy for the histogram-based clustering algorithm of [16] to our KDE-based clustering algorithm. To this end, let $\Delta \subset (0,1]$ be finite and $n \geq 1$, $\varsigma \geq 1$. For $\delta \in \Delta$, we fix $\sigma_{\delta,n} > 0$ and $\tau_{\delta,n} > 0$ such that (29) and $\tau_{\delta,n} \geq 2\psi(2\sigma_{\delta,n})$ are satisfied. In addition, we define

$$\varepsilon_{\delta,n} := C_u \sqrt{\frac{|\log \delta|(\varsigma + \log|\Delta|)\log\log n}{\delta^d n}} + \max\{1, 2d^2 \operatorname{vol_d}\} \cdot c \cdot \delta^{|\log \delta| - d} \tag{43}$$

where $C_u \geq 1$ is some user-specified constant and the second term can actually be omitted if the used kernel K has bounded support. Now assume that, for each $\delta \in \Delta$, we run Algorithm 1 with the parameters $\varepsilon_{\delta,n}$ and $\tau_{\delta,n}$, with the start level $\rho_0 := \varepsilon_{\delta,n}$, and with the level set estimators $(L_{D,\rho})_{\rho \geq 0}$ given by (20). Let us consider a width $\delta^*_{D,\Delta} \in \Delta$ that achieves the smallest returned level, i.e.

$$\delta^*_{D,\Delta} \in \arg\min_{\delta \in \Delta} \rho_{D,\delta,\mathrm{out}} \,. \tag{44}$$

Note that in general, this width may not be uniquely determined, so that in the following we need to additionally assume that we have a well-defined choice, e.g. the smallest $\delta \in \Delta$ satisfying (44). Moreover, we write

$$\rho^*_{D,\Delta} := \min_{\delta \in \Delta} \rho_{D,\delta,\mathrm{out}} \tag{45}$$

for the smallest returned level. Note that unlike the width $\delta^*_{D,\Delta}$, the level $\rho^*_{D,\Delta}$ is always unique. Finally, we define $\varepsilon_{D,\Delta} := \varepsilon_{\delta^*_{D,\Delta},n}$ and $\tau_{D,\Delta} := \tau_{\delta^*_{D,\Delta},n}$. With these preparation we can now present the following finite sample bound for $\rho^*_{D,\Delta}$.

Theorem 5.8. *Let P be a distribution for which* Assumption M *is satisfied and whose Lebesgue density is bounded. Moreover, consider a symmetric kernel $K : \mathbb{R}^d \to [0,\infty)$ with exponential tail behavior, for which the assumptions of Theorem 4.8 hold. In addition, assume that the two clusters of P have separation exponent $\kappa \in (0,\infty]$. For a fixed finite $\Delta \subset (0, \mathrm{e}^{-1}]$, and $n \geq 1$, $\varsigma \geq 1$, and $C_u \geq 1$, we define $\varepsilon_{\delta,n}$ by (43) and $\sigma_{\delta,n} > 0$ and $\tau_{\delta,n} > 0$ such that (29), $\tau_{\delta,n} \geq 2\psi(2\sigma_{\delta,n})$, and $2\sigma_{\delta,n} \leq \delta_{\mathrm{thick}}$ are satisfied for all $\delta \in \Delta$. Furthermore, assume that $4C_u^2 \log\log n \geq C\|h\|_\infty$, where C is the constant in (28) and $\varepsilon_{\delta,n} + (\tau_{\delta,n}/\underline{c}_{\mathrm{sep}})^\kappa \leq (\rho^{**} - \rho^*)/9$ for all $\delta \in \Delta$. Then we have*

$$P^n\Big(\Big\{D \in X^n : \varepsilon_{D,\Delta} \,<\, \rho^*_{D,\Delta} - \rho^* \,\leq\, \min_{\delta \in \Delta}\big((\tau_{\delta,n}/\underline{c}_{\mathrm{sep}})^\kappa + 6\varepsilon_{\delta,n}\big)\Big\}\Big) \geq 1 - e^{-\varsigma} \,.$$

Moreover, if the separation exponent κ is exact and $\kappa < \infty$, then the assumptions above actually guarantee

$$P^n\Big(D : \min_{\delta \in \Delta}\big(c_1 \tau^\kappa_{\delta,n} + \varepsilon_{\delta,n}\big) \,<\, \rho^*_{D,\Delta} - \rho^* \,\leq\, \min_{\delta \in \Delta}\big(c_2 \tau^\kappa_{\delta,n} + 6\varepsilon_{\delta,n}\big)\Big) \geq 1 - e^{-\varsigma} \,,$$

where $c_1 := \frac{1}{4}(6\overline{c}_{\mathrm{sep}})^{-\kappa}$ and $c_2 := \underline{c}_{\mathrm{sep}}^{-\kappa}$, and similarly

$$P^n\Big(\Big\{D \in X^n : c_1 \tau^\kappa_{D,\Delta} + \varepsilon_{D,\Delta} \,<\, \rho^*_{D,\Delta} - \rho^* \,\leq\, c_2 \tau^\kappa_{D,\Delta} + 6\varepsilon_{D,\Delta}\Big\}\Big) \geq 1 - e^{-\varsigma} \,.$$

To achieve our goal of an adaptive parameter selection strategy, it now suffices in view of Theorem 5.8 to define appropriate Δ, $\sigma_{\delta,n}$, and $\tau_{\delta,n}$. Here we proceed as in [16, Section 5]. Namely, for $n \geq 16$, we consider the interval

$$I_n := \left[\Big(\frac{\log n \cdot (\log\log n)^2}{n}\Big)^{\frac{1}{d}}, \Big(\frac{1}{\log\log n}\Big)^{\frac{1}{d}}\right]$$

and fix some $n^{-1/d}$-net $\Delta_n \subset I_n$ of I_n with $|\Delta_n| \leq n$. Furthermore, for some fixed $C_u \geq 1$ and $n \geq 16$, we define $\sigma_{\delta,n}$ by (29), write $\tau_{\delta,n} := \sigma_{\delta,n}^{\gamma} \log\log\log n$, and define $\varepsilon_{\delta,n}$ by (43) for all $\delta \in \Delta_n$ and $\varsigma = \log n$. Following the ideas of the proofs of [16, Corollaries 5.2 and 5.3] we then obtain a constant $\overline{K}$ such that for all sufficiently large $n \geq 16$ we have

$$P^n\biggl(D : \rho^*_{D,\Delta_n} - \rho^* \leq \overline{K}\Bigl(\frac{(\log n)^3 \cdot (\log\log n)^2}{n}\Bigr)^{\frac{\gamma\kappa}{2\gamma\kappa+d}}\biggr) \geq 1 - \frac{1}{n}\,. \tag{46}$$

Here, (46) holds if P has separation exponent $\kappa \in (0,\infty)$, and if the kernel K has bounded support, it remains true for $\kappa = \infty$. In addition, the upper bound in (46) can be matched by a lower bound that only differs by a double logarithmic factor provided that the separation exponent $\kappa \in (0,\infty)$ is exact. Finally, if Assumption R is satisfied, we further find

$$P^n\biggl(D : \sum_{i=1}^{2} \mu(B_i(D) \triangle A_i^*) \leq \hat{K}\Bigl(\frac{(\log n)^3 \cdot (\log\log n)^2}{n}\Bigr)^{\frac{\vartheta\gamma\kappa}{2\gamma\kappa+\vartheta d}}\biggr) \geq 1 - \frac{1}{n}\,,$$

for all sufficiently large $n \geq 16$, where $\hat{K}$ is another constant independent of n.

6 Proofs

6.1 Proofs for the Generic Algorithm in Section 3

Lemma 6.1. *Let (X,d) be a connected metric space and $A \subset X$ be an open or closed subset with $\emptyset \neq A \neq X$. Then, for all $\delta > 0$ we have $\psi_A^*(\delta) \geq \delta$.*

Proof of Lemma 6.1: In view of [17, (A.5.1)] it suffices to show $d(A, X \setminus A) = 0$. Let us assume the converse, that is $\varepsilon := d(A, X \setminus A) > 0$. Then, if A is closed we know that $X \setminus A$ is open, and for $x \in A$ and $y \in X$ with $d(x,y) < \varepsilon$ our assumption yields $y \in A$. In other words, the open ball with center x and radius ε is contained in A, and therefore A is open, too. However, this gives a partition of X into two open and non-empty sets, which contradicts the assumption that X is connected. Clearly, by interchanging the roles of A and $X \setminus A$, we analogously find $d(A, X \setminus A) = 0$ for A open. □

Theorem 6.2. *Let $\rho_* \geq 0$ and Assumption P be satisfied with $|\mathcal{C}(M_\rho)| \leq 1$ for all $\rho \geq \rho_*$. Moreover, let $(L_\rho)_{\rho\geq 0}$ be a decreasing family of sets $L_\rho \subset X$ such that*

$$M_{\rho+\varepsilon}^{-\delta} \subset L_\rho \subset M_{\rho-\varepsilon}^{+\delta} \tag{47}$$

for some fixed $\varepsilon, \delta > 0$ and all $\rho \geq \rho_*$. For all $\rho \geq \rho_*$ we then have:

i) If $M_{\rho+3\varepsilon}^{-\delta} \neq \emptyset$, then for all $\tau > 2\psi^*_{M_{\rho+\varepsilon}}(\delta)$ we have $|\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta})| = 1$ and the corresponding CRM $\zeta : \mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}) \to \mathcal{C}_\tau(L_\rho)$ satisfies

$$\mathcal{C}_\tau(L_\rho) = \zeta\bigl(\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta})\bigr) \cup \bigl\{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \emptyset\bigr\}\,. \tag{48}$$

ii) If $M_{\rho+\varepsilon}^{-\delta} = \emptyset$, Assumption S is satisfied, and $\delta \in (0, \delta_{\text{thick}}]$, then we have

$$\Bigl|\bigl\{B \in \mathcal{C}_\tau(L_\rho) : B \cap L_{\rho+2\varepsilon} \neq \emptyset\bigr\}\Bigr| \leq 1\,, \qquad\qquad \tau > 2c_{\text{thick}}\delta^\gamma\,. \tag{49}$$

Proof of Theorem 6.2: *i).* We first note that $M^{-\delta}_{\rho+3\varepsilon} \neq \emptyset$ implies $M^{-\delta}_{\rho+\varepsilon} \neq \emptyset$. By (5) we thus find $|\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})| \leq |\mathcal{C}(M_{\rho+\varepsilon})| \leq 1$, and by the already observed $M_{\rho+\varepsilon} \neq \emptyset$ we conclude that $|\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})| = 1$. To establish (48) we now write $A := M^{-\delta}_{\rho+\varepsilon}$ and $B := \zeta(A)$. Our first intermediate goal is to establish the following *disjoint* union:

$$\begin{aligned}\mathcal{C}_\tau(L_\rho) = \{B\} &\cup \bigl\{B' \in \mathcal{C}_\tau(L_\rho) \setminus \{B\} : B' \cap L_{\rho+2\varepsilon} \neq \emptyset\bigr\} \\ &\cup \bigl\{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \emptyset\bigr\}\,. \end{aligned} \tag{50}$$

To this end, note that $M^{-\delta}_{\rho+3\varepsilon} \neq \emptyset$ and $M^{-\delta}_{\rho+3\varepsilon} \subset A$ together with $A \subset \zeta(A) = B$ implies

$$\emptyset \neq M^{-\delta}_{\rho+3\varepsilon} = A \cap M^{-\delta}_{\rho+3\varepsilon} \subset B \cap L_{\rho+2\varepsilon}\,.$$

This yields $\{B' \in \mathcal{C}_\tau(L_\rho) \setminus \{B\} : B' \cap L_{\rho+2\varepsilon} = \emptyset\} = \{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \emptyset\}$, which in turn implies (50).

Let us now show (48). To this end, we first observe that $|\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})| = 1$ implies $\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon}) = \{A\}$ and hence $\zeta\bigl(\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})\bigr) = \{B\}$. In view of (50) it thus remains to show

$$B' \cap L_{\rho+2\varepsilon} = \emptyset\,,$$

for all $B' \in \mathcal{C}_\tau(L_\rho)$ with $B' \neq B$. Let us assume the converse, that is, there is a $B' \in \mathcal{C}_\tau(L_\rho)$ with $B' \neq B$ and $B' \cap L_{\rho+2\varepsilon} \neq \emptyset$. Since $L_{\rho+2\varepsilon} \subset M^{+\delta}_{\rho+\varepsilon}$, there then exists an $x \in B' \cap M^{+\delta}_{\rho+\varepsilon}$. By part *i)* of [17, Lemma A.3.1] this gives an $x' \in M_{\rho+\varepsilon}$ with $d(x, x') \leq \delta$, and hence we obtain

$$d(x', M^{-\delta}_{\rho+\varepsilon}) \leq \psi^*_{M_{\rho+\varepsilon}}(\delta) < \frac{\tau}{2}\,.$$

From this inequality we conclude that there is an $x'' \in M^{-\delta}_{\rho+\varepsilon}$ satisfying $d(x', x'') < \tau/2$. The CRM property then yields $x'' \in M^{-\delta}_{\rho+\varepsilon} = A \subset B$ and therefore Lemma 6.1 yields

$$d(B', B) \leq d(x, x'') \leq d(x, x') + d(x', x'') < \delta + \tau/2 \leq \psi^*_{M_{\rho+\varepsilon}}(\delta) + \tau/2 < \tau$$

On the other hand, $B' \neq B$ implies $d(B', B'') \geq \tau$ by [17, Lemma A.2.4], and hence we have found a contradiction.

ii). Clearly, if $L_{\rho+2\varepsilon} = \emptyset$ there is nothing to prove, and hence we may assume that $L_{\rho+2\varepsilon} \neq \emptyset$. Now assume that (49) is false. Then there exist $B_1, B_2 \in \mathcal{C}_\tau(L_\rho)$ with $B_1 \neq B_2$ and $B_i \cap L_{\rho+2\varepsilon} \neq \emptyset$ for $i = 1, 2$. For $i = 1, 2$ we consequently find $x_i \in B_i \cap L_{\rho+2\varepsilon}$, and for these there exist $A_i \in \mathcal{C}_\tau(L_{\rho+2\varepsilon})$ with $x_i \in A_i$. Now recall from [17, Lemma A.2.7] that $L_{\rho+2\varepsilon} \subset L_\rho$ implies $\mathcal{C}_\tau(L_{\rho+2\varepsilon}) \sqsubset \mathcal{C}_\tau(L_\rho)$, and therefore we have a CRM $\zeta : \mathcal{C}_\tau(L_{\rho+2\varepsilon}) \to \mathcal{C}_\tau(L_\rho)$. Our construction then gives

$$x_i \in A_i \cap B_i \subset \zeta(A_i) \cap B_i\,,$$

and therefore we have $\zeta(A_i) \cap B_i \neq \emptyset$ for $i = 1, 2$. However, $\zeta(A_i)$ and B_i are both elements of the partition $\mathcal{C}_\tau(L_\rho)$ and hence we conclude $\zeta(A_i) = B_i$ for $i = 1, 2$. Moreover, ζ is a map, and therefore $B_1 \neq B_2$ implies $A_1 \neq A_2$. Let us write $A := A_1 \cup A_2$. Since we know from [17, Lemma A.2.4] that $d(A_1, A_2) \geq \tau$, we conclude by [17, Lemma A.2.8] that $\mathcal{C}_\tau(A) = \{A_1, A_2\}$, and thus $|\mathcal{C}_\tau(A)| = 2$. However, we also have

$$A \subset L_{\rho+2\varepsilon} \subset M^{+\delta}_{\rho+\varepsilon}\,,$$

and since $M^{-\delta}_{\rho+\varepsilon} = \emptyset$ holds, Assumption S together with $\delta \in (0, \delta_{\mathrm{thick}}]$ and $\tau > 2c_{\mathrm{thick}}\delta^\gamma$ ensures $|\mathcal{C}_\tau(A)| = 1$. Since this contradicts $|\mathcal{C}_\tau(A)| = 2$ we have proven (49). $\square$

Proof of Theorem 3.2: For $i \geq 0$ we write $\rho_i := \rho_0 + i\varepsilon$ for the sequence of potential levels Algorithm 1 visits. Moreover, let $i^* := \max\{i \geq 0 : M^{-\delta}_{\rho_i+3\varepsilon} \neq \emptyset\}$, where we note that this maximum is finite part *iv)* of Assumption S. For $i = 0, \dots, i^*$, part *i)* of Theorem 6.2 then shows that Algorithm 1 identifies exactly one component in its Line 3, and therefore it only identifies more than one component in Line 3, if $i \geq i^* + 1$. If it finishes the loop at Line 5, we thus know that $\rho \geq \rho_{i^*+2}$, and therefore the level ρ considered in Line 7 satisfies $\rho \geq \rho_{i^*+4}$. Now the definition of i^* yields $M^{-\delta}_{\rho_{i^*+1}+3\varepsilon} = \emptyset$, and since $\rho_{i^*+1} + 3\varepsilon = (i^* + 1)\varepsilon + 3\varepsilon = (i^* + 4)\varepsilon = \rho_{i^*+4}$, we find $M^{-\delta}_\rho = \emptyset$ for the ρ considered in Line 7. This implies $M^{-\delta}_{\rho+\varepsilon} = \emptyset$, and hence part *ii)* of Theorem 6.2 shows that Algorithm 1 identifies at most one component in Line 7. □

Lemma 6.3. *Let (X, d) be a metric space, $A, B \subset X$ be two subsets, and $\delta > 0$. Then we have*

$$\big(A \cap B\big)^{-\delta} = A^{-\delta} \cap B^{-\delta} .$$

Proof of Lemma 6.3: From part *iv)* of [17, Lemma A.3.1] we know that $(\tilde{A} \cup \tilde{B})^{+\delta} = \tilde{A}^{+\delta} \cup \tilde{B}^{+\delta}$ for all $\tilde{A}, \tilde{B} \subset X$ and $\delta > 0$. This gives

$$\begin{aligned}\big(A \cap B\big)^{-\delta} = X \setminus \big(X \setminus (A \cap B)\big)^{+\delta} &= X \setminus \big((X \setminus A) \cup (X \setminus B)\big)^{+\delta} \\ &= X \setminus \big((X \setminus A)^{+\delta} \cup (X \setminus B)^{+\delta}\big) \\ &= \big(X \setminus (X \setminus A)^{+\delta}\big) \cap \big(X \setminus (X \setminus B)^{+\delta}\big) \\ &= A^{-\delta} \cap B^{-\delta} ,\end{aligned}$$

and hence we have show the assertion. □

Lemma 6.4. *Let* Assumption M *be satisfied, $\rho \in (\rho^*, \rho^{**}]$, $\varepsilon := \rho - \rho^*$, and $A_{1,\rho}$, and $A_{2,\rho}$ be the two connected components of M_ρ, i.e. $\mathcal{C}(M_\rho) = \{A_{1,\rho}, A_{2,\rho}\}$. Then the following statements hold:*

i) For all $0 < \tau \leq 3\tau^*(\varepsilon)$ we have $\mathcal{C}_\tau(M_\rho) = \mathcal{C}(M_\rho)$.

ii) For all $0 < \delta < \tau \leq \tau^*(\varepsilon)$ we have $\mathcal{C}_\tau(M_\rho) \sqsubseteq \mathcal{C}_\tau(M^{+\delta}_\rho) = \{A^{+\delta}_{1,\rho}, A^{+\delta}_{2,\rho}\}$.

iii) For all $\delta \in (0, \delta_{\text{thick}}]$ and $\psi(\delta) < \tau \leq \tau^*(\varepsilon)$ we have $|\mathcal{C}_\tau(M^{-\delta}_\rho)| = 2$ with $\mathcal{C}_\tau(M^{-\delta}_\rho) = \{A^{-\delta}_{1,\rho}, A^{-\delta}_{2,\rho}\}$.

Proof of Lemma 6.4: To adapt to the notation of [16, 17] we write $\tau^*_{M_\rho} := d(A_{1,\rho}, A_{2,\rho})$. Note that this definition gives $\tau^*_{M_\rho} = 3\tau^*(\rho - \rho^*) = 3\tau^*(\varepsilon)$.

i). The assertion directly follows part *ii)* from [17, Proposition A.2.10].

ii). The assertion has been shown in part *iii)* of [17, Lemma A.4.1].

iii). We first note that using part *ii)* of [16, Theorem 2.7] with $\varepsilon^* := \varepsilon$ and $\rho = \rho^* + \varepsilon^*$ we find $|\mathcal{C}_\tau(M^{-\delta}_\rho)| = |\mathcal{C}(M_\rho)| = 2$. Moreover, we have

$$d(A_{1,\rho}, A_{2,\rho}) = \tau^*_{M_\rho} = 3\tau^*(\varepsilon) \geq \tau > \psi(\delta) = 3c_{\text{thick}}\delta^\gamma > \psi^*_{M_\rho}(\delta) \geq \delta , \tag{51}$$

where the last inequality follows from Lemma 6.1 since Assumption M implies Assumption P, and hence X is connected. Consequently, part *v)* of [17, Lemma A.3.1] yields

$$M^{-\delta}_\rho = \big(A_{1,\rho} \cup A_{2,\rho}\big)^{-\delta} = A^{-\delta}_{1,\rho} \cup A^{-\delta}_{2,\rho} , \tag{52}$$

and we additionally note that (51) implies

$$d(A_{1,\rho}^{-\delta}, A_{2,\rho}^{-\delta}) \geq d(A_{1,\rho}, A_{2,\rho}) \geq \tau \,. \tag{53}$$

Now let A_1 and A_2 be the two τ-connected components of $\mathcal{C}_\tau(M_\rho^{-\delta})$. Let us assume that $A_1 \neq A_{1,\rho}^{-\delta}$ and $A_1 \neq A_{2,\rho}^{-\delta}$. Then (52) shows that there exist $x' \in A_1 \cap A_{1,\rho}^{-\delta}$ and $x'' \in A_1 \cap A_{2,\rho}^{-\delta}$. Since A_1 is τ-connected, there further exist $x_1, \dots, x_n \in A_1$ with $x_1 = x'$, $x_n = x''$ and $d(x_i, x_{i+1}) < \tau$ for all $i = 1, \dots, n-1$. By $x_1 \in A_{1,\rho}^{-\delta}$, $x_n \in A_{2,\rho}^{-\delta}$, and (52) we conclude that there is an $i = \{1, \dots, n-1\}$ with $x_i \in A_{1,\rho}^{-\delta}$ and $x_{i+1} \in A_{2,\rho}^{-\delta}$. This gives

$$d(A_{1,\rho}^{-\delta}, A_{2,\rho}^{-\delta}) \leq d(x_i, x_{i+1}) < \tau \,,$$

which clearly contradicts (53). □

Lemma 6.5. *Let* Assumption M *be satisfied, and* P_1 *and* P_2 *be defined by* (10) *for some fixed* $\rho^\dagger \in (\rho^*, \rho^{**}]$. *Then for* $i = 1, 2$ *and* $\rho \geq \rho^\dagger$ *we have*

$$M_{i,\rho} = M_\rho \cap A_{i,\rho^\dagger} \,. \tag{54}$$

Proof of Lemma 6.5: We first note that since P is normal at all levels $\rho > 0$, we have $\mu(M_\rho \vartriangle \{h \geq \rho\}) = 0$ for all μ-densities h of P and all $\rho > 0$. For a fixed $\rho \geq \rho^\dagger > 0$. we can thus find a μ-density h of P such that $M_\rho = \{h \geq \rho\}$ and $M_{\rho^\dagger} = \{h \geq \rho^\dagger\}$. Let us define $h_i := \mathbf{1}_{A_{i,\rho^\dagger}} h$. Then h_i is a μ-density of P_i and we have

$$\{h_i \geq \rho\} = M_\rho \cap A_{i,\rho^\dagger} \,. \tag{55}$$

Moreover, by our definitions we find

$$M_{i,\rho} = \operatorname{supp} \mu\big(\,\cdot\, \cap \{h_i \geq \rho\}\big) \,,$$

and hence it suffices to show that $M_{i,\rho} = \{h_i \geq \rho\}$.

For the proof of the inclusion "$\subset$" we fix an $x \in M_{i,\rho}$ and an open $U \subset X$ with $x \in U$. The definition of the support of a measure then yields

$$\mu(U \cap M_\rho) = \mu\big(U \cap \{h \geq \rho\}\big) \geq \mu\big(U \cap \{h_i \geq \rho\}\big) > 0 \,,$$

which in turn implies $x \in M_\rho$. This shows $M_{i,\rho} \subset M_\rho$. Moreover, $A_{i,\rho^\dagger}$ is closed by definition and we further have

$$\mu\big(A_{i,\rho^\dagger} \cap \{h_i \geq \rho\}\big) = \mu(\{h_i \geq \rho\}) = \mu\big(X \cap \{h_i \geq \rho\}\big) \,.$$

Since the support of a finite measure is also the smallest closed subset having full measure, we conclude that $M_{i,\rho} \subset A_{i,\rho^\dagger}$. Combining the two found inclusions $M_{i,\rho} \subset M_\rho$ and $M_{i,\rho} \subset A_{i,\rho^\dagger}$ with (55) we have thus found the desired $M_{i,\rho} \subset \{h_i \geq \rho\}$.

For the proof of the converse inclusion we fix an $x \in \{h_i \geq \rho\} = M_\rho \cap A_{i,\rho^\dagger}$. Moreover, we fix an open $U \subset X$ with $x \in U$, so that it suffices to show $\mu(U \cap \{h_i \geq \rho\}) > 0$. To this end, we may assume without loss of generality that $i = 1$. Moreover, since $d(A_{1,\rho^\dagger}, A_{2,\rho^\dagger}) > 0$ and $x \in A_{1,\rho^\dagger}$ we may additionally assume that $U \cap A_{2,\rho^\dagger} = \emptyset$. Now, $x \in M_\rho$ implies $\mu(U \cap M_\rho) > 0$. Let us write $A_k := M_\rho \cap A_{k,\rho^\dagger} = \{h_k \geq \rho\}$. This yields $M_\rho = A_1 \cup A_2$, $A_1 \cap A_2 = \emptyset$, and

$$\mu(U \cap A_2) \leq \mu(U \cap A_{2,\rho^\dagger}) = 0 \,.$$

Using the disjoint union $U \cap M_\rho = (U \cap A_1) \cup (U \cap A_2)$, we conclude that

$$\mu(U \cap \{h_1 \geq \rho\}) = \mu(U \cap A_1) = \mu(U \cap M_\rho) > 0 \,.$$

As mentioned above this shows $x \in M_{1,\rho}$. □

Lemma 6.6. *Let* Assumption M *be satisfied, and* P_1 *and* P_2 *be defined by* (10) *for some fixed* $\rho^\dagger \in (\rho^*, \rho^{**}]$*. Then, for* $i = 1, 2$*, the following statements are true:*

i) For $\varepsilon^\dagger := \rho^\dagger - \rho^*$ *and all* $0 < \delta < \tau^*(\varepsilon^\dagger)$ *and* $\rho \geq \rho^\dagger$ *we have* $M_{i,\rho}^{+\delta} = M_\rho^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta}$.

ii) For all $\delta > 0$ *and* $\rho \geq \rho^\dagger$ *we have* $M_{i,\rho}^{-\delta} = M_\rho^{-\delta} \cap A_{i,\rho^\dagger}^{-\delta}$.

Proof of Lemma 6.6: *i).* Let $\xi : \mathcal{C}(M_\rho) \to \mathcal{C}(M_{\rho^\dagger})$ be the CRM and $B_1, \ldots, B_n$ be the connected components of $\mathcal{C}(M_\rho)$. Without loss of generality we may assume there is an $m \in \{0, \ldots, n\}$ such that $\xi(B_j) \subset A_{1,\rho^\dagger}$ for all $j = 1, \ldots, m$ and $\xi(B_j) \subset A_{2,\rho^\dagger}$ for all $j = m+1, \ldots, n$. We define $A_1 := B_1 \cup \cdots \cup B_m$ and $A_2 := B_{m+1} \cup \cdots \cup B_n$. Clearly, this construction ensures

$$A_k \subset \xi(A_k) \subset A_{k,\rho^\dagger}\,, \qquad\qquad k = 1, 2. \tag{56}$$

Moreover, we have $M_\rho = A_1 \cup A_2$, and hence we find

$$M_\rho^{+\delta} = A_1^{+\delta} \cup A_2^{+\delta}$$

by part *iv)* of [17, Lemma A.3.1]. In view of (54), we consequently need to prove that

$$\big((A_1 \cup A_2) \cap A_{i,\rho^\dagger}\big)^{+\delta} = \big(A_1^{+\delta} \cup A_2^{+\delta}\big) \cap A_{i,\rho^\dagger}^{+\delta}\,. \tag{57}$$

Be begin by observing that

$$(A_1 \cup A_2) \cap A_{i,\rho^\dagger} = (A_1 \ \cap A_{i,\rho^\dagger}) \cup (A_2 \ \cap A_{i,\rho^\dagger}) = A_i\,, \tag{58}$$

where we used both (56) and $A_{1,\rho^\dagger} \cap A_{2,\rho^\dagger} = \emptyset$. Similarly, the right-hand side of (57) can be written as

$$\big(A_1^{+\delta} \cup A_2^{+\delta}\big) \cap A_{i,\rho^\dagger}^{+\delta} = \big(A_1^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta}\big) \cup \big(A_2^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta}\big)\,. \tag{59}$$

In addition, (56) ensures $A_i^{+\delta} \subset A_{i,\rho^\dagger}^{+\delta}$, and by continuing (59) we thus find for $k \in \{1, 2\}$ with $k \neq i$ that

$$\big(A_1^{+\delta} \cup A_2^{+\delta}\big) \cap A_{i,\rho^\dagger}^{+\delta} = A_i^{+\delta} \cup \big(A_k^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta}\big)\,. \tag{60}$$

Moreover, by part *ii)* of Lemma 6.4 we know that $A_{1,\rho^\dagger}^{+\delta}$ and $A_{2,\rho^\dagger}^{+\delta}$ are the two τ-connected components of $M_{\rho^\dagger}^{+\delta}$ for any τ with $\delta < \tau < \tau^*(\varepsilon^\dagger)$. Consequently, we have $A_{1,\rho^\dagger}^{+\delta} \cap A_{2,\rho^\dagger}^{+\delta} = \emptyset$, and since (56) ensures $A_k^{+\delta} \subset A_{k,\rho^\dagger}^{+\delta}$ we conclude that $A_k^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta} = \emptyset$. Inserting the latter into (60) gives

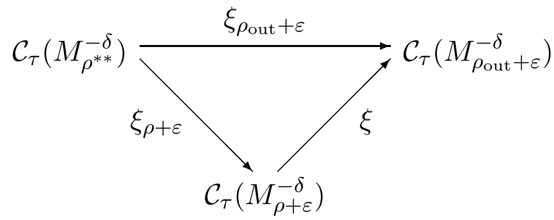
$$\big(A_1^{+\delta} \cup A_2^{+\delta}\big) \cap A_{i,\rho^\dagger}^{+\delta} = A_i^{+\delta}\,. \tag{61}$$

Now, (57) follows from combining (58) with (61).

ii). This directly follows from combining (54) with Lemma 6.3. □

Proof of Theorem 3.4: *i).* We first note that $\varepsilon^* \leq \varepsilon^{**} := \rho^{**} - \rho^*$ implies $\tau^*(\varepsilon^*) \leq \tau^*(\varepsilon^{**})$. Consequently, part *iii)* of Lemma 6.4 applied for $\rho := \rho^{**}$ gives the assertion.

ii). We begin by showing that the CRMs $\xi_{\rho+\varepsilon} : \mathcal{C}_\tau(M^{-\delta}_{\rho^{**}}) \to \mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})$ and $\xi : \mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon}) \to \mathcal{C}_\tau(M^{-\delta}_{\rho_{\mathrm{out}}+\varepsilon})$ are bijective. To this end we consider the following commutative diagram of CRMs:



Now, part *iv)* of [17, Theorem A.6.2] shows that $\xi_{\rho_{\mathrm{out}}+\varepsilon}$ is bijective, and consequently, $\xi_{\rho+\varepsilon}$ is injective. Moreover, part *i)* of [16, Theorem 2.7] shows that $1 \le |\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})| \le 2$, and since we already know that $|\mathcal{C}_\tau(M^{-\delta}_{\rho^{**}})| = 2$ and that $\xi_{\rho+\varepsilon}$ is injective, we conclude that $|\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})| = 2$ and that $\xi_{\rho+\varepsilon}$ is bijective. Using the diagram we then see that the CRM ξ is also bijective.

Our next goal is to show that the CRM $\widehat{\xi}_\rho : \mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon}) \to \widehat{\mathcal{C}}_\tau(L_\rho)$ is well-defined and bijective. To this end, we first recall that our assumption $\rho \le \rho^{**} - 3\varepsilon$ together with [16, Theorem 2.8] gives the following *disjoint* union:

$$\mathcal{C}_\tau(L_\rho) = \widehat{\xi}_\rho\big(\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})\big) \cup \big\{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \emptyset\big\} .$$

Consequently, we have $\widehat{\xi}_\rho\big(\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})\big) = \widehat{\mathcal{C}}_\tau(L_\rho)$, that is, we can view $\widehat{\xi}_\rho$ as a *surjective CRM* $\widehat{\xi}_\rho : \mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon}) \to \widehat{\mathcal{C}}_\tau(L_\rho)$. Similarly, part *i)* of Theorem 3.1 ensures

$$\rho_{\mathrm{out}} \le \rho^* + \varepsilon^* + 5\varepsilon \le \rho^* + 6\varepsilon^* \le \rho^{**} - 3\varepsilon ,$$

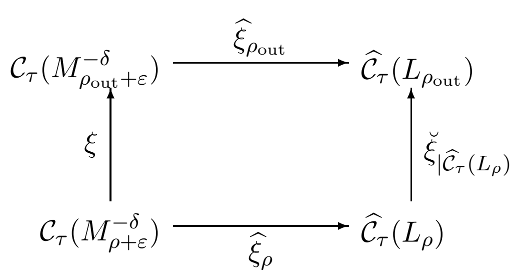
and repeating the reasoning above we see that the CRM $\widehat{\xi}_{\rho_{\mathrm{out}}} : \mathcal{C}_\tau(M^{-\delta}_{\rho_{\mathrm{out}}+\varepsilon}) \to \mathcal{C}_\tau(L_{\rho_{\mathrm{out}}})$ can be viewed as a *surjective CRM* $\widehat{\xi}_{\rho_{\mathrm{out}}} : \mathcal{C}_\tau(M^{-\delta}_{\rho_{\mathrm{out}}+\varepsilon}) \to \widehat{\mathcal{C}}_\tau(L_{\rho_{\mathrm{out}}})$. Finally, consider the CRM $\breve{\xi} : \mathcal{C}_\tau(L_\rho) \to \mathcal{C}_\tau(L_{\rho_{\mathrm{out}}})$. For $B \in \widehat{\mathcal{C}}_\tau(L_\rho)$ we then have

$$\emptyset \ne B \cap L_{\rho+2\varepsilon} \subset \breve{\xi}(B) \cap L_{\rho+2\varepsilon} \subset \breve{\xi}(B) \cap L_{\rho_{\mathrm{out}}+2\varepsilon} ,$$

i.e. we have shown $\breve{\xi}(B) \in \widehat{\mathcal{C}}_\tau(L_{\rho_{\mathrm{out}}})$. Consequently, the restriction

$$\breve{\xi}_{|\widehat{\mathcal{C}}_\tau(L_\rho)} : \widehat{\mathcal{C}}_\tau(L_\rho) \to \widehat{\mathcal{C}}_\tau(L_{\rho_{\mathrm{out}}})$$

is well-defined, and obviously also a CRM. Combining these considerations we obtain the following commutative diagram of CRMs



Now, we have already seen that ξ is bijective, and in addition, part *ii)* of [17, Theorem A.6.2] shows that $\widehat{\xi}_{\rho_{\mathrm{out}}}$ is injective. Moreover, our considerations above showed that $\widehat{\xi}_{\rho_{\mathrm{out}}} : \mathcal{C}_\tau(M^{-\delta}_{\rho_{\mathrm{out}}+\varepsilon}) \to$

$\widehat{\mathcal{C}}_\tau(L_{\rho_{\rm out}})$ is surjective, and hence the latter CRM is bijective. Using the diagram we conclude that $\widehat{\xi}_\rho : \mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon}) \to \widehat{\mathcal{C}}_\tau(L_\rho)$ is injective. Since we have already seen that it is surjective, we conclude that it is indeed bijective.

With the help of these preparations, the first assertion now easily follows from *i)* and the bijectivity of $\widehat{\xi}_\rho$ and $\xi_{\rho+\varepsilon}$, namely

$$|\widehat{\mathcal{C}}_\tau(L_\rho)| = \big|\widehat{\xi}_\rho \circ \xi_{\rho+\varepsilon}\big((\mathcal{C}_\tau(M^{-\delta}_{\rho^{**}})\big)\big| = \big|\mathcal{C}_\tau(M^{-\delta}_{\rho^{**}})\big| = 2\,.$$

To show the second assertion, we write $B_i^\rho := \widehat{\xi}_\rho \circ \xi_{\rho+\varepsilon}(V_i)$. This immediately gives $V_i \subset B_i^\rho$ for $i = 1, 2$. Moreover, using the diagram we find

$$B_i^\rho \subset \breve{\xi}_{|\widehat{\mathcal{C}}_\tau(L_\rho)}(B_i^\rho) = \breve{\xi}_{|\widehat{\mathcal{C}}_\tau(L_\rho)} \circ \widehat{\xi}_\rho \circ \xi_{\rho+\varepsilon}(V_i) = \widehat{\xi}_{\rho_{\rm out}} \circ \xi \circ \xi_{\rho+\varepsilon}(V_i) = B_i\,,$$

where the latter identity follows from part *iii)* of [17, Theorem A.6.2].

iii). We first observe that $\varepsilon^\dagger := \rho^\dagger - \rho^*$ satisfies $\varepsilon^\dagger \geq \varepsilon^*$ and by Lemma 6.1 we hence find $\delta \leq \psi^*(\delta) < \tau^*(\varepsilon^*) \leq \tau^*(\varepsilon^\dagger)$. Lemma 6.6 then shows

$$M^{-\delta}_{i,\rho} = M^{-\delta}_\rho \cap A^{-\delta}_{i,\rho^\dagger} \qquad \text{and} \qquad M^{+\delta}_{i,\rho} = M^{+\delta}_\rho \cap A^{+\delta}_{i,\rho^\dagger}\,.$$

By the definition of $L_{i,\rho}$ we thus have to show the following two inclusions

$$M^{-\delta}_{\rho+\varepsilon} \cap A^{-\delta}_{i,\rho^\dagger} \subset L_\rho \cap B_i \tag{62}$$

$$L_\rho \cap B_i \subset M^{+\delta}_{\rho-\varepsilon} \cap A^{+\delta}_{i,\rho^\dagger}\,. \tag{63}$$

We begin by proving (62). To this end, we first observe that (7) ensures $M^{-\delta}_{\rho+\varepsilon} \subset L_\rho$ and hence it suffices to establish $A^{-\delta}_{i,\rho^\dagger} \subset B_i$. Now, we have already observed that $\tau \leq \tau^*(\varepsilon^*) \leq \tau^*(\varepsilon^\dagger)$, and consequently part *iii)* of Lemma 6.4 shows that $A^{-\delta}_{1,\rho^\dagger}$ and $A^{-\delta}_{2,\rho^\dagger}$ are the two τ-connected components of $M^{-\delta}_{\rho^\dagger}$. Moreover, part *i)* of Theorem 3.1 shows $\rho_{\rm out} \leq \rho^* + \varepsilon^* + 5\varepsilon \leq \rho^\dagger - \varepsilon$, and hence we have $\rho^\dagger - \varepsilon \in [\rho_{\rm out}, \rho^{**} - 3\varepsilon]$. Applying [16, Theorem 2.8] and the already established part *ii)* to the level $\rho^\dagger - \varepsilon$ we then obtain

$$A^{-\delta}_{i,\rho^\dagger} \subset B_i^{\rho^\dagger-\varepsilon} \subset B_i\,, \qquad\qquad i = 1, 2\,.$$

Let us now establish $L_{i,\rho} \subset B_i^{\rho^\dagger+2\varepsilon}$. Without loss of generality we may assume $i = 1$. Now, consider the CRM $\xi : \mathcal{C}_\tau(L_\rho \cap B_1) \to \mathcal{C}_\tau(L_{\rho^\dagger+2\varepsilon} \cap B_1)$, which is possible since $\rho \geq \rho^\dagger + 2\varepsilon$. Let us assume that there was a $B' \in \mathcal{C}_\tau(L_\rho \cap B_1)$ with

$$\xi(B') \not\subset B_1^{\rho^\dagger+2\varepsilon}\,.$$

Since $B_1^{\rho^\dagger+2\varepsilon}$ is a τ-connected component of $L_{\rho^\dagger+2\varepsilon} \cap B_1$ by part *ii)* applied to the level $\rho^\dagger + 2\varepsilon \in [\rho_{\rm out}, \rho^{**} - 3\varepsilon]$ and $\xi(B')$ is another such τ-connected component we conclude that $\xi(B') \cap B_1^{\rho^\dagger+2\varepsilon} = \emptyset$. Moreover, our construction and part *ii)* give

$$\xi(B') \cap B_2^{\rho^\dagger+2\varepsilon} \subset B_1 \cap B_2^{\rho^\dagger+2\varepsilon} \subset B_1 \cap B_2 = \emptyset\,,$$

and therefore part *ii)* shows $\xi(B') \notin \widehat{\mathcal{C}}_\tau(L_{\rho^\dagger} + 2\varepsilon)$. Together with $\rho \geq \rho^\dagger + 4\varepsilon$ the latter implies

$$B' \cap L_\rho \subset B' \cap L_{\rho^\dagger+4\varepsilon} \subset \xi(B') \cap L_{\rho^\dagger+4\varepsilon} = \emptyset\,.$$

Consequently, we have found a contradiction, and therefore we have $\xi(B') \subset B_1^{\rho^\dagger+2\varepsilon}$ for all τ-connected components of $L_{1,\rho} = L_\rho \cap B_1$. Since $B' \subset \xi(B')$ we have thus found $L_{1,\rho} \subset B_1^{\rho^\dagger+2\varepsilon}$.

Let us now show (63). To this end, we note that (7) ensures $L_\rho \subset M_{\rho-\varepsilon}^{+\delta}$, and hence it suffices to prove $L_\rho \cap B_i \subset A_{i,\rho^\dagger}^{+\delta}$. Moreover, we have already shown that $L_\rho \cap B_i \subset B_i^{\rho^\dagger+2\varepsilon}$, and therefore, it suffices to establish

$$B_i^{\rho^\dagger+2\varepsilon} \subset A_{i,\rho^\dagger}^{+\delta} .$$

To this end, recall that we have already observed $\tau \leq \tau^*(\varepsilon^*) \leq \tau^*(\varepsilon^\dagger)$. Part *ii)* of Lemma 6.4 thus shows that $A_{1,\rho^\dagger}^{+\delta}$ and $A_{2,\rho^\dagger}^{+\delta}$ are the two τ-connected components of $M_{\rho^\dagger}^{+\delta}$. Now consider the CRM $\xi : \mathcal{C}_\tau(L_{\rho^\dagger+2\varepsilon}) \to \mathcal{C}_\tau(M_{\rho^\dagger}^{+\delta})$. Then the τ-connected component $B_1^{\rho^\dagger+2\varepsilon}$ of $L_{\rho^\dagger+2\varepsilon}$ satisfies $B_1^{\rho^\dagger+2\varepsilon} \subset \xi(B_1^{\rho^\dagger+2\varepsilon})$, and therefore, exactly one of the following two conditions is satisfied

$$B_1^{\rho^\dagger+2\varepsilon} \subset A_{1,\rho^\dagger}^{+\delta} , \tag{64}$$

$$B_1^{\rho^\dagger+2\varepsilon} \subset A_{2,\rho^\dagger}^{+\delta} . \tag{65}$$

However, our construction ensures $V_1 \subset A_{1,\rho^\dagger}^{+\delta}$, and part *ii)* gives $V_1 \subset B_1^{\rho^\dagger+2\varepsilon}$. This gives $\emptyset \neq V_1 \subset B_1^{\rho^\dagger+2\varepsilon} \cap A_{1,\rho^\dagger}^{+\delta}$, and therefore we can exclude (65). Consequently (64) is true. The inclusion $B_2^{\rho^\dagger+2\varepsilon} \subset A_{2,\rho^\dagger}^{+\delta}$ can be shown analogously. □

6.2 Proofs for Section 4

Proof of Lemma 4.2: For the tail function $\kappa_1(\cdot)$ the estimate follows from

$$\begin{aligned} \kappa_1(r) &= \int_{\mathbb{R}^d \setminus B(0,r)} K(x) \, \mathrm{d}\lambda^d(x) \leq c \int_{\mathbb{R}^d \setminus B(0,r)} \exp\bigl(-\|x\|_2\bigr) \, \mathrm{d}\lambda^d(x) \\ &= cd \, \mathrm{vol_d} \int_r^\infty e^{-s} s^{d-1} ds \leq cd^2 \, \mathrm{vol_d} \, e^{-r} r^{d-1} , \end{aligned}$$

where the last estimate for the incomplete gamma function is taken from [19, Lemma A.1.1]. The second inequality follows from the monotonicity of the function $r \mapsto e^{-r}$. □

Lemma 6.7. *Let $K : \mathbb{R}^d \to [0, \infty)$ be a symmetric kernel with tail function $\kappa_1(\cdot)$. Moreover, let P be a λ^d-absolutely continuous distribution on $\mathbb{R}^d$ that is normal at some level $\rho \geq 0$. Then for all $x \in \mathbb{R}^d$ and $\sigma > 0$ with $B(x, \sigma) \subset M_\rho$ and all $\delta > 0$ we have*

$$h_{P,\delta}(x) \geq \rho - \rho\kappa_1(\tfrac{\sigma}{\delta}) \tag{66}$$

while for all $x \in \mathbb{R}^d$ and $\sigma > 0$ with $B(x, \sigma) \subset X \setminus M_\rho$ and all $\delta > 0$ we have

$$h_{P,\delta}(x) < \rho + \delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta}) . \tag{67}$$

Finally, if P has a bounded density h, then the inequality (66) *can be replaced by*

$$h_{P,\delta}(x) \geq \rho - \kappa_1(\tfrac{\sigma}{\delta}) \cdot \|h\|_\infty \tag{68}$$

whenever $0 \leq \rho \leq \|h\|_\infty$ and (67) *can be replaced, for all $\rho \geq 0$, by*

$$h_{P,\delta}(x) < \rho + \kappa_1(\tfrac{\sigma}{\delta}) \cdot \|h\|_\infty . \tag{69}$$

Proof of Lemma 6.7: Let h be a λ^d-density of P. We begin by proving (66). To this end, we first observe that $\lambda^d(B(x,\sigma) \setminus \{h \geq \rho\}) \leq \lambda^d(M_\rho \setminus \{h \geq \rho\}) = 0$, since P is normal at level ρ. Therefore, we obtain

$$\begin{aligned}\int_{B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) &= \int_{B(x,\sigma)\cap\{h\geq\rho\}} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) \\ &\geq \rho \int_{B(x,\sigma)\cap\{h\geq\rho\}} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) \\ &= \rho \int_{B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y)\,, \end{aligned} \tag{70}$$

and this leads to

$$\begin{aligned} h_{P,\delta}(x) &= \int_{\mathbb{R}^d} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) \\ &\geq \rho \int_{B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) + \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) \\ &= \rho \int_{B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) + \rho \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) \\ &\qquad - \rho \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) + \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) \\ &\geq \rho - \rho \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y)\,, \end{aligned}$$

where in the last step we used (15). In the case of a general density h the assertion now follows from (16), and for a bounded density h and $\rho \leq \|h\|_\infty$ the inequality (68) is a direct consequence of (66).

To show (67) we first note that (2) yields

$$\lambda^d\big(B(x,\sigma) \setminus \{h<\rho\}\big) \leq \lambda^d\big((\mathbb{R}^d \setminus M_\rho) \setminus \{h<\rho\}\big) = \lambda^d\big(\{h \geq \rho\} \setminus M_\rho\big) = 0\,.$$

Analogously to (70) we then obtain

$$\begin{aligned}\int_{B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) &= \int_{B(x,\sigma)\cap\{h<\rho\}} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) \\ &< \rho \int_{B(x,\sigma)\cap\{h<\rho\}} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) \\ &= \rho \int_{B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y)\,, \end{aligned}$$

where for the strict inequality we used our assumption that K is strictly positive in a neighborhood of 0. Adapting the last estimate of the proof of (66) we then find

$$\begin{aligned} h_{P,\delta}(x) &< \rho \int_{B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) + \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) \\ &= \rho \int_{B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) + \rho \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) \\ &\qquad - \rho \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y) + \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) \\ &\leq \rho + \int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y)\,. \end{aligned}$$

Now, in the case of a bounded density h the inequality (69) follows from (16), while in the general case the estimate

$$\int_{\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y)\,h(y)\,\mathrm{d}\lambda^d(y) \leq \sup_{y\in\mathbb{R}^d\setminus B(x,\sigma)} K_\delta(x-y) = \delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta})$$

leads to (67). □

Proof of Theorem 4.3: We begin by proving the first inclusion. To this end, we fix an $x \in M^{-2\sigma}_{\rho+\varepsilon+\epsilon} = \mathbb{R}^d \setminus (\mathbb{R}^d \setminus M_{\rho+\varepsilon+\epsilon})^{+2\sigma}$. This means $x \notin (\mathbb{R}^d \setminus M_{\rho+\varepsilon+\epsilon})^{+2\sigma}$, i.e., for all $x' \in \mathbb{R}^d \setminus M_{\rho+\varepsilon+\varepsilon}$ we have $\|x-x'\| > 2\sigma$. In other words, for all $x' \in \mathbb{R}^d$ with $\|x-x'\| \leq 2\sigma$, we have $x' \in M_{\rho+\varepsilon+\epsilon}$, i.e., we have found

$$B(x,2\sigma) \subset M_{\rho+\varepsilon+\epsilon}\,. \tag{71}$$

Let us now suppose that there exists a sample $x_i \in D$ such that $h_{D,\delta}(x_i) < \rho$ and $\|x-x_i\| \leq \sigma$. By $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon$ we then find

$$h_{P,\delta}(x_i) < \rho+\varepsilon\,. \tag{72}$$

On the other hand, $\|x-x_i\| \leq \sigma$ together with the already shown (71) implies $B(x_i,\sigma) \subset M_{\rho+\varepsilon+\epsilon}$ by a simple application of the triangle inequality. Consequently, (66) together with $\epsilon \geq \rho\kappa_1(\frac{\sigma}{\delta})$ gives $h_{P,\delta}(x_i) \geq \rho+\varepsilon$, which contradicts (72). For all samples $x_i \in D$, we thus have $h_{D,\delta}(x_i) \geq \rho$ or $\|x-x_i\| > \sigma$. Let us assume that we have $\|x-x_i\| > \sigma$ for all $x_i \in D$. Then we find

$$h_{D,\delta}(x) = \frac{1}{n}\sum_{i=1}^n \delta^{-d} K\Big(\frac{x-x_i}{\delta}\Big) \leq \frac{1}{n}\sum_{i=1}^n \delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta}) = \delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta}) \leq \rho\,.$$

On the other hand, we have $B(x,\sigma) \subset B(x,2\sigma) \subset M_{\rho+\varepsilon+\epsilon}$ and therefore (66) together with $\epsilon \geq \rho\kappa_1(\frac{\sigma}{\delta})$ gives $h_{P,\delta}(x) \geq \rho+\varepsilon$. By $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon$ we conclude that $h_{D,\delta}(x) > \rho$, and hence we have found a contradiction. Therefore there does exist a sample $x_i \in D$ with $\|x-x_i\| \leq \sigma$. Using the inclusion (71) together with the triangle inequality we then again find $B(x_i,\delta) \subset M_{\rho+\varepsilon+\epsilon}$, and hence (66) yields $h_{P,\delta}(x_i) \geq \rho+\varepsilon$. This leads to $h_{D,\delta}(x_i) \geq \rho$, and hence we finally obtain

$$x \in \{x' \in D : h_{D,\delta}(x') \geq \rho\}^{+\sigma} = L_{D,\rho}\,.$$

Finally, if h has a bounded density, then we have $M_\rho = \emptyset$ for $\rho > \|h\|_\infty$ and therefore $M^{-2\sigma}_{\rho+\varepsilon+\epsilon} \subset L_{D,\rho}$ is trivially satisfied. Moreover, to show the assertion for $\rho \leq \|h\|_\infty$, we simply need to replace (66) with (68) in the proof above.

Let us now prove the second inclusion. To this end, we pick an $x \in L_{D,\rho}$. By the definition of $L_{D,\rho}$, there then exists an $x_i \in D$ such that $\|x-x_i\| \leq \sigma$ and $h_{D,\delta}(x_i) \geq \rho$. The latter implies $h_{P,\delta}(x_i) > \rho-\varepsilon$.

Our first goal is to show that $M_{\rho-\varepsilon-\epsilon} \cap B(x_i,\sigma) \neq \emptyset$. To this end, let us assume the converse, that is $B(x_i,\sigma) \subset \mathbb{R}^d \setminus M_{\rho-\varepsilon-\epsilon}$. By (67) and $\epsilon \geq \delta^{-d}\kappa_\infty(\frac{\sigma}{\delta})$ we then find $h_{P,\delta}(x_i) < \rho-\varepsilon$, which contradicts the earlier established $h_{P,\delta}(x_i) > \rho-\varepsilon$. Consequently, there exists an $\tilde{x} \in M_{\rho-\varepsilon-\epsilon} \cap B(x_i,\sigma)$, which in turn leads to

$$d(x, M_{\rho-\varepsilon-\epsilon}) \leq \|x-\tilde{x}\| \leq \|x-x_i\| + \|x_i-\tilde{x}\| \leq 2\sigma\,.$$

This shows the desired $x \in M^{+2\sigma}_{\rho-\varepsilon-\epsilon}$. Finally, to show the assertion for bounded densities, we simply need to replace (67) with (69) in the proof above. □

For the proof of Lemma 4.6 we need to recall the following classical result, which is a reformulation of [22, Theorem 2.6.4].

Theorem 6.8. *Let $\mathcal{A}$ be a set of subsets of Z that has finite VC-dimension V. Then the set of indicator functions $\mathcal{G} := \{\mathbf{1}_A : A \in \mathcal{A}\}$ is a uniformly bounded VC-class for which we have $B = 1$ and the constants A and ν in (22) only depend on V.*

We also need the next result, which investigates the effect of scaling in the input space.

Lemma 6.9. *Let $\mathcal{G}$ be set of measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ such that there exists a constant $B \geq 0$ with $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}$. For $\delta > 0$, we define $g_\delta : \mathbb{R}^d \to \mathbb{R}$ by $g_\delta(x) := g(x/\delta)$, $x \in \mathbb{R}^d$ Furthermore, we write $\mathcal{G}_\delta := \{g_\delta : g \in \mathcal{G}\}$. Then, for all $\epsilon \in (0, B]$ and all $\delta > 0$, we have*

$$\sup_P \mathcal{N}(\mathcal{G}, L_2(P), \epsilon) = \sup_P \mathcal{N}(\mathcal{G}_\delta, L_2(P), \epsilon)\,,$$

where the suprema are taken over all probability measures P on $\mathbb{R}^d$.

Proof of Lemma 6.9: Because of symmetry we only prove "$\leq$". Let us fix $\epsilon, \delta > 0$ and a distribution P on $\mathbb{R}^d$. We define a new distribution P' on $\mathbb{R}^d$ by $P'(A) := P(\frac{1}{\delta}A)$ for all measurable $A \subset \mathbb{R}^d$. Furthermore, let $\mathcal{C}'$ be an ϵ-net of $\mathcal{G}_\delta$ with respect to $L_2(P')$. For $\mathcal{C} := \mathcal{C}'_{1/\delta}$, we then have $|\mathcal{C}| = |\mathcal{C}'|$, and hence it suffices to show that $\mathcal{C}$ is an ϵ-net of $\mathcal{G}$ with respect to $L_2(P)$. To this end, we fix a $g \in \mathcal{G}$. Then $g_\delta \in \mathcal{G}_\delta$, and hence there exists an $h' \in \mathcal{C}'$ with $\|g_\delta - h'\|_{L_2(P')} \leq \epsilon$. Moreover, we have $h := h'_{1/\delta} \in \mathcal{C}$, and since the definition of P' ensures $\mathbb{E}_{P'} f_\delta = \mathbb{E}_P f$ for all measurable $f : \mathbb{R}^d \to [0, \infty)$, we obtain

$$\|g - h\|_{L_2(P)} = \|g_\delta - h_\delta\|_{L_2(P')} = \|g_\delta - h'\|_{L_2(P')} \leq \epsilon,$$

i.e. $\mathcal{C}$ is an ϵ-net of $\mathcal{G}$ with respect to $L_2(P)$. □

Proof of Lemma 4.6: The collection $\mathcal{A} := \{x + B_{\|\cdot\|} : x \in \mathbb{R}^d\}$ of closed balls with radius 1 has finite VC-dimension by [4, Corollary 4.2] or [4, Lemma 4.1], respectively. In both cases, Theorem 6.8 thus shows that

$$\mathcal{G} := \bigl\{K(x - \cdot) : x \in \mathbb{R}^d\bigr\}$$

there are constants A and ν only depending on the VC-dimension of $\mathcal{A}$ such that

$$\mathcal{N}\bigl(\mathcal{G}, L_2(P), \|K\|_\infty \epsilon\bigr) = \mathcal{N}\bigl(\|K\|_\infty^{-1}\mathcal{G}, L_2(P), \epsilon\bigr) \leq \left(\frac{A}{\epsilon}\right)^\nu$$

for all $\epsilon \in (0, 1]$ and all distributions P on $\mathbb{R}^d$. Our next step is to apply Lemma 6.9. To this end, we first observe that

$$\mathcal{G}_\delta = \bigl\{K(x - \delta^{-1}\cdot\,) : x \in \mathbb{R}^d\bigr\} = \left\{K\Bigl(\frac{x' - \cdot}{\delta}\Bigr) : x' \in \mathbb{R}^d\right\} = \delta^d \mathcal{K}_\delta\,.$$

Consequently, Lemma 6.9 leads to

$$\begin{aligned}\sup_P \mathcal{N}(\mathcal{K}_\delta, L_2(P), \delta^{-d}\epsilon) &= \sup_P \mathcal{N}(\delta^d \mathcal{K}_\delta, L_2(P), \epsilon) \\ &= \sup_P \mathcal{N}(\mathcal{G}, L_2(P), \epsilon) \leq \left(\frac{A\|K\|_\infty}{\epsilon}\right)^\nu\end{aligned}$$

for all $\epsilon \in (0, \|K\|_\infty]$. A simple variable transformation then yields the assertion. □

Proof of Lemma 4.7: We first observe that $A \subset E$ is a compact subset of some Banach space E and $T : A \to F$ is a α-Hölder continuous map into another Banach space F with Lipschitz norm then, for all $\epsilon > 0$, we have

$$\mathcal{N}\big(T(A), \|\cdot\|_F, |T|_\alpha \epsilon^\alpha\big) \le \mathcal{N}\big(A, \|\cdot\|_E, \epsilon\big)\,,$$

where $|T|_\alpha$ denotes the α-Hölder constant of T. In the following we fix a $\delta > 0$ with $\delta \le \big(\frac{|K|_\alpha}{\|K\|_\infty}\big)^{1/\alpha} \operatorname{diam}_{\|\cdot\|}(X)$ and a probability measure P on $\mathbb{R}^d$. For $x \in X$ we now consider the map $k_{x,\delta} : \mathbb{R}^d \to [0, \infty]$ defined by

$$k_{x,\delta}(y) := K_\delta(x - y) = \delta^{-d} K\Big(\frac{x - y}{\delta}\Big)\,, \qquad\qquad y \in \mathbb{R}^d.$$

Since K is bounded and measurable, so is $k_{x,\delta}$, and hence we obtain a map $T : X \to L_\infty(P)$ defined by $T(x) := k_{x,\delta}$. Our next goal is to show that T is α-Hölder continuous. To this end, we pick $x, x' \in X$. A simple estimate then yields

$$\begin{aligned}\|T(x) - T(x')\|_\infty &= \sup_{y\in\mathbb{R}^d}\Big|\delta^{-d} K\Big(\frac{x - y}{\delta}\Big) - \delta^{-d} K\Big(\frac{x' - y}{\delta}\Big)\Big| \\ &\le \delta^{-(\alpha+d)} |K|_\alpha\, \|x - x'\|^\alpha\,,\end{aligned}$$

i.e. T is indeed α-Hölder continuous with $|T|_\alpha \le \delta^{-(\alpha+d)}|K|_\alpha$. By our initial observation and (24) we then conclude that

$$\begin{aligned}\mathcal{N}\big(\mathcal{K}_\delta, \|\cdot\|_{L_2(P)}, |T|_\alpha\epsilon^\alpha\big) &= \mathcal{N}\big(T(X), \|\cdot\|_{L_2(P)}, |T|_\alpha\epsilon^\alpha\big) \\ &\le \mathcal{N}\big(X, \|\cdot\|_\infty, \epsilon\big) \le C_{\|\cdot\|}(X)\epsilon^{-d}\end{aligned}$$

for all $0 < \epsilon \le \operatorname{diam}_{\|\cdot\|}(X)$. A simple variable transformation together with our bound on $|T|_\alpha$ thus yields

$$\mathcal{N}\big(\mathcal{K}_\delta, \|\cdot\|_{L_2(P)}, \epsilon\big) \le C_{\|\cdot\|}(X)\bigg(\frac{|T|_\alpha}{\epsilon}\bigg)^{d/\alpha} \le C_{\|\cdot\|}(X)\bigg(\frac{|K|_\alpha}{\delta^{\alpha+d}\epsilon}\bigg)^{d/\alpha}$$

for all $0 < \epsilon \le \delta^{-(\alpha+d)}|K|_\alpha \operatorname{diam}_{\|\cdot\|}(X)$. Since the assumed

$$\delta \le \big(\frac{|K|_\alpha}{\|K\|_\infty}\big)^{1/\alpha} \operatorname{diam}_{\|\cdot\|}(X)$$

implies

$$\delta^{-d}\|K\|_\infty \le \big(\operatorname{diam}_{\|\cdot\|}(X)\big)^\alpha \delta^{-(\alpha+d)}|K|_\alpha$$

we then see that (25) does hold for all $0 < \epsilon \le \delta^{-d}\|K\|_\infty$. □

For the proof of Theorem 4.8 we quote a version of Talagrand's inequality due to [1] from [19, Theorem 7.5].

Theorem 6.10. *Let (Z, P) be a probability space and $\mathcal{G}$ be a set of measurable functions from Z to $\mathbb{R}$. Furthermore, let $B \ge 0$ and $\sigma \ge 0$ be constants such that $\mathbb{E}_P g = 0$, $\mathbb{E}_P g^2 \le \sigma^2$, and $\|g\|_\infty \le B$ for all $g \in \mathcal{G}$. For $n \ge 1$, define $G : Z^n \to \mathbb{R}$ by*

$$G(z) := \sup_{g\in\mathcal{G}}\bigg|\frac{1}{n}\sum_{j=1}^n g(z_j)\bigg|\,, \quad z = (z_1, \ldots, z_n) \in Z^n.$$

Then, for all $\varsigma > 0$, we have

$$P^n\bigg(\bigg\{z \in Z^n \,:\, G(z) \ge 4\mathbb{E}_{P^n} G + \sqrt{\frac{2\varsigma\sigma^2}{n}} + \frac{\varsigma B}{n}\bigg\}\bigg) \le e^{-\varsigma}.$$

For the proof of Theorem 4.8 we also need [5, Proposition 2.1], which bounds the expected suprema of empirical processes indexed by uniformly bounded VC-classes. The following theorem provides a slightly simplified version of that proposition.

Theorem 6.11. *Let (Z, P) be a probability space and $\mathcal{G}$ be a uniformly bounded VC-class on Z with constants A, B, and ν. Furthermore, let $\sigma > 0$ be a constant such that $\sigma \leq B$ and $\mathbb{E}_P g^2 \leq \sigma^2$ for all $g \in \mathcal{G}$. Then there exists a universal constant C such that G defined as in Theorem 6.10 satisfies*

$$\mathbb{E}_{P^n} G \leq C \left(\frac{\nu B}{n} \log \frac{AB}{\sigma} + \sqrt{\frac{\nu \sigma^2}{n} \log \frac{AB}{\sigma}} \right) . \tag{73}$$

We are now able to establish the following generalization of Theorem 4.8.

Proposition 6.12. *Let $X \subset \mathbb{R}^d$ and P be a probability measure on X that has a Lebesgue density $h \in L_1(\mathbb{R}^d) \cap L_p(\mathbb{R}^d)$ for some $p \in (1, \infty]$. Moreover, let $\frac{1}{p} + \frac{1}{p'} = 1$ and $q := \frac{1}{2p'} = \frac{1}{2} - \frac{1}{2p}$ and $K : \mathbb{R}^d \to [0, \infty)$ be a symmetric kernel. Suppose further that the set $\mathcal{K}_\delta$ defined in (23) satisfies (26) for all $\delta \in (0, \delta_0]$, where $\delta_0 \in (0, 1]$. Then, there exists a positive constant C only depending on d, p, and K such that, for all $n \geq 1$, all $\delta \in (0, \delta_0]$ satisfying $\delta \|h\|_p^{p'} \leq 4^{p'} \|K\|_\infty$, and all $\varsigma \geq 1$ we have*

$$P^n \left(\left\{ D : \|h_{D,\delta} - h_{P,\delta}\|_{\ell_\infty(X)} < \frac{C\varsigma}{n\delta^d} \log \frac{C}{\delta^{a+dq} \|h\|_p^{1/2}} + \sqrt{\frac{C\|h\|_p \varsigma}{\delta^{d(1+1/p)} n} \log \frac{C}{\delta^{a+dq} \|h\|_p^{1/2}}} \right\} \right) \geq 1 - e^{-\varsigma}.$$

Proof of Proposition 6.12: We define $\theta := 1 - \frac{1}{2p'} = \frac{1}{2} + \frac{1}{2p}$. Then $K \in L_1(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ leads to

$$\|K\|_{2p'} \leq \|K\|_1^{1-\theta} \|K\|_\infty^\theta = \|K\|_\infty^\theta .$$

We further define

$$k_{x,\delta} := \delta^{-d} K \left(\frac{x - \cdot}{\delta} \right)$$

and $f_{x,\delta} := k_{x,\delta} - \mathbb{E}_P k_{x,\delta}$. Then it is easy to check that $\mathbb{E}_P f_{x,\delta} = 0$ and $\|f_{x,\delta}\|_\infty \leq 2\|K\|_\infty \delta^{-d}$ for all $x \in X$ and $\delta > 0$. Moreover, we have $\mathbb{E}_P f_{x,\delta}^2 \leq \mathbb{E}_P k_{x,\delta}^2$ and thus

$$\begin{aligned}
\mathbb{E}_P f_{x,\delta}^2 &= \delta^{-2d} \int_{\mathbb{R}^d} K^2 \left(\frac{x - y}{\delta} \right) h(y) \, \mathrm{d}\lambda^d(y) \\
&\leq \delta^{-2d} \|h\|_p \left(\int_{\mathbb{R}^d} K^{2p'} \left(\frac{x - y}{\delta} \right) \mathrm{d}\lambda^d(y) \right)^{1/p'} \\
&= \delta^{-2d} \|h\|_p \left(\delta^d \int_{\mathbb{R}^d} K^{2p'}(x - y) \, \mathrm{d}\lambda^d(y) \right)^{1/p'} \\
&\leq \delta^{-d(1+1/p)} \|h\|_p \|K\|_\infty^{2\theta} \\
&=: \sigma_\delta^2
\end{aligned}$$

for all $x \in X$ and $\delta > 0$. In addition, for all $D \in X^n$ we have

$$\mathbb{E}_D f_{x,\delta} = \frac{1}{n} \sum_{i=1}^n f_{x,\delta}(x_i) = h_{D,\delta}(x) - h_{P,\delta}(x) . \tag{74}$$

Applying Theorem 6.10 to $\mathcal{G} := \{f_{x,\delta} : x \in X\}$, we hence obtain, for all $\delta > 0$, $\varsigma > 0$, and $n \geq 1$, that

$$\|h_{D,\delta} - h_{P,\delta}\|_{\ell_\infty(X)} < 4\mathbb{E}_{D'\sim P^n}\|h_{D',\delta} - h_{P,\delta}\|_{\ell_\infty(X)} + \frac{2\varsigma}{n\delta^d} + \sqrt{\frac{2\varsigma\|h\|_p\|K\|_\infty^{2\theta}}{n\delta^{d(1+1/p)}}} \qquad (75)$$

holds with probability P^n not smaller than $1 - e^{-\varsigma}$. It thus remains to bound the term

$$\mathbb{E}_{D'\sim P^n}\|h_{D',\delta} - h_{P,\delta}\|_{\ell_\infty(X)} = \mathbb{E}_{D'\sim P^n} \sup_{x\in X}\bigl|\mathbb{E}_D f_{x,\delta}\bigr| \,.$$

To this end, we first note that $|\mathbb{E}_P k_{x,\delta}| \leq \|k_{x,\delta}\|_\infty = \delta^{-d}\|K\|_\infty =: B_\delta$. Consequently, we have

$$\mathcal{F}_\delta := \{f_{x,\delta} : x \in X\} \subset \bigl\{k_{x,\delta} - b : k_{x,\delta} \in \mathcal{K}_\delta, |b| \leq B_\delta\bigr\}\,,$$

and since $\mathcal{N}([-B_\delta, B_\delta], |\cdot|, \epsilon) \leq 2B_\delta\epsilon^{-1}$ we conclude that for $\tilde{A} := \max\{1, A_0\}$ we have

$$\begin{aligned}\sup_Q \mathcal{N}\bigl(\mathcal{F}_\delta, L_2(Q), \epsilon\bigr) &\leq 2\biggl(\frac{A_0\|K\|_\infty\delta^{-(d+a)}}{\epsilon}\biggr)^\nu \cdot \frac{\|K\|_\infty\delta^{-d}}{\epsilon}\\ &\leq \biggl(\frac{2\tilde{A}\|K\|_\infty\delta^{-(d+a)}}{\epsilon}\biggr)^{\nu+1}\end{aligned}$$

for all $\delta \in (0, \delta_0]$ and all $\epsilon \in (0, B_\delta]$, where the supremum runs over all distributions Q on X. Now, our very first estimates showed $\|f_{x,\delta}\|_\infty \leq 2B_\delta$ and $\mathbb{E}_P f_{x,\delta}^2 \leq \sigma_\delta^2$ and since $\sigma_\delta \leq 2B_\delta$ is equivalent to

$$\delta \leq 4^{p'}\frac{\|K\|_\infty^{p'(2-2\theta)}}{\|h\|_p^{p'}} = 4^{p'}\frac{\|K\|_\infty}{\|h\|_p^{p'}}$$

Theorem 6.11 together with $2\theta = 1 + 1/p$ thus yields

$$\mathbb{E}_{D'\sim P^n}\sup_{x\in X}\bigl|\mathbb{E}_D f_{x,\delta}\bigr| \leq C\left(\frac{2(\nu{+}1)\|K\|_\infty}{n\delta^d}\log\frac{2\tilde{A}\|K\|_\infty}{\sigma_\delta\delta^{d+a}} + \sqrt{\frac{(\nu{+}1)\|h\|_p\|K\|_\infty^{1+1/p}}{2\delta^{d(1+1/p)}n}\log\frac{2\tilde{A}\|K\|_\infty}{\sigma_\delta\delta^{d+a}}}\right).$$

for such δ. Moreover, we have

$$\log\frac{2\tilde{A}\|K\|_\infty}{\sigma_\delta\delta^{d+a}} = \log\frac{2\tilde{A}\|K\|_\infty}{\delta^{-d(1+1/p)/2}\|h\|_p^{1/2}\|K\|_\infty^\theta\delta^{d+a}} = \log\frac{2\tilde{A}\|K\|_\infty^q}{\delta^{a+dq}\|h\|_p^{1/2}}\,,$$

and hence the previous estimate can be simplified to

$$\mathbb{E}_{D'\sim P^n}\sup_{x\in X}\bigl|\mathbb{E}_D f_{x,\delta}\bigr| \leq C\left(\frac{4\nu\|K\|_\infty}{n\delta^d}\log\frac{2\tilde{A}\|K\|_\infty^q}{\delta^{a+dq}\|h\|_p^{1/2}} + \sqrt{\frac{\nu\|h\|_p\|K\|_\infty^{1+1/p}}{\delta^{d(1+1/p)}n}\log\frac{2\tilde{A}\|K\|_\infty^q}{\delta^{a+dq}\|h\|_p^{1/2}}}\right).$$

Combining this with (75) gives

$$
\begin{aligned}
\|h_{D,\delta} - h_{P,\delta}\|_{\ell_\infty(X)} &< 4\mathbb{E}_{D'\sim P^n}\|h_{D',\delta} - h_{P,\delta}\|_{\ell_\infty(X)} + \frac{2\varsigma}{n\delta^d} + \sqrt{\frac{2\varsigma\|h\|_p\|K\|_\infty^{1+1/p}}{n\delta^{d(1+1/p)}}} \\
&\le 4C\left(\frac{4\nu\|K\|_\infty}{n\delta^d}\log\frac{2\tilde{A}\|K\|_\infty^q}{\delta^{a+dq}\|h\|_p^{1/2}} + \sqrt{\frac{\nu\|h\|_p\|K\|_\infty^{1+1/p}}{\delta^{d(1+1/p)}n}\log\frac{2\tilde{A}\|K\|_\infty^q}{\delta^{a+dq}\|h\|_p^{1/2}}}\right) \\
&\quad + \frac{2\varsigma}{n\delta^d} + \sqrt{\frac{2\varsigma\|h\|_p\|K\|_\infty^{1+1/p}}{n\delta^{d(1+1/p)}}} \\
&\le \frac{\tilde{C}\varsigma}{n\delta^d}\log\frac{\tilde{C}}{\delta^{a+dq}\|h\|_p^{1/2}} + \sqrt{\frac{\tilde{C}\|h\|_p\varsigma}{\delta^{d(1+1/p)}n}\log\frac{\tilde{C}}{\delta^{a+dq}\|h\|_p^{1/2}}}
\end{aligned}
$$

with probability P^n not smaller than $1 - e^{-\varsigma}$. □

Proof of Theorem 4.8: By Proposition 6.12, it suffices to find a constant C' such that

$$
\frac{C\varsigma}{n\delta^d}\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}} + \sqrt{\frac{C\|h\|_p\varsigma}{\delta^{d(1+1/p)}n}\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}}} \le C'\sqrt{\frac{\|h\|_p\,|\log\delta|\,\varsigma}{n\delta^{d(1+1/p)}}}\,. \tag{76}
$$

To this end, we first observe that $\delta^{a+dq} \le \frac{\|h\|_p^{1/2}}{C}$ implies

$$
\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}} \le \delta^{-2a-2dq}\,,
$$

and thus we obtain $\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}} \le (2a+2dq)\log\delta^{-1}$. For $C'' := (2a+2dq)C$ we therefore find

$$
\frac{C\varsigma}{n\delta^d}\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}} + \sqrt{\frac{C\|h\|_p\varsigma}{\delta^{d(1+1/p)}n}\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}}} \;\le\; \frac{C''\varsigma}{n\delta^d}\log\delta^{-1} + \sqrt{\frac{C''\|h\|_p\varsigma}{n\delta^{d(1+1/p)}}\log\delta^{-1}}\,.
$$

Moreover, it is easy to check that the assumption

$$
\frac{|\log\delta|}{n\delta^{d/p'}} \le \frac{\|h\|_p}{C''\varsigma}
$$

ensures that

$$
\frac{C''\varsigma}{n\delta^d}\log\delta^{-1} \le \sqrt{\frac{C''\|h\|_p\varsigma}{n\delta^{d(1+1/p)}}\log\delta^{-1}}\,,
$$

and from the latter we conclude that (76) holds for $C' := 2\sqrt{C''}$. The assertion now follows for the constant $C''' := \max\{C, C', C''\}$. □

6.3 Proofs for the KDE-Based Clustering in Section 5

Lemma 6.13. *For all $\delta \in (0, \mathrm{e}^{-1}]$ and $d \ge 1$ we have $\delta^{|\log\delta|}|\log\delta|^{2d-2} \le \delta^{|\log\delta|-d}$.*

Proof of Lemma 6.13: The derivative of the function $h(\delta) := \delta^{-1/2} + \log\delta$ is given by

$$h'(\delta) = \frac{1}{2} \cdot \frac{2\sqrt{\delta} - 1}{\delta^{3/2}}$$

and from this we conclude that h has a global minimum at $\delta = 1/4$. Since $h(1/4) = 2 - 2\log 2 > 0$, we thus find $|\log\delta| = -\log\delta < \delta^{-1/2}$ for all $\delta \in (0, 1]$. The latter yields $|\log\delta|^{2d} < \delta^{-d}$, and since $|\log\delta| \geq 1$ for $\delta \in (0, \mathrm{e}^{-1}]$, we then obtain the assertion. □

Proof of Theorem 5.1: Let us fix a $D \in X^n$ with $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon/2$. By (28) we see that the probability P^n of such a D is not smaller than $1 - e^{-\varsigma}$. We define $\epsilon := \|h\|_\infty \kappa_1(\frac{\sigma}{\delta})$. In the case of $\operatorname{supp} K \subset B_{\|\cdot\|}$ this leads to $\epsilon = 0$ $\delta^{-d}\kappa_\infty(\frac{\sigma}{\delta}) = 0 \leq \rho_0$ as noted after Theorem 4.3. Furthermore, in the case of (14), Lemma 4.2 shows

$$\begin{aligned}\epsilon = \|h\|_\infty \kappa_1(\tfrac{\sigma}{\delta}) \leq \|h\|_\infty \kappa_1(|\log\delta|^2) &\leq cd^2 \operatorname{vol_d} e^{-|\log\delta|^2} |\log\delta|^{2d-2} \\ &\leq cd^2 \operatorname{vol_d} \delta^{|\log\delta| - d} \leq \varepsilon/2\,,\end{aligned}$$

where in the third to last step we used $0 < \delta \leq 1$ and in the second to last step we used Lemma 6.13. In addition, we have

$$\delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta}) \leq c\delta^{-d} e^{-|\log\delta|^2} = c\delta^{|\log\delta| - d} \leq \varepsilon \leq \rho_0\,.$$

Consequently, Theorem 4.3 shows, for all $\rho \geq \rho_0$, that

$$M_{\rho+\varepsilon}^{-2\sigma} \subset L_{D,\rho} \subset M_{\rho-\varepsilon}^{+2\sigma}\,. \tag{77}$$

i). The assertion follows from Theorem 3.2 applied in the case $\tilde{d} := 2\sigma$. Indeed, we have just seen that (47) holds for all $\rho \geq \rho_0$, if we replace δ by $\tilde{\delta}$, and our assumptions guarantee $\tilde{\delta} \in (0, \delta_{\mathrm{thick}}]$, $\rho_0 \geq \rho_*$, and

$$\tau > \psi(\tilde{\delta}) = 3c_{\mathrm{thick}}\tilde{\delta}^\gamma > 2c_{\mathrm{thick}}\tilde{\delta}^\gamma\,.$$

Moreover, (32) is a simple consequence of (77).

ii). Let us check that the remaining assumptions of Theorem 3.1 are also satisfied for $\tilde{\delta} := 2\sigma$, if $\varepsilon^* \leq (\rho^{**} - \rho^*)/9$. Clearly, we have $\tilde{\delta} \in (0, \delta_{\mathrm{thick}}]$, $\varepsilon \in (0, \varepsilon^*]$, and $\psi(\tilde{\delta}) < \tau$. To show $\tau \leq \tau^*(\varepsilon^*)$ we write

$$E := \{\varepsilon' \in (0, \rho^{**} - \rho^*] : \tau^*(\varepsilon') \geq \tau\}.$$

Since we assumed $\varepsilon^* < \infty$, we obtain $E \neq \emptyset$ by the definition of ε^*. There thus exists an $\varepsilon' \in E$ with $\varepsilon' \leq \inf E + \varepsilon \leq \varepsilon^*$. Using the monotonicity of τ^* established in [16, Theorem A.4.2] we then conclude that $\tau \leq \tau^*(\varepsilon') \leq \tau^*(\varepsilon^*)$, and hence all assumptions of [16, Theorem 2.9] are indeed satisfied with δ replaced by $\tilde{\delta}$. The assertions now immediately follow from this theorem. □

Proof of Corollary 5.2: Using Theorem 5.1 the proof of *ii)* is a literal copy of the proof of [16, Theorem 4.1] and the proof of *i)* is an easy adaptation of this proof. □

Proof of Corollary 5.4: Using Theorem 5.1 the proof is a simple combination and adaptation of the proofs of [16, Theorem 4.3] and [16, Corollary 4.4]. □

Proof of Corollary 5.7: Using Theorem 5.1 the proof is a simple combination and adaptation of the proofs of [16, Theorem 4.7] and [16, Corollary 4.8]. □

Proof of Theorem 5.8: The definition of $\varepsilon_{\delta,n}$ in (43) together with $4C_u^2 \log\log n \geq C\|h\|_\infty$ ensures that (30) and (31) are satisfied for all $\delta \in \Delta$. In addition, the assumptions of Theorem 5.8 further ensure that the remaining conditions of Theorem 5.1 are also satisfied. Now the assertion follows by some standard union bound arguments, which are analogous to those of the proof of [16, Theorem 5.1]. □

Acknowledgment

The work of I. Steinwart and P. Thomann for this article was funded by the DFG Grant STE 1074/2-1.

References

[1] O. Bousquet. Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. Ph.D. thesis, Ecole Polytechnique, 2002.

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

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

[4] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer, New York, 2001.

[5] E. Giné and A. Guillou. 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, 2001.

[6] E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 38:907–921, 2002. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

[7] J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, New York, 1975.

[8] J.A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76:388–394, 1981.

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

[10] M. Maier, M. Hein, and U. von Luxburg. Optimal construction of k-nearest neighbor graphs for identifying noisy clusters. *Theoret. Comput. Sci*, 410:1749–1764, 2009.

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

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

[13] A. Rinaldo and L. Wasserman. Generalized density clustering. *Ann. Statist.*, 38:2678–2722, 2010.

[14] B.K. Sriperumbudur and I. Steinwart. Consistency and rates for clustering with DBSCAN. In N. Lawrence and M. Girolami, editors, *JMLR Workshop and Conference Proceedings Volume 22: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics 2012*, pages 1090–1098, 2012.

[15] I. Steinwart. Adaptive density level set clustering. In S. Kakade and U. von Luxburg, editors, *JMLR Workshop and Conference Proceedings Volume 19: Proceedings of the 24th Conference on Learning Theory 2011*, pages 703–738, 2011.

[16] I. Steinwart. Fully adaptive density-based clustering. *Ann. Statist.*, 43:2132–2167, 2015.

[17] I. Steinwart. Supplement A to "Fully adaptive density-based clustering". *Ann. Statist.*, 43, 2015.

[18] I. Steinwart. Supplement B to "Fully adaptive density-based clustering". *Ann. Statist.*, 43, 2015.

[19] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

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

[21] W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *J. Comput. Graph. Statist.*, 19:397–418, 2010.

[22] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

[23] D. Wang, X. Lu, and A. Rinaldo. Optimal rates for cluster tree estimation using kernel density estimators. Technical report, https://arxiv.org/abs/1706.03113, 2017.