Journal of Machine Learning Research 24 (2023) 1-56 Submitted 12/21; Revised 8/23; Published 8/23

Adaptive Clustering Using Kernel Density Estimators

Ingo Steinwart INGO.STEINWART@MATHEMATIK.UNI-STUTTGART.DE
University of Stuttgart
Department of Mathematics
D-70569 Stuttgart, Germany

Bharath K. Sriperumbudur BKS18@PSU.EDU
Pennsylvania State University
Department of Statistics
University Park, PA 16802, USA

Philipp Thomann PHILIPP.THOMANN@D-ONE.AI
D ONE Solutions AG
Sihlfeldstrasse 58
8003 Zürich, Switzerland

Editor: Sivan Sabato

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. In addition, we compare our results to other guarantees found in the literature and also present some experiments comparing our algorithm to k-means and hierarchical clustering.

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 (1975) and is known as *density-based clustering*, assumes i.i.d. data $D = (x_1, \ldots, x_n)$ generated by some unknown distribution P on $X \subset \mathbb{R}^d$. Given some *level* $\rho \geq 0$, the clusters of P are then defined to be the connected components of the level set $\{h \geq \rho\} := \{x \in X : h(x) \geq \rho\}$, where h is the density of P with respect to the Lebesgue measure. This *single level approach* has been studied, for example by Hartigan (1975); Cuevas and Fraiman (1997); Rigollet (2007); Maier et al. (2009); Rinaldo and Wasserman (2010). However, one of the conceptual drawbacks of the single level approach is that different values of ρ may lead to different (numbers of) clusters, see Figure 1, and in addition, there is no general rule for choosing ρ. To address this conceptual shortcoming, one often considers the so-called *cluster tree approach* instead, which tries to consider all

©2023 Ingo Steinwart, Bharath K. Sriperumbudur, and Philipp Thomann.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v24/21-1532.html`.

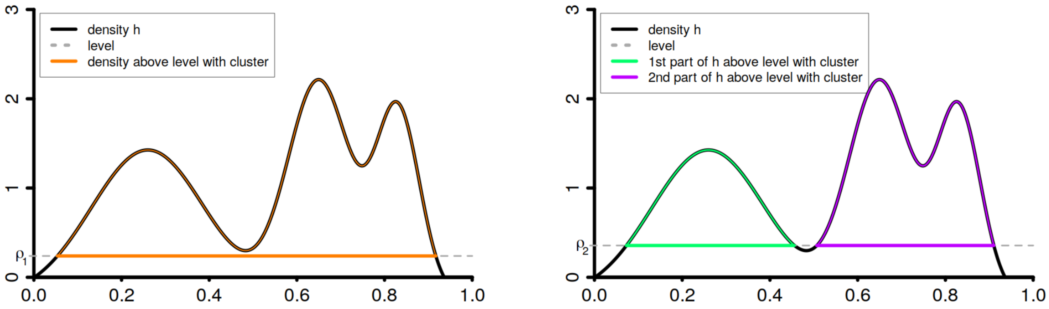

Figure 1: The single level approach applied to a probability density h on $[0, 1]$. **Left.** The level set $\{h \geq \rho_1\}$ displayed by the orange horizontal line consists of one connected component, which results in one cluster for ρ_1. For illustrational purposes, the part of the graph of h that belongs to this level set is also colored in orange. **Right.** For a slightly larger ρ_2, the level set $\{h \geq \rho_2\}$ consists of two connected components, which are indicated by the green and purple horizontal lines. As a result, we obtain two clusters for this level ρ_2.

levels and the corresponding connected components simultaneously, see the left picture in Figure 2 for an illustration.

There exists a variety of articles investigating properties of the cluster tree approach, see e.g. (Hartigan, 1975; Stuetzle, 2003; Chaudhuri and Dasgupta, 2010; Stuetzle and Nugent, 2010; Kpotufe and von Luxburg, 2011; Chaudhuri et al., 2014; Wang et al., 2019) for details. For example, Chaudhuri and Dasgupta (2010) show, under some assumptions on h, that a modified single linkage algorithm recovers this tree in the sense of Hartigan (1981), and Kpotufe and von Luxburg (2011); Chaudhuri et al. (2014) obtain similar results for an underlying k-NN density estimator. In addition, Kpotufe and von Luxburg (2011); Chaudhuri et al. (2014) propose a simple pruning strategy, that removes connected components that artificially occur because of finite sample variability. However, the notion of recovery taken from Hartigan (1981) only focuses on the correct estimation of the cluster tree structure and not on the estimation of the clusters itself, cf. the discussion by Steinwart (2011). Finally, the most recent paper (Wang et al., 2019) 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 smoothness such as α-Hölder continuity.

A third approach taken by Steinwart (2011); Sriperumbudur and Steinwart (2012); Steinwart (2015a) tries to estimate both the *first split* ρ^* *in the cluster tree*, and the corresponding clusters, see the right picture in Figure 2 for an illustration. As in the previously discussed papers, finite sample bounds are derived, which in (Steinwart, 2015a) are extended to learning rates. For example, these learning rates for estimating ρ^* are of the probabilistic form

$$P^n\left(\left\{D \in X^n : 0 \leq \rho^*_D - \rho^* \leq K a_n\right\}\right) \geq 1 - \frac{1}{n},$$

where ρ^*_D is the sample based estimate constructed by the considered algorithm and K is a constant depending on some assumptions on P. In other words, with high probability, the algorithm estimates ρ^* up to precision $K a_n$. The learning rates for estimating the resulting

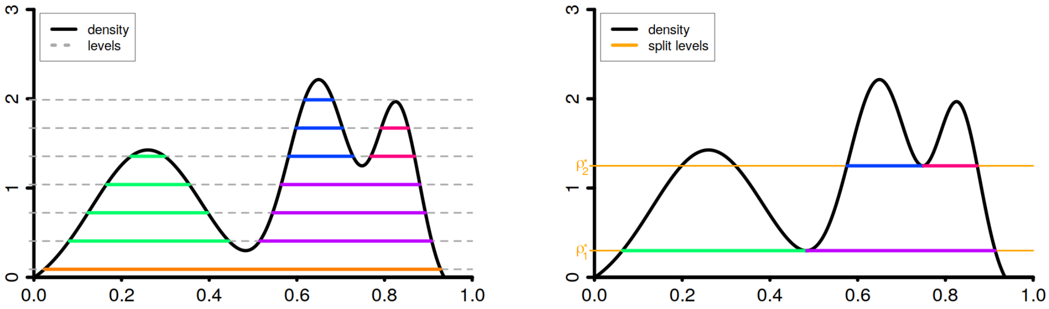

Figure 2: A density h on $[0, 1]$ and different density-based clustering approaches. **Left.** The cluster tree approach considers various, or ideally even all, levels (in grey) simultaneously, making it superfluous to choose one level in advance. However, for a given algorithm the difficulty of detecting connected components may change with the considered level, since, for example, the distances of related connected components, i.e. clusters having the same color, change with the level as it can be seen for the second and third level from below. **Right.** The split tree approach tries to estimate both the levels at which a split in the infinite, ideal cluster tree of h occurs and the resulting clusters at these levels. For the depicted h this leads to the problem of estimating the split levels ρ_1^*, ρ_2^* (drawn in yellow) as well as the resulting clusters for ρ_1^* (drawn in green and purple) and for ρ_2^* (drawn in blue and dark pink). Our previous work, see (Steinwart, 2011; Sriperumbudur and Steinwart, 2012; Steinwart, 2015a) only considered the first split ρ_1^* and its clusters.

clusters are of similar probabilistic nature. Moreover, Steinwart (2015a) shows that these learning rates can also be obtained by an adaptive, fully data-driven hyper-parameter selection strategy. Unfortunately, however, Steinwart (2011, 2015a) only considers the simplest possible density estimator, namely a histogram approach, and Sriperumbudur and Steinwart (2012) restrict their considerations to compactly supported moving window density estimates for α-Hölder-continuous densities. In addition, the method by Sriperumbudur and Steinwart (2012) requires the user to know α, and hence 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, and for *multi cluster* distributions, i.e. for distributions that have more than one split in the cluster tree. As a consequence, it remained unclear whether and how a suitably modified version of this algorithm can be used to estimate the *split-tree*, i.e. the combination of all levels at which a split in the cluster tree occurs together with the resulting clusters at these splits. We refer to Figure 2 for an illustration of such a split tree.

The goal of this paper is to address the discussed issues of Steinwart (2011); Sriperumbudur and Steinwart (2012); Steinwart (2015a). 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

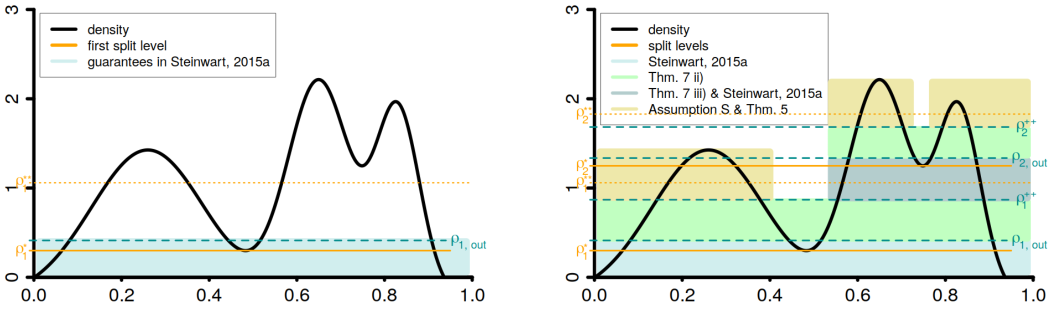

Figure 3: Guarantees provided in previous papers such as Steinwart (2015a) and in this paper. **Left.** Steinwart (2015a) etc. only provide guarantees for estimating the first split level ρ_1^* and the corresponding clusters. In other words only the light blue area below the algorithm's split level estimate $\rho_{1,\text{out}}$ is covered. **Right.** In this paper we provide guarantees for estimating the entire split tree. To be more specific, Part *ii)* of Theorem 9 ensures that the generic cluster algorithm works correctly in the green area between $\rho_{1,\text{out}}$ and a distribution dependent value $\rho_1^{\dagger\dagger}$. On the left hand in the yellow area, Theorem 7 then guarantees, that the cluster algorithm correctly recognizes that there is no further split in the splitting tree. In contrast, on the right hand side, Part *iii)* of Theorem 9 makes it possible to reuse the guarantees of Steinwart (2015a) after some minor technical modifications. As a consequence, it can be ensured in the gray-blue area on the right hand side that the algorithm outputs an estimate $\rho_{2,\text{out}}$ of the second split level ρ_2^* together with estimates for the corresponding clusters. Then in the green area on top of this, Part *ii)* of Theorem 9 again guarantees that the cluster algorithm works correctly up to some distribution dependent $\rho_2^{\dagger\dagger}$. Finally, Theorem 7 ensures in the two yellow areas at the very top that the cluster algorithm correctly recognizes that there is no further split in the cluster tree.

levels ρ close to the maximum $\|h\|_\infty$ of the density h. Such assumptions were missing in (Steinwart, 2011; Sriperumbudur and Steinwart, 2012; Steinwart, 2015a).

ii) We present a simple modification of the output behavior of the generic cluster algorithm of Steinwart (2015a) 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 Steinwart (2015a), we then show that this new cluster algorithm is able to:

a) provide an estimate ρ_{out} of the first split level ρ^* if there is one;

b) correctly detect distributions for which there is no such split level;

c) construct estimates B_i of the clusters A_i^* occurring at the first split level.

Note that from a technical side both *a)* with finite sample bounds on $|\rho_{\text{out}} - \rho^*|$ and *c)* with finite sample bounds on $\lambda^d(B_i \vartriangle A_i^*)$ directly follow from Steinwart (2015a), since our modification of the generic cluster algorithm scans through candidate levels ρ in exactly the same way as the original algorithm of Steinwart (2015a) does. Therefore, the surprising, and compared to (Steinwart, 2015a) new part of our finite sample guarantees 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 by Steinwart (2015a).

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 (Steinwart, 2011; Sriperumbudur and Steinwart, 2012; Steinwart, 2015a), it turns out, that the algorithm behaves correctly until it reaches a level $\rho^{\dagger\dagger}$ for which the connected components of $\{h \geq \rho\}$ are sufficiently apart from each other. Above this level $\rho^{\dagger\dagger}$, we further show that the results of *ii)* can then be recursively applied, leading to guarantees for the entire split-tree. We refer to Figure 3 for a detailed description on how the different guarantees can be combined and how these guarantees differ from our previous work (Steinwart, 2011; Sriperumbudur and Steinwart, 2012; Steinwart, 2015a).

Besides these improvements for the generic cluster algorithm we additionally present the following results:

iv) We show that the new generic cluster algorithm does not only work with an underlying histogram density estimator (HDEs) as in (Steinwart, 2011, 2015a), but also for a variety of kernel density estimators (KDEs). Here it turns out that the results of Steinwart (2015a), 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 Steinwart (2015a), all the results for KDEs are new. Moreover, the results for KDEs substantially extend the results of Sriperumbudur and Steinwart (2012), since there *a)* only moving window kernels were treated, and *b)* only α-Hölder continuous densities with known α were considered. 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 by Wang et al. (2019). The latter improvement is partially made possible, because we can rely on the tools of Steinwart (2015a). However, compared to the HDEs in (Steinwart, 2015a) considering KDEs still requires significant technical efforts such as finite sample bound for the $\|\cdot\|_\infty$-distance between a KDE and its population version.

v) We discuss in some detail the differences of our results and those of Chaudhuri et al. (2014) and Wang et al. (2019), which in some sense are the articles closest to ours.

Here it turns out that, depending on the set of assumptions, sometimes the learning rates for estimating the split levels obtained by Wang et al. (2019) are better and sometimes the ones obtained by us are better. However, adaptivity, for example, is not achieved by Wang et al. (2019). The latter is also true for the paper by Chaudhuri et al. (2014), but in addition their rates are worse than ours, and in addition, their set of assumptions is more restrictive.

vi) We present some experiments comparing our algorithm run with either the moving window kernel or the Epanechnikov kernel to both k-means and hierarchical clustering. Here, it turns out that our algorithm outperforms the latter two as soon as the sample size is sufficiently large. In addition, our algorithm is less sensitive to clusters having different spatial scales.

This paper is organized as follows: In Section 2 we recall the key concepts of Steinwart (2015a). 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. In Section 6 we present the comparison to the articles by Chaudhuri et al. (2014) and Wang et al. (2019), and Section 7 contains the experiments. All proofs can be found in Section 8.

2. Preliminaries

In this section we recall the setup for defining density-based clusters in a general context from Steinwart (2015a). 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 $B_{\|\cdot\|}(x,\delta) := x + \delta B_{\|\cdot\|}$. 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 $\mathrm{vol_d}$. 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 $\delta \geq 0$ we define the δ-tube and δ-trim of A in X by

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

where $d(x,A) := \inf_{x' \in A} \|x - x'\|$. We refer to Figure 4 for an illustration of these concepts as well as to some possible topological changes when going from A to either $A^{+\delta}$ or $A^{-\delta}$. 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. Obviously, we have $A^{+0} = \overline{A}$, and hence also $A^{-0} = \mathring{A}$. Furthermore, since $x \mapsto d(x,A)$ is continuous, $A^{+\delta}$ is always closed in X and $A^{-\delta}$ is always open in X. Note that if A is bounded, we always find compact and convex $X \subset \mathbb{R}^d$ with $A^{+\delta}_X = A^{+\delta}_{\mathbb{R}^d}$ and $A^{-\delta}_X = A^{-\delta}_{\mathbb{R}^d}$. Based on this observation and the fact that we usually consider $\delta \in (0,1]$ in combination with some suitably chosen X, see *Assumption P* below for details, we often ignore the surrounding set X. In addition to this notations, we denote the inradius

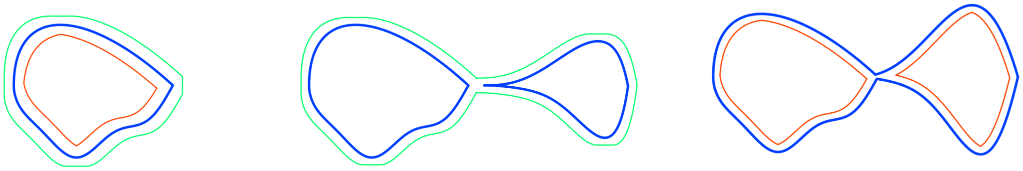

Figure 4: **Left.** A set A in blue together with a δ-tube $A^{+\delta}$ of it in green and a δ-trim $A^{-\delta}$ of it in orange, where in both cases we considered the supremum norm in $\mathbb{R}^2$. All three sets are connected. **Middle.** A set A in blue that consists of two connected components. This time, however, its δ-tube $A^{+\delta}$ is connected. **Right.** A connected set A in blue, for which its δ-trim $A^{-\delta}$ has two connected components. In summary, we see that the connected component structure of A does not necessarily relate to those of $A^{+\delta}$ and $A^{-\delta}$.

and diameter of a bounded $A \subset \mathbb{R}^d$ by inrad A and diam A, respectively, that is

$$\begin{aligned}\operatorname{inrad} A &:= \sup\bigl\{r > 0 : \exists x \in A \text{ with } B(x,r) \subset A\bigr\}\\ \operatorname{diam} A &:= \sup\bigl\{\|x - x'\| : x, x' \in A\bigr\}\,.\end{aligned}$$

Some interesting properties of these quantities can be found in (Steinwart et al., 2021, Lemma 8.1 and Lemma 8.2).

Throughout this work, $\mathbf{1}_A$ denotes the indicator function of a set A, and $A \vartriangle B$ the symmetric difference of two sets A and B. Let us now assume that P is a probability measure 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 one could define the clusters of P to be the connected components of the level set $\{h \geq \rho\}$, where $\rho \geq 0$ 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 (Steinwart, 2015a, Section 2.1). To address this issue, Steinwart (2015a) considered, for $\rho \geq 0$, the measures

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

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

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

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

$$\lambda^d\bigl(\{h \geq \rho\} \setminus M_\rho\bigr) = \lambda^d\bigl(\{h \geq \rho\} \cap (\mathbb{R}^d \setminus M_\rho)\bigr) = \mu_\rho(\mathbb{R}^d \setminus M_\rho) = 0\,, \tag{1}$$

i.e. modulo λ^d-zero sets, the level sets $\{h \geq \rho\}$ are not larger than M_ρ. In fact, M_ρ turns out to be the smallest closed set satisfying (1) and it is shown in (Steinwart, 2015b, Lemma A.1.2), that

$$\{h \overset{\circ}{\geq} \rho\} \subset M_\rho \subset \overline{\{h \geq \rho\}} \qquad \text{and} \qquad M_\rho \vartriangle \{h \geq \rho\} \subset \partial\{h \geq \rho\}\,.$$

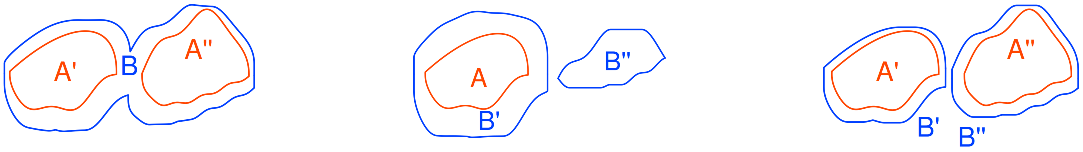

Figure 5: Examples of injective, surjective, and bijective of CRMs, which also illustrate partitions generated by connected components. **Left.** For $A := A' \cup A''$, $\mathcal{P}(A) := \{A', A''\}$, and $\mathcal{P}(B) := \{B\}$, the CRM $\zeta : \mathcal{P}(A) \to \mathcal{P}(B)$ is surjective but not injective. **Middle.** For $B := B' \cup B''$ and the obvious definitions of $\mathcal{P}(A)$ and $\mathcal{P}(B)$, the CRM $\zeta : \mathcal{P}(A) \to \mathcal{P}(B)$ is injective but not surjective. **Right.** Analogously, for $A := A' \cup A''$ and $B := B' \cup B''$, the CRM $\zeta : \mathcal{P}(A) \to \mathcal{P}(B)$ is bijective and hence $\mathcal{P}(A)$ is persistent in $\mathcal{P}(B)$. In this case the structure of the cells in A and B have a clear one-to-one relation to each other. Note that in all cases we further have $\mathcal{P}(A) = \mathcal{C}(A)$ and $\mathcal{P}(B) = \mathcal{C}(B)$.

In order to ensure inclusions that are "inverse" to (1), Steinwart (2015a) said that P is normal at level ρ if there exist two λ^d-densities h_1 and h_2 of P such that

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

In particular, if P is normal at level ρ, then M_ρ and $\{h_1 \geq \rho\}$ only differ by a Lebesgue zero set. Moreover, it is shown in (Steinwart, 2015b, Lemma A.1.3)[1] 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. Furthermore, if P has a λ^d-density h such that $\lambda^d(\partial\{h \geq \rho\}) = 0$, then the same lemma shows that P is normal at level ρ. This implies that essentially all P one usually thinks of are normal. Finally, 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, as mentioned above, we have $\lambda^d(M_\rho \vartriangle \{h \geq \rho\}) = 0$.

The next assumption collects the concepts introduced so far.

Assumption P. We have a λ^d-absolutely continuous, probability measure P that is normal at every level. In addition, $\operatorname{supp} P$ is compact, and $X \subset \mathbb{R}^d$ is a compact and connected set with $(\operatorname{supp} P)^{+2}_{\mathbb{R}^d} \subset X$.

Note that the assumption $(\operatorname{supp} P)^{+2}_{\mathbb{R}^d} \subset X$ ensures both $A^{+\delta}_X = A^{+\delta}_{\mathbb{R}^d}$ and $A^{-\delta}_X = A^{-\delta}_{\mathbb{R}^d}$ for all $\delta \in (0,1]$ and $A \subset \operatorname{supp} P$. In this case, we therefore usually ignore the surrounding X.

Let us now recall the definition of clusters from Steinwart (2015a). We begin with the following definition.

Definition 1 *Let $\mathcal{P}(A)$ and $\mathcal{P}(B)$ be partitions of $A \subset B$ with $A \neq \emptyset$. 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

[1] In this lemma, the term "upper normal at level ρ" means that $\lambda^d(M_\rho \setminus \{h_1 \geq \rho\}) = 0$ 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 $h_2 := \mathrm{d}P/\,\mathrm{d}\lambda^d$.

only if $\mathcal{P}_1$ is finer than $\mathcal{P}_2$. Now assume that $\mathcal{P}(A)$ and $\mathcal{P}(B)$ are two partitions with $\mathcal{P}(A) \sqsubset \mathcal{P}(B)$. Then (Steinwart, 2015b, Lemma A.2.1) shows that there exists a unique map $\zeta : \mathcal{P}(A) \to \mathcal{P}(B)$ with

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

Following Steinwart (2011, 2015a), we call ζ the cell relating map (CRM) between A and B. Moreover, if ζ is bijective, we say that $\mathcal{P}(A)$ is *persistent* in $\mathcal{P}(B)$ and write $\mathcal{P}(A) \sqsubseteq \mathcal{P}(B)$. We refer to Figure 5 for illustrations of persistent and non-persistent partitions.

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 (Steinwart, 2015b, Lemma A.2.3) as well as Figure 5.

Following Steinwart (2015a), 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 $x_1, \ldots, x_n \in A$ such that $x_1 = x$, $x_n = x'$ and $\|x_i - x_{i+1}\| < \tau$ for all $i = 1, \ldots, n-1$. 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 (Steinwart, 2015b, 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 (Steinwart, 2015a, Section 2.2) for details.

Following Steinwart (2015a), we can now describe probability measures that can be clustered.

Definition 2 *Let* Assumption P *be satisfied. Then P can be clustered between $\rho^* \geq 0$ and $\rho^{**} > \rho^*$, if for all $\rho \in [0, \rho^{**}]$, we have $|\mathcal{C}(M_\rho)| \in \{1, 2\}$ and the following two conditions are met:*

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

ii) 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

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

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

$$\tau^*(\varepsilon) := \frac{1}{3} \cdot d\big(\zeta_{\rho^*+\varepsilon}(A_1), \zeta_{\rho^*+\varepsilon}(A_2)\big)\,, \qquad\qquad \varepsilon \in (0, \rho^{**} - \rho^*]. \tag{2}$$

Definition 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 persistent, see the right picture in Figure 6 for an illustration. Thus, the topological structure of M_r between the *split level* ρ^* and an *anchor level* ρ^{**} for ρ^* equals that of $M_{\rho^{**}}$. Consequently we can use the connected components of $M_{\rho^{**}}$ to number the connected

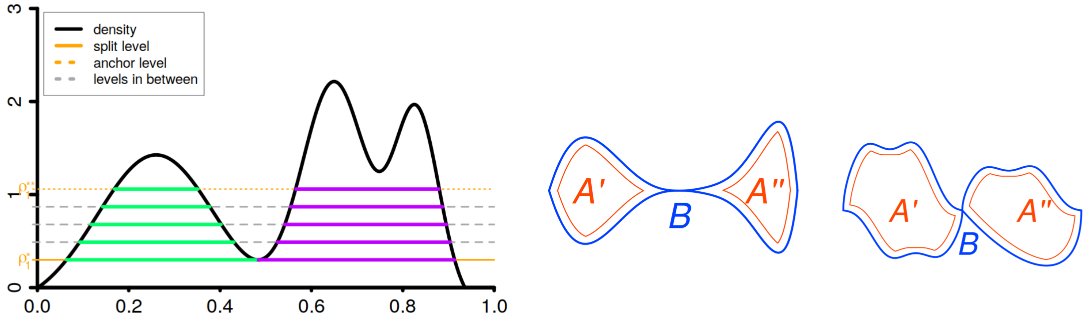

Figure 6: **Left.** A density h on $[0,1]$ that can be clustered between ρ_1^* and ρ_1^{**}. Below the split level ρ_1^*, all level sets M_ρ only have one connected component. Between the split level ρ_1^* and an anchor level ρ_1^{**} of it, that is for all $\rho \in (\rho_1^*, \rho_1^{**}]$, all level sets M_ρ have two connected components (in green and purple). Moreover, we have persistence $\mathcal{C}(M_{\rho_1^{**}}) \sqsubseteq \mathcal{C}(M_\rho)$ as indicated by the same color of the related cells. Taking their unions results in the two clusters A_1^* and A_2^* at the split level ρ_1^*, which are again indicated in green and purple. Finally, note that $\tau^*(\varepsilon)$ equals the distance of the green and purple component at level $\rho^* + \varepsilon$ modulo the factor 3. **Middle.** A connected set B with $M_\rho^{-\delta} = A' \cup A''$. The resulting CRM $\zeta : \mathcal{C}(A) \to \mathcal{C}(B)$ is only surjective and since the bridge in the middle narrows faster than linearly, the thickness function has an exponent $\gamma < 1$. **Right.** Again, a connected set B with $M_\rho^{-\delta} = A' \cup A''$, for which the resulting CRM $\zeta : \mathcal{C}(A) \to \mathcal{C}(B)$ is only surjective. This time, however, the thickness function has an exponent $\gamma = 1$.

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 $\rho^* + \varepsilon$. We again refer to Figure 6 for some visual impressions.

The major goal of Steinwart (2011, 2015a) 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^*. 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 $A \subset \mathbb{R}^d$, Steinwart (2011, 2015a) considered the function $\psi_A^* : (0,\infty) \to [0,\infty]$ defined by

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

Roughly speaking, $\psi_A^*(\delta)$ describes the smallest radius ε needed to "recover" A from $A^{-\delta}$ in the sense of $A \subset (A^{-\delta})^{+\varepsilon}$, see (Steinwart, 2015b, 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 fixed $c \geq 1$, $\gamma \in (0,1]$ and all sufficiently small $\delta > 0$. However, the discussion in (Steinwart, 2015b, Section A.5) also showed that $\gamma = 1$ can be viewed as the most normal case. Finally, we refer to Figure 6 for some examples of linear and nonlinear behavior of ψ^* that is relevant for the following definition taken from Steinwart (2015a).

Definition 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*

$$\psi^*_{M_\rho}(\delta) \leq c_{\text{thick}}\, \delta^\gamma\,.$$

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

Now, the following assumption describes the probability measures we wish to cluster.

Assumption M. The probability measure P can be clustered between ρ^* and ρ^{**}, see Definition 2. 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 (2).

The theory developed by Steinwart (2011); Sriperumbudur and Steinwart (2012); Steinwart (2015a) focused on the question, whether it is possible to estimate ρ^* and the resulting clusters for distributions that can be clustered. To this end, it was assumed that we had an estimation algorithm that constructs, for a given data sets D, level set estimates $L_{D,\rho}$ for all $\rho > 0$. Moreover, it was assumed that these level set estimates satisfy the guarantees

$$M^{-\delta}_{\rho+\varepsilon} \subset L_{D,\rho} \subset M^{+\delta}_{\rho-\varepsilon} \tag{3}$$

for all $\rho \in [0, \rho^{**}]$ and some $\varepsilon, \delta > 0$. Based on these assumptions, a generic cluster algorithm was developed, where "generic" refers to the fact, that the cluster algorithm does not not need to know specifics about the construction of $L_{D,\rho}$. Instead, only the guarantees (3) with known values for ε and δ are needed. The key result (Steinwart, 2015a, 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^*. Here the main difficulty of establishing this result arose from the fact no pair of $M_{\rho+\varepsilon}$, and $M_{\rho-\varepsilon}$, $M^{-\delta}_{\rho+\varepsilon}$, and $M^{+\delta}_{\rho-\varepsilon}$ need to have the same topological structure in terms of their connected components. Indeed, parts of these incompatibilities can already be seen in Figures 1 and 4.

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 probability 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 (Steinwart, 2015a, 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}*, as the following lemma shows.

Lemma 4 *Let $A \subset \mathbb{R}^d$ be bounded and $\delta > 0$. Then $A^{-\delta}_{\mathbb{R}^d} \neq \emptyset$ if and only if $\delta < \operatorname{inrad} A$.*

3. A Generic Algorithm for Estimating the Split-Tree

In this section we 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 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 (Steinwart, 2015a) 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 (Steinwart, 2015a), correctly identifies a split in the cluster tree. Moreover, we demonstrate that, unlike the one in (Steinwart, 2015a), 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 a 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 $|\mathcal{C}(M_\rho)| \leq 1$ as well as:

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

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

Note that $|\mathcal{C}(M_\rho)| \leq 1$ 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 *i)* ensures that the level set M_ρ is still thick in the sense of Definition 3, while in the other case $M_\rho^{-\delta} = \emptyset$, condition *ii)* 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. In this respect note that (Steinwart et al., 2021, Lemma 8.3) shows that for all $\delta \in (0, \delta_{\text{thick}}]$, there exists a $\rho \geq \rho_*$ with $M_\rho^{-\delta} = \emptyset$, and therefore dealing with the case *ii)* cannot be avoided.

In the case $\gamma = 1$, Condition *ii)* can also be interpreted in terms on inradius and diameter as the following lemma shows:

Lemma 5 *For compact $M \subset \mathbb{R}^d$ and $c \geq 1$ the following statements are equivalent:*

i) For all $\delta > 0$ with $M^{-\delta} = \emptyset$ and all $\emptyset \neq A \subset M^{+\delta}$ and $\tau > 2c\delta$, we have $|\mathcal{C}_\tau(A)| = 1$.

ii) We have $2(c-1)\operatorname{inrad} M \geq \operatorname{diam} M$.

Since by Lemma 4 there only exists some $\delta \in (0, \delta_{\text{thick}}]$ with $M_\rho^{-\delta} = \emptyset$ if $\operatorname{inrad} M_\rho \leq \delta_{\text{thick}}$, we thus see that for $\gamma = 1$, Condition *ii)* is equivalent to

$$2(c_{\text{thick}} - 1)\operatorname{inrad} M_\rho \geq \operatorname{diam} M_\rho \tag{4}$$

for all $\rho \geq \rho^*$ with $\operatorname{inrad} M_\rho \leq \delta_{\text{thick}}$. Inequality (4) essentially states that the diameter-inradius ratio must be bounded for increasing ρ. Consequently, (4) is satisfied for $c_{\text{thick}} := 2$ if all M_ρ above ρ_* are balls with respect to the considered norm since in this case we have

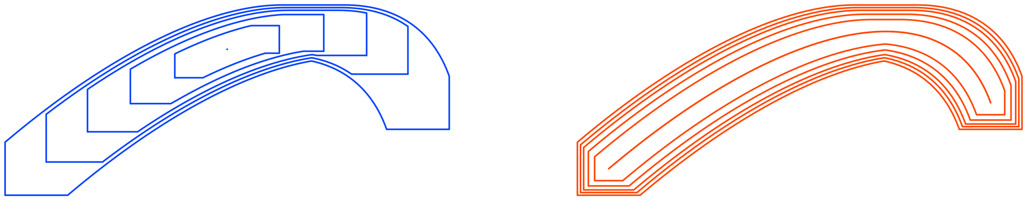

Figure 7: Contour lines of two continuous densities. The levels ρ are chosen such that we have $\operatorname{inrad} M_\rho \in \{0,1,2,3,4,5\}$ with respect to the $\|\cdot\|_\infty$. **Left.** *Assumption S* is satisfied for $\gamma = 1$, since, for increasing ρ, $\operatorname{diam} M_\rho$ decays faster than $\operatorname{inrad} M_\rho$ does, and hence (4) holds. **Right.** *Assumption S* is not satisfied for $\gamma = 1$, since the highest level set satisfies $\operatorname{inrad} M_\rho = 0$ and $\operatorname{diam} M_\rho > 0$, which violates (4).

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

```
**Require:** Some $\varepsilon, \tau > 0$ and a start level $\rho_0 \geq 0$. A decreasing family $(L_\rho)_{\rho \geq 0}$ of subsets of $X$.
**Ensure:** An estimate of $\rho_*$ or $\rho^*$ and the corresponding clusters.
$\rho \leftarrow \rho_0$
repeat
    Identify the $\tau$-connected components $B_1, \dots, B_M$ of $L_\rho$ 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 $\tau$-connected components $B_1, \dots, B_M$ of $L_\rho$ 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 $L_{\rho_0}$.
end if
```

$2 \operatorname{inrad} M_\rho = \operatorname{diam} M_\rho$. Moreover, by increasing c_{thick} we see that (4) remains true if we distort these balls by bi-Lipschitz continuous maps with constants that can be bounded independently of ρ. Analogously, (4) is satisfied, if M_ρ are balls with respect to a *fixed* norm that is different to the one used for the δ-tubes and δ-trims in condition *ii)*. In addition, if we have a flat plateau at the highest level $\rho_{\max} := \sup\{\rho > 0 : M_\rho \neq \emptyset\} < \infty$, that is $\operatorname{inrad} M_{\rho_{\max}} > 0$, then (4) is always satisfied for some $c_{\text{thick}} > 1$ because of $\operatorname{diam} M_\rho \leq \operatorname{diam} X < \infty$. In contrast, (4) is violated, if, for example, $\operatorname{inrad} M_{\rho_{\max}} = 0$ and $\operatorname{diam} M_{\rho_{\max}} > 0$. Finally, for $d = 1$, we always have $2 \operatorname{inrad} M_\rho = \operatorname{diam} M_\rho$ for all non-empty level sets M_ρ, since for these $|\mathcal{C}(M_\rho)| = 1$ ensures that M_ρ is an interval. Consequently, for $d = 1$ *Assumption S* is satisfied with $\gamma = 1$, $c_{\text{thick}} = 2$, and $\delta_{\text{thick}} = 1$ whenever *Assumption P* is satisfied. For two-dimensional examples we refer to Figure 7 for an illustration.

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 also able to detect distributions that only have one cluster 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 by Steinwart (2015a) is the more flexible start level ρ_0, compared to $\rho_0 = 0$ in (Steinwart, 2015a), and the modified output in Lines 8-12. Indeed, the algorithm in (Steinwart, 2015a) 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 7 that the latter case typically occurs for P satisfying *Assumption S*. In this respect recall 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.

The next theorem adapts (Steinwart, 2015a, Theorem 2.9) to Algorithm 1. Since the proof of (Steinwart, 2015a, 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 6.

Theorem 6 *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* (3) *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*

$$\tau - \psi(\delta) < 3\tau^*\bigl(\rho_{\text{out}} - \rho^* + \varepsilon\bigr) .$$

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

$$\sum_{i=1}^{2} \lambda^d\bigl(B_i \vartriangle A_i^*\bigr) \leq 2\sum_{i=1}^{2} \lambda^d\bigl(A_i^* \setminus (A^i_{\rho_{\text{out}}+\varepsilon})^{-\delta}\bigr) + \lambda^d\bigl(M^{+\delta}_{\rho_{\text{out}}-\varepsilon} \setminus \{h > \rho^*\}\bigr) .$$

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 6 shows that Algorithm 1 is still able to estimate ρ^* and the corresponding clusters if the P 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 P that only have one cluster in the sense of *Assumption S*. The next theorem ensures such a behavior.

Theorem 7 *Let* Assumption S *be satisfied and* $(L_\rho)_{\rho \geq 0}$ *be a decreasing family of sets* $L_\rho \subset X$ *such that* (3) *holds for some fixed* $\varepsilon, \delta > 0$ *and all* $\rho \geq \rho_0$. *If* $\rho_0 \geq \rho_*$, $\delta \in (0, \delta_{\text{thick}}]$, *and* $\tau > 2c_{\text{thick}}\delta^\gamma$, *then Algorithm 1 returns* ρ_0 *and* L_0.

Note that Theorem 6 requires $\tau > \psi(d) = 3c_{\text{thick}}\delta^\gamma$, while Theorem 7 even holds under the milder assumption $\tau > 2c_{\text{thick}}\delta^\gamma$. Consequently, if we choose a τ with $\tau > 3c_{\text{thick}}\delta^\gamma$, then the corresponding assumptions of both theorems are satisfied. Moreover, the additional assumption $\tau < \tau^*(\varepsilon^*)$ in Theorem 6 is actually more an assumption on ε^* than on τ as we will see later.

Now assume that the assumptions of Theorem 6 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

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 $\rho_0$ and $(L_\rho)_{\rho\geq 0}$
if $\rho_{\mathrm{out}} > \rho_0$ then
    Store the return values of Algorithm 1 in the split-tree
    Call Algorithm 2 with $\rho_{\mathrm{out}} + \varepsilon$ and $(L_{1,\rho})_{\rho\geq 0}$
    Call Algorithm 2 with $\rho_{\mathrm{out}} + \varepsilon$ and $(L_{2,\rho})_{\rho\geq 0}$
else if $\rho_{\mathrm{out}} = 0$ then
    Store the return values of Algorithm 1 in the split-tree
end if
return split-tree
```

two detected clusters B_1 and B_2 separately, 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, and for this reason, we need to replace (3) 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 $M_{\rho^\dagger}$. We then define two new "children" distributions P_1 and P_2 by

$$P_i(B) := \frac{P(B \cap A_{i,\rho^\dagger})}{P(A_{i,\rho^\dagger})}\,, \qquad\qquad i = 1, 2, \tag{5}$$

for all measurable $B \subset X$. 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. We can now introduce distributions having a finite split tree.

Definition 8 *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 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 $\rho^\dagger := (\rho^{**} + \rho^*)/2$ the two probability measures P_1 and P_2 defined by* (5) *have a finite split-tree with minimal step size $\epsilon > 0$.*

For example, the distribution shown repeatedly in the introduction, see e.g. Figure 2 for the split tree of this distribution, can be clustered. 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 the rather technical Theorem 9 below, which in its formulation requires the sets of τ-connected components of L_ρ that are identified in Lines 4 and 7 of Algorithm 1, that is, the sets

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

Theorem 9 *Let* Assumption M *be satisfied. Furthermore, let* $\varepsilon^* \leq (\rho^{**} - \rho^*)/16$, $\delta \in (0, \delta_{\mathrm{thick}}]$, $\tau \in (\psi(\delta), \tau^*(\varepsilon^*)]$, *and* $\varepsilon \in (0, \varepsilon^*]$, *and* $\rho_0 \leq \rho^*$. *In addition, let* $(L_\rho)_{\rho \geq 0}$ *be a decreasing family satisfying* (3) *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 $\rho \in [\rho_{\mathrm{out}}, \rho^{**} - 3\varepsilon]$ *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*

$$V_i \subset B_i^\rho \subset B_i \, , \qquad i = 1, 2 \, . \tag{6}$$

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*

$$M^{-\delta}_{i,\rho+\varepsilon} \subset L_{i,\rho} \subset M^{+\delta}_{i,\rho-\varepsilon}. \tag{7}$$

To illustrate Theorem 9, 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

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

Then we have, $\rho^{**} - 3\varepsilon > \rho^{**} - 4\varepsilon^* \geq \rho^{**} - (\rho^{**} - \rho^*)/4 = \rho^{\dagger\dagger}$, and part *ii)* of Theorem 9 thus shows (6) for all $\rho \in [\rho_{\mathrm{out}}, \rho^{\dagger\dagger}]$. Consequently, Algorithm 1, when working with the level sets $(L_{i,\rho})_{\rho \in [\rho_{\mathrm{out}}, \rho^{\dagger\dagger}]}$, does identify exactly one connected component in its Line 3. In other words, the loop between its Lines 2 and 5 is not left for such ρ. Moreover, for $\rho \geq \rho^{\dagger\dagger} \geq \rho^\dagger + 4\varepsilon$ part *iii)* of Theorem 9 ensures (7). Consequently, Theorems 6 and 7 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. We refer to Figure 8 for a detailed description of how these guarantees work together. In summary, these considerations show that Algorithm 2 can be recursively analyzed with the help of Theorems 6 and 7 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 $\epsilon > 0$. In particular, for all quantitative guarantees it actually suffices to describe the behavior of Algorithm 1 for distributions satisfying *Assumption S* and *Assumption M.* This insight will be adopted later in the statistical analysis of Section 5.

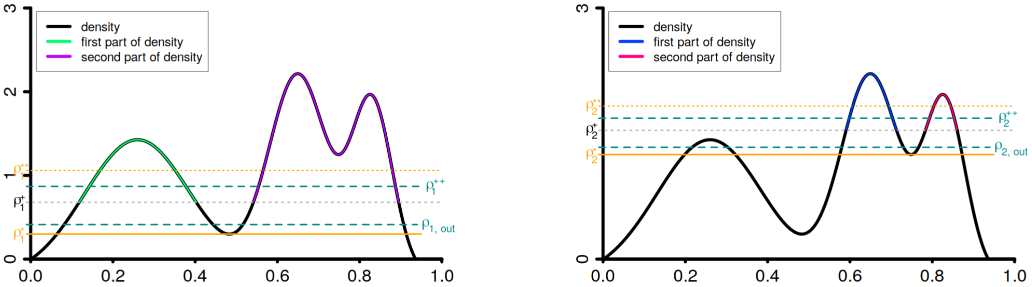

Figure 8: A density of a P that has a finite split-tree with minimal step size $\epsilon > 0$ and the resulting guarantees. **Left.** Situation at the lowest split level ρ_1^*. The distribution P can be clustered between ρ_1^* and ρ_1^{**} and the (unnormalized) densities of the "children" distributions $P_{1,1}$ (green) and $P_{1,2}$ (purple) above $\rho_1^\dagger := (\rho_1^{**} + \rho_1^*)/2$ are shown. *Assumption S* is satisfied by $P_{1,1}$, while $P_{1,2}$ satisfies *Assumption M*. Algorithm 2 initialized with $\rho_0 := 0$ calls Algorithm 1 in its Line 1, which in turn returns $\rho_{1,\text{out}} \approx \rho_1^*$ with $\rho_{1,\text{out}} > \rho_1^*$ and two corresponding clusters denoted by $B_{1,1}$ and $B_{1,2}$ as guaranteed by Theorem 6. Algorithm 2 stores these in its Line 3 and continues by calling a second instance of Algorithm 2 with start level $\rho_{1,\text{out}} + \varepsilon$ and the family $(L_{1,\rho})_{\rho\geq 0}$ in Line 4. This instance of Algorithm 2 in turn calls Algorithm 1 with $\rho_{1,\text{out}} + \varepsilon$ and the family $(L_{1,\rho})_{\rho\geq 0}$. For $\rho \in [\rho_{1,\text{out}}, \rho_1^{\dagger\dagger}]$ with $\rho_1^{\dagger\dagger} := 0.75\rho_1^* + 0.25\rho_1^{**}$, that is, for levels between the two cyan lines, Part *ii)* of Theorem 9 ensures that Algorithm 1 only detects one cluster B_1^ρ in its loop, see also the calculations following Theorem 9. Consequently, this loop is not left below $\rho_1^{\dagger\dagger}$. Moreover, for $\rho \geq \rho_1^{\dagger\dagger}$, Part *iii)* of Theorem 9 ensures $M_{1,\rho+\varepsilon}^{-\delta} \subset L_{1,\rho} \subset M_{1,\rho-\varepsilon}^{+\delta}$, and hence Theorem 7 shows that Algorithm 1 returns its start level $\rho_{1,\text{out}} + \varepsilon$ and $L_{\rho_{1,\text{out}}+\varepsilon}$. As a consequence, both if-clauses of the second instance of Algorithm 2 are not satisfied and the overall program continues at Line 5 of the first instance of Algorithm 2. There, a third instance of Algorithm 2 is called with $\rho_{1,\text{out}} + \varepsilon$ and $(L_{2,\rho})_{\rho\geq 0}$, which in turn begins by calling Algorithm 1 with the same values. Again, Theorem 9 ensures, that for $\rho \in [\rho_{1,\text{out}}, \rho_1^{\dagger\dagger}]$ Algorithm 1 only detects one cluster and that for $\rho \geq \rho_1^{\dagger\dagger}$, the crucial inclusions $M_{2,\rho+\varepsilon}^{-\delta} \subset L_{2,\rho} \subset M_{2,\rho-\varepsilon}^{+\delta}$ are satisfied. Theorem 6 hence ensures that Algorithm 1 returns a $\rho_{2,\text{out}} \approx \rho_2^*$ with $\rho_{2,\text{out}} > \rho_2^* > \rho_{1,\text{out}} + \varepsilon$ and corresponding clusters denoted by $B_{2,1}$ and $B_{2,2}$ to the third instance of Algorithm 2. These values are then stored in the split tree and a fourth and fifth instance of Algorithm 2 are called with $\rho_{2,\text{out}} + \varepsilon$ and the newly defined $(L_{(2,1),\rho})_{\rho\geq 0}$, respectively $(L_{(2,2),\rho})_{\rho\geq 0}$, given by $L_{(2,i),\rho} := L_{2,\rho} \cap B_{2,i}$. **Right.** *Assumption S* is satisfied for both children measures occurring at the split ρ_2^*. As for B_1 on the left image, the combination of Theorem 9 and Theorem 7 ensures that Algorithm 1 called by the fourth and fifth instance of Algorithm 2 with $\rho_{2,\text{out}} + \varepsilon$ and $(L_{(2,i),\rho})_{\rho\geq 0}$ returns these values to these instances of Algorithm 2. Thus, the fourth and fifth instance of Algorithm 2 return to the third instance of Algorithm 2 without any further action, and in turn the third instance returns to the first instance of Algorithm 2. This instance then reaches its Line 9, and therefore the overall program terminates with $(\rho_{1,\text{out}}, B_1, B_2)$ and $(\rho_{2,\text{out}}, B_{2,1}, B_{2,2})$ being stored in its split tree.

4. Uncertainty Control for Kernel Density Estimators

The results of Section 3 provide guarantees as soon as the input level sets satisfy (3). Steinwart (2015a) has shown that guarantees of the form (3) can be established for the level sets of histogram-based density estimators. The goal of this section is to show that (3) can also be established for a variety of kernel density estimators. Our first definition introduces the considered kernels.

Definition 10 *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*

$$\int_{\mathbb{R}^d} K(x)\,\mathrm{d}\lambda^d(x) = 1\,. \tag{8}$$

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

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

We call $\kappa_1(\cdot)$ and $\kappa_\infty(\cdot)$ tail functions*. Finally, we say that K has a* bounded support *if* $\operatorname{supp} K \subset B_{\|\cdot\|}$*, and that K has an* exponential tail behavior*, if there exists a constant $c > 0$ such that*

$$K(x) \le c\exp\bigl(-\|x\|_2\bigr)\,, \qquad\qquad x \in \mathbb{R}^d. \tag{9}$$

Recall that the integrability condition (8) is standard for kernel density estimators. Moreover, kernels of the form $K(x) = k(\|x\|)$ are always symmetric and if the representing $k : [0,\infty) \to [0,\infty)$ 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 $K(x) > 0$ in some neighborhood of 0. In particular, for $k = c\mathbf{1}_{[0,1]}$ we obtain the "rectangular window kernel", which is a symmetric kernel with bounded support, 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.

Given a kernel function K, it is standard in kernel density estimation to consider the modified versions K_δ of K with $\delta \to 0$ for increasing sample sizes n, see for example (Devroye and Lugosi, 2001). Roughly speaking, this is because one can show that the infinite-sample kernel density estimator $h_{P,\delta}$ defined below in (13) converges to the density h for $\delta \to 0$. Also note that for the kernels discussed above and $\delta \in (0,1)$, these versions K_δ are more narrow than the original function K.

Before we proceed with our main goal of establishing (3) let us briefly discuss a couple of simple properties of symmetric kernels K in the sense of Definition 10. To this end, we first note that the properties of the Lebesgue measure λ^d ensure that

$$\int_{\mathbb{R}^d} K_\delta(x-y)\,\mathrm{d}\lambda^d(y) = \int_{\mathbb{R}^d} K(x-y)\,\mathrm{d}\lambda^d(y) = \int_{\mathbb{R}^d} K(y-x)\,\mathrm{d}\lambda^d(y) = 1 \tag{10}$$

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

$$\int_{\mathbb{R}^d \setminus B(x,\sigma)} K_\delta(x-y) \,\mathrm{d}\lambda^d(y) = \int_{\mathbb{R}^d \setminus B(0,\sigma/\delta)} K(y) \,\mathrm{d}\lambda^d(y) = \kappa_1(\tfrac{\sigma}{\delta})\,. \tag{11}$$

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

$$\kappa_1(r) = \kappa_\infty(r) = 0\,, \qquad\qquad r \geq 1\,. \tag{12}$$

Moreover, for kernels with exponential tail, (Steinwart et al., 2021, Lemma 4.2) shows that the behavior of the tail functions can be bounded by $\kappa_1(r) \leq cd^2 \operatorname{vol_d} e^{-r} r^{d-1}$ and $\kappa_\infty(r) \leq ce^{-r}$.

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

$$h_{P,\delta}(x) := \delta^{-d} \int_{\mathbb{R}^d} K\Big(\frac{x-y}{\delta}\Big) \,\mathrm{d}P(y) \qquad\qquad x \in \mathbb{R}^d. \tag{13}$$

It is easy to see that $h_{P,\delta} \geq 0$ 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 (10) yields $\|h_{P,\delta}\|_{L_1(\lambda^d)} = 1$, 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. (Devroye and Lugosi, 2001, 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 (10) yields

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

Clearly, if $D = (x_1, \dots, x_n) \in X^n$ 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

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

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

$$L_{D,\rho} := \{h_{D,\delta} \geq \rho\}\,. \tag{14}$$

While from a statistical perspective, this level set estimator is perfectly fine, it is also 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 (14) 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

$$L_{D,\rho} := \{x \in D : h_{D,\delta}(x) \geq \rho\}^{+\sigma}\,, \tag{15}$$

where $\sigma > 0$. Note that computing connected components of (15) 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 (3) for level set estimates of the form (15).

Theorem 11 *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

$$M^{-2\sigma}_{\rho+\varepsilon+\epsilon} \subset L_{D,\rho} \subset M^{+2\sigma}_{\rho-\varepsilon-\epsilon}\,. \tag{16}$$

Moreover, if P has a bounded density h, then (16) *also holds for $\epsilon = \|h\|_\infty \kappa_1(\frac{\sigma}{\delta})$.*

If K has bounded support for the norm considered in Theorem 11, Equation (12) shows that (16) actually holds for $\epsilon = 0$ and all $\rho \geq 0$ and all $\sigma \geq \delta$. Therefore, we have indeed (3) with δ replaced by 2σ. In general, however, we have an additional horizontal uncertainty ϵ that affects the guarantees of Theorem 6. 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 11 tells us that the uncertainty control (16) is satisfied as soon as we have a data set D with $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon$. Recall that rates for $\|h_{D,\delta} - h_{P,\delta}\|_\infty \to 0$ have already been proven by Giné and Guillou (2002). However, these rates only hold for $n \geq n_0$, where n_0 may depend on D. In addition one needs to choose a sequence (δ_n) of bandwidths a-priori, which excludes adaptivity as we will see below. Finally, the theory developed by Steinwart (2015a) requires bounds of the form $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon(\delta, n, \varsigma)$ that hold with probability not smaller than $1 - e^{-\varsigma}$. Thus, the results of Giné and Guillou (2002) are not suitable for our purposes. To establish suitable bounds, we need to recall some notions first.

Definition 12 *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*

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

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 13 *Let $(Z, \mathcal{A})$ be a measurable space and $\mathcal{G}$ be a set of measurable functions from Z to $\mathbb{R}$ for which there is a $B > 0$ with $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}$. Then $\mathcal{G}$ is called a* uniformly bounded VC-class*, if there are $A > 0$ and $\nu > 0$ such that, for every distribution P on Z we have*

$$\mathcal{N}(\mathcal{G}, L_2(P), \epsilon) \leq \left(\frac{AB}{\epsilon}\right)^{\nu}, \qquad 0 < \epsilon \leq B. \tag{17}$$

Let us briefly look at two important, sufficient criteria for ensuring that the set of functions

$$\mathcal{K}_\delta := \left\{K_\delta(x - \cdot) : x \in X\right\} \tag{18}$$

is a uniformly bounded VC-class. The first such result considers moving window kernels.

Lemma 14 *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* (18) *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 (18) 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 $0 < \varepsilon \leq \mathrm{diam}_{\|\cdot\|}(X)$ we have

$$\mathcal{N}(X, \|\cdot\|, \epsilon) \leq C_{\|\cdot\|}(X)\epsilon^{-d}. \tag{19}$$

Lemma 15 *Let $K : \mathbb{R}^d \to [0, \infty)$ be a symmetric, α-Hölder continuous kernel and $\|\cdot\|$ be a norm 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* (18)*. Then for all $\delta > 0$ with $\delta \leq \left(\frac{|K|_\alpha}{\|K\|_\infty}\right)^{1/\alpha} \mathrm{diam}_{\|\cdot\|}(X)$, all $0 < \epsilon \leq B := \delta^{-d}\|K\|_\infty$, and all distributions P on $\mathbb{R}^d$ we have*

$$\mathcal{N}\big(\mathcal{K}_\delta, L_2(P), \epsilon\big) \leq C_{\|\cdot\|}(X)\left(\frac{|K|_\alpha}{\delta^{\alpha+d}\epsilon}\right)^{d/\alpha}. \tag{20}$$

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

Now, the second main result of this section establishes a finite sample bound on the norm $\|h_{D,\delta} - h_{P,\delta}\|_\infty$. Later we will combine it with Theorem 11.

Theorem 16 *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 (18) *is a uniformly bounded VC-class with constants of the form* $B_\delta = \delta^{-d}\|K\|_\infty$, $A_\delta = A_0\delta^{-a}$, *and* $A_0 > 0, a \geq 0, \nu \geq 1$ *being independent of* δ, *that is,*

$$\mathcal{N}\big(\mathcal{K}_\delta, L_2(Q), \epsilon\big) \leq \left(\frac{A_0\|K\|_\infty \delta^{-(d+a)}}{\epsilon}\right)^\nu \tag{21}$$

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*

$$\delta \leq \min\left\{\delta_0, \frac{4^{p'}\|K\|_\infty}{\|h\|_p^{p'}}, \frac{\|h\|_p^{\frac{1}{2a+d/p'}}}{C}\right\} \qquad \textit{and} \qquad \frac{|\log\delta|}{n\delta^{d/p'}} \leq \frac{\|h\|_p}{C\varsigma} \tag{22}$$

we have

$$P^n\left(\left\{D : \|h_{D,\delta} - h_{P,\delta}\|_{\ell_\infty(X)} < C\sqrt{\frac{\|h\|_p \,|\log\delta|\,\varsigma}{n\delta^{d(1+1/p)}}}\right\}\right) \geq 1 - e^{-\varsigma}. \tag{23}$$

For bounded densities, Theorem 16 recovers the same rates as Giné and Guillou (2002). However, Giné and Guillou (2002) established the rates in an almost sure asymptotic form, whereas Theorem 16 provides a finite sample bound. Moreover, unlike Giné and Guillou (2002), Theorem 16 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 split levels ρ^* and the corresponding clusters. In this respect recall the discussion following Theorem 9, which showed that for deriving guarantees for estimating the split tree with the help of Algorithm 2 it actually suffices to analyze the behavior of Algorithm 1 for distributions satisfying *Assumption S* and *Assumption M*. Following this insight, we will focus on such guarantees for Algorithm 1.

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 discussed by Steinwart et al. (2021).

Theorem 17 *Let* P *be a distribution with bounded Lebesgue density and with* Assumption P *being satisfied. Moreover, let* K *be symmetric kernel with exponential tail behavior, for which the assumptions of Theorem 16 hold. For fixed* $\delta \in (0, \mathrm{e}^{-1}]$ *and* $\tau > 0$, *we choose a* $\sigma > 0$ *with*

$$\sigma \geq \begin{cases} \delta & \textit{if } \operatorname{supp} K \subset B_{\|\cdot\|}\,, \\ \delta\,|\log\delta|^2 & \textit{otherwise.} \end{cases} \tag{24}$$

and we further assume that this σ satisfies both $\sigma \leq \delta_{\text{thick}}/2$ and $\tau > \psi(2\sigma)$. Moreover, for fixed $\varsigma \geq 1$, $n \geq 1$ satisfying the assumptions (22), we pick an $\varepsilon > 0$ satisfying the bound

$$\varepsilon \geq \frac{C}{2}\sqrt{\frac{\|h\|_\infty \,|\log \delta|\, \varsigma}{n\delta^d}}\,, \tag{25}$$

and if K does not have bounded support, also

$$\varepsilon \geq \max\bigl\{1, 2d^2 \operatorname{vol_d}\bigr\} \cdot c \cdot \delta^{|\log \delta| - d}\,. \tag{26}$$

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 (15), 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 with $\hat{M}_{\rho_*} := \bigcup_{\rho > \rho_*} M_\rho$ we have*

$$\lambda^d\bigl(L_{\rho_0} \vartriangle \hat{M}_{\rho_*}\bigr) \leq \lambda^d\bigl(M^{+2\sigma}_{\rho_0 - \varepsilon} \setminus \hat{M}_{\rho_*}\bigr) + \lambda^d\bigl(\hat{M}_{\rho_*} \setminus M^{-2\sigma}_{\rho_0 + \varepsilon}\bigr)\,. \tag{27}$$

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

$$\varepsilon^* \geq \varepsilon + \inf\bigl\{\varepsilon' \in (0, \rho^{**} - \rho^*] : \tau^*(\varepsilon') \geq \tau\bigr\}\,. \tag{28}$$

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 $\rho_{D,\text{out}}$ satisfies both $\rho_{D,\text{out}} \in [\rho^* + 2\varepsilon, \rho^* + \varepsilon^* + 5\varepsilon]$ and

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

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

$$\sum_{i=1}^{2} \lambda^d\bigl(B_i(D) \vartriangle A^*_i\bigr) \leq 2\sum_{i=1}^{2} \lambda^d\bigl(A^*_i \setminus (A^i_{\rho_{D,\text{out}}+\varepsilon})^{-2\sigma}\bigr) + \lambda^d\bigl(M^{+2\sigma}_{\rho_{D,\text{out}}-\varepsilon} \setminus \{h > \rho^*\}\bigr)\,. \tag{29}$$

For our subsequent asymptotic analysis we note that the assumptions $\delta \in (0, \mathrm{e}^{-1}]$ and $\varsigma \geq 1$ of Theorem 17 show that (26) is satisfied if

$$\max\bigl\{1, 2d^2 \operatorname{vol_d}\bigr\} \cdot c \cdot \delta^{|\log \delta| + d/2} \leq \frac{C}{2}\sqrt{\frac{\|h\|_\infty}{n}}\,, \tag{30}$$

and if we choose δ in terms of n, i.e., $\delta = \delta_n$, then (30) 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.

While (24) only provides a lower bound on possible values for σ, Theorem 17 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 (28). Namely, larger values of σ lead to larger τ and thus to larger 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 (9). 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 (26) 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 (26). Arguing as around (30) 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 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 (16) become worse in the horizontal direction, i.e., 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 17 can be used to derive both consistency and rates. We begin with the following result.

Corollary 18 *Let P be a distribution with bounded Lebesgue density and with* Assumption P *being satisfied. Moreover, let K be a symmetric kernel with exponential tail behavior, for which the assumptions of Theorem 16 hold. Let (δ_n) be a positive sequence with $\delta_n \preceq n^{-a}$ for some $a > 0$ and pick a null sequence (σ_n) satisfying* (24) *for all sufficiently large n. Moreover, let (ε_n) and (τ_n) be positive null sequences with $\psi(2\sigma_n) < \tau_n$ for all sufficiently large n, and*

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

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 (15)*, 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 $\lambda^d(\overline{\{h > 0\}} \setminus \{h > 0\}) = 0$ we also have

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

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 $\lambda^d(\overline{A^*_i \cup A^*_2} \setminus (A^*_1 \cup A^*_2)) = 0$, we also have, for $B_1(D)$, $B_2(D)$ as in (29)*:*

$$\lim_{n\to\infty} P^n\Big(\big\{D \in X^n : \lambda^d(B_1(D) \vartriangle A^*_1) + \lambda^d(B_2(D) \vartriangle A^*_2) \leq \epsilon\big\}\Big) = 1\,.$$

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 Steinwart (2015a) that describes how well the clusters are separated above ρ^*.

Definition 19 *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 for all $\varepsilon \in (0, \rho^{**} - \rho^*]$ we have*

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

Moreover, the separation exponent κ is exact*, if there is a $\overline{c}_{\mathrm{sep}} > 0$ such that*

$$\tau^*(\varepsilon) \leq \overline{c}_{\mathrm{sep}}\, \varepsilon^{1/\kappa}\,, \qquad\qquad \varepsilon \in (0, \rho^{**} - \rho^*]\,.$$

The separation exponent describes how fast the connected components of M_ρ approach each other for $\rho \searrow \rho^*$. The "best" separation exponent is $\kappa = \infty$ and in this case 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 17. Indeed, the proof of (Steinwart, 2015a, Theorem 4.3) shows that the value $\varepsilon^* := \varepsilon + (\tau/\underline{c}_{\mathrm{sep}})^\kappa$ satisfies (28) as soon as we have $9\varepsilon^* \leq \rho^{**} - \rho^*$. Consequently, the bound in part *ii) (a)* of Theorem 17 becomes

$$2\varepsilon \leq \rho_{D,\mathrm{out}} - \rho^* \leq 6\varepsilon + \Big(\frac{\tau}{\underline{c}_{\mathrm{sep}}}\Big)^\kappa \tag{31}$$

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 (31) can be improved to

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

as the proof of Theorem (Steinwart, 2015a, Theorem 4.3) shows. To establish rates, it thus suffices to find null sequences (ε_n), (δ_n), (σ_n), and (τ_n) that satisfy (24) and (25), and additionally $\delta_n \in \mathcal{O}(n^{-a})$ for some $a > 0$, if K does not have bounded support. The following corollary presents the resulting rates that are, modulo logarithmic terms, the best ones we can obtain from this approach.

Corollary 20 *Let P be a distribution for which* Assumption M *is satisfied and whose Lebesgue density is bounded. Moreover, consider a symmetric kernel K with exponential tail behavior, for which the assumptions of Theorem 16 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*

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

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

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 (15), 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

$$P^n\biggl(\Bigl\{D \in X^n : 0 \leq \rho_D^* - \rho^* \leq \overline{K}\varepsilon_n\Bigr\}\biggr) \geq 1 - \frac{1}{n}\,.$$

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

$$P^n\biggl(\in X^n : \underline{K}\varepsilon_n \leq \rho_D^* - \rho^* \leq \overline{K}\varepsilon_n\biggr) \geq 1 - \frac{1}{n}\,. \tag{32}$$

Finally, if $\kappa = \infty$ and $\operatorname{supp} K \subset B_{\|\cdot\|}$, then (32) holds for all sufficiently large n, if $\sigma_n = \delta_n$ and

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

Note that the rates obtained in Corollary 20 only differ by the factor $(\log n)^2$ from the rates in (Steinwart, 2015a, 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 (Steinwart, 2015a, Corollary 4.4).

Our next goal is to establish rates for $\lambda^d(B_i(D) \bigtriangleup A_i^*) \to 0$. Since this is a modified level set estimation problem, we need to recall some assumptions used in this context. The first assumption in this direction is one-sided variant of a well-known condition introduced by Polonik (1995).

Definition 21 *Let P be a distribution on $X \subset \mathbb{R}^d$ that has a Lebesgue-density h. For a level $\rho \geq 0$, we say that P has* flatness exponent *$\vartheta \in (0,\infty]$, if there is a $c_{\text{flat}} > 0$ with*

$$\lambda^d\bigl(\{0 < h - \rho < s\}\bigr) \leq (c_{\text{flat}} s)^\vartheta\,, \qquad s > 0.$$

Note that 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.

The second definition describes in some sense the roughness of the boundary of the clusters.

Definition 22 *Let* Assumption M *be satisfied and $\alpha \in (0,1]$. Then the clusters have an* α-smooth boundary*, if there is a $c_{\text{bound}} > 0$ such that, for all $\rho \in (\rho^*, \rho^{**}]$ and $\delta \in (0, \delta_{\text{thick}}]$ we have*

$$\lambda^d\bigl((A_\rho^i)^{+\delta} \setminus (A_\rho^i)^{-\delta}\bigr) \leq c_{\text{bound}}\delta^\alpha\,,$$

where $i \in \{1,2\}$ and A_ρ^1 and A_ρ^2 denote the two connected components of the level set M_ρ.

Note that 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 (Steinwart, 2015b, Lemma A.10.4).

The following assumption collects the different properties of P we have introduced.

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 $\kappa \in (0, \infty]$.

With the help of *Assumption R* we can now establish rates for estimating the clusters, that is, for $\lambda^d(B_i(D) \vartriangle A_i^*) \to 0$.

Corollary 23 *Let* Assumption R *be satisfied and* K *be as in Corollary 20. 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 20. Then there is a constant* $\overline{K} \geq 1$ *such that, for all* $n \geq 1$ *and the ordering as in* (29), *we have*

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

Again, the rates obtained in Corollary 23 only differ by the factor $(\log n)^2$ from the rates in (Steinwart, 2015a, Corollary 4.8). Moreover, if K has a bounded support, then an easy modification of Corollary 23 again yields exactly the same rates as in (Steinwart, 2015a, Corollary 4.8).

Our final goal is to modify the adaptive parameter selection strategy for the histogram-based clustering algorithm of Steinwart (2015a) 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 (24) and $\tau_{\delta,n} \geq 2\psi(2\sigma_{\delta,n})$ are satisfied. In addition, we define

$$\varepsilon_{\delta,n} := C_u \sqrt{\frac{|\log\delta|(\varsigma + \log|\Delta|)\log\log n}{\delta^d n}} + \max\{1, 2d^2 \operatorname{vol_d}\} \cdot c \cdot \delta^{|\log\delta| - d}\,, \tag{33}$$

where $C_u \geq 1$ is some user-specified constant and the second term can 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 (15). Let us consider a width $\delta^*_{D,\Delta} \in \Delta$ that achieves the smallest returned level, i.e.

$$\delta^*_{D,\Delta} \in \arg\min_{\delta \in \Delta} \rho_{D,\delta,\mathrm{out}}\,. \tag{34}$$

In general, this width may not be unique, and hence we assume in the following that we have a well-defined choice, e.g. the smallest $\delta \in \Delta$ satisfying (34). Moreover, we write

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

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 24 *Let P be a distribution for which* Assumption M *is satisfied and whose Lebesgue density is bounded. Moreover, consider a symmetric kernel K with exponential tail behavior, for which the assumptions of Theorem 16 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* (33) *and $\sigma_{\delta,n} > 0$ and $\tau_{\delta,n} > 0$ such that* (24)*, $\tau_{\delta,n} \geq 2\psi(2\sigma_{\delta,n})$, 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* (23) *and $\varepsilon_{\delta,n} + (\tau_{\delta,n}/\underline{c}_{\mathrm{sep}})^\kappa \leq (\rho^{**} - \rho^*)/9$ for all $\delta \in \Delta$. Then we have*

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

Moreover, if the separation exponent κ is exact and $\kappa < \infty$, then we even have

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

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

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

For an adaptive parameter selection strategy, it thus suffices to define appropriate Δ, $\sigma_{\delta,n}$, and $\tau_{\delta,n}$. Here we proceed as in (Steinwart, 2015a, Section 5). Namely, for $n \geq 16$, we consider the interval

$$I_n := \biggl[\Bigl(\frac{\log n \cdot (\log\log n)^2}{n}\Bigr)^{\frac{1}{d}}, \Bigl(\frac{1}{\log\log n}\Bigr)^{\frac{1}{d}}\biggr] \tag{35}$$

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 (24), write $\tau_{\delta,n} := \sigma^\gamma_{\delta,n} \log\log\log n$, and define $\varepsilon_{\delta,n}$ by (33) for all $\delta \in \Delta_n$ and $\varsigma = \log n$. Following the ideas of the proofs of (Steinwart, 2015a, Corollaries 5.2 and 5.3) we then obtain a constant $\overline{K}$ such that for all sufficiently large $n \geq 16$ we have

$$P^n\biggl(D : \rho^*_{D,\Delta_n} - \rho^* \leq \overline{K}\Bigl(\frac{(\log n)^3 \cdot (\log\log n)^2}{n}\Bigr)^{\frac{\gamma\kappa}{2\gamma\kappa+d}}\biggr) \geq 1 - \frac{1}{n}\,. \tag{36}$$

Here, (36) 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 (36) 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

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

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

Finally, we like to mention that the presented adaptive strategy works optimally as long as all split levels have the same exact separation exponent κ. Indeed, as described above, our strategy detects, modulo some logarithmic terms, an asymptotically optimal choice $\delta^*_{D,\Delta}$ for the first split level, and since for all other split levels this choice is also asymptotically optimal as long as they have the same exact separation exponent κ, we see that Algorithm 2 is adaptive when using $\delta^*_{D,\Delta}$.

6. Comparisons

In this section we compare our findings to the most closely related papers, namely (Wang et al., 2019) and (Chaudhuri et al., 2014). In particular, we discuss the different assumptions as well as the obtained statistical guarantees.

To begin with, let us have a look at the assumptions on P, respectively its density h made by Wang et al. (2019). Here, we first note that h is assumed to be α-smooth for some $\alpha > 0$, that is, h is $s := \lfloor \alpha \rfloor$-times continuously differentiable and all partial derivatives of order s are $(\alpha - s)$-Hölder continuous. In general, h does not need to have compact support, instead, Wang et al. (2019) only assume that $\{h \geq \rho\}$ is compact for all $\rho > \rho^*$, where ρ^* is the smallest split level, see their Assumptions **C** and **C'**, respectively. In this respect we note that for P having a continuous density their notion of split levels is closely related to ours and that their Assumption $\mathbf{S}(\kappa)$ equals our separation exponent κ. Finally, Wang et al. (2019) do not impose a thickness assumption, instead a so-called *inner cone condition* is considered. Recall that this cone condition assumes that constants $\varepsilon_I > 0$, $c_I > 0$, and $r_I > 0$ exist such that for all split levels ρ^* and all $\rho \in (\rho^* - \varepsilon_I, \rho^* + \varepsilon_I)$ we have

$$\lambda^d\bigl(B(x,r) \cap \{h \geq \rho\}\bigr) \geq c_I r^d\,, \qquad\qquad x \in \{h \geq \rho\}, r \in (0, r_I]\,.$$

In general, it seems unclear how this condition relates to our thickness assumption, but in many natural situations the inner cone condition and the thickness assumption with $\gamma = 1$ are simultaneously satisfied. For details, we refer to the discussion in (Steinwart, 2015b, Appendix A.5), the rather generic examples considered in (Steinwart, 2015c, Appendix B.2), and the discussion in (Wang et al., 2019, page 15). In the following comparison we therefore assume that the inner cone condition and the thickness assumption with $\gamma = 1$ are simultaneously satisfied. In addition, we assume that h has finitely many split levels.

Like our results, the clustering algorithm of Wang et al. (2019) is also based on a kernel density estimator. However, their central Algorithm 2 is using so-called α-valid kernels, whose KDEs enjoy the ideal approximation error behavior $\|h_{P,\delta} - h\|_\infty \preceq \delta^\alpha$ for $\delta \to 0$ and α-smooth densities h. We refer to (Wang et al., 2019, page 9) for details. Finally, their split level estimator uses a verification strategy that is similar to Line 3 of our Algorithm 1, see (Wang et al., 2019, Definition 6).

To compare the split level guarantee of Wang et al. (2019) to ours, we assume that h is α-smooth, that all split levels have separation exponent α, and that h satisfies the additional assumptions discussed above. Also, we assume that their Algorithm 2 uses an α-valid kernel. Then (Wang et al., 2019, Proposition 3) shows that, modulo some logarithmic factors, all split levels can be estimated with rate $n^{-\frac{\alpha}{2\alpha+d}}$. Since we can choose $\alpha = \kappa$ and $\gamma = 1$, these rate coincide with ours established in Corollary 20 if we again ignore logarithmic factors. Given any $\kappa > 0$, however, our results do not require h to be κ-smooth, while (Wang et al., 2019, Proposition 3) only holds for κ-smooth densities. Moreover note that there are α-smooth densities with $\kappa \gg \alpha$, in fact we may have $\alpha = 1$ and $\kappa = \infty$, and for such densities, our rates are significantly better than those of Wang et al. (2019). In addition, our clustering algorithm does not need to use kernels that are aligned with the smoothness of h, and unlike Wang et al. (2019) we also present a way to make our algorithm adaptive to the unknown κ. Finally, we like to emphasize that apart from Proposition 3, (Wang et al., 2019) contains several other interesting results, which are, however, not comparable to ours.

Unlike the comparison to (Wang et al., 2019), the comparison to (Chaudhuri et al., 2014) turns out to be much more involved, as the latter paper uses assumptions that are quite different to ours. Providing a detailed comparison is therefore beyond the scope of this paper and we refer the interested reader to (Steinwart et al., 2021, Appendix A) for a detailed comparison. To summarize the key points of this comparison, we show that, up to logarithmic factors, the best possible convergence rate achieved by the central Theorem VII.5 of Chaudhuri et al. (2014) for estimating the split levels is $n^{-\frac{\alpha}{3\alpha+d}}$ while our algorithm achieves a rate of $n^{-\frac{\alpha}{2\alpha+d}}$, which is strictly better for all dimensions, with α being the Hölder-continuity of the underlying density. In particular, our rates can be achieved without knowing α, while Chaudhuri et al. (2014) do not offer such adaptivity. In addition, our results can handle discontinuous densities, e.g. step functions on rectangles with mutually positive distances, while their Theorem VII.5 does not provide any guarantee at all for such densities. In particular, the consistency results for their unpruned algorithm require “mild uniform continuity conditions”, see the end of Section III.B in (Chaudhuri et al., 2014), and the guarantees for the pruned algorithm stated in their Theorem VII.5 explicitly require control on the uniform modulus of continuity. Finally, apart from split level estimation, our results also provide guarantees for corresponding clusters in measure while no such guarantees are in (Chaudhuri et al., 2014).

7. Experiments

In this section, we illustrate the behavior of our generic KDE-based clustering algorithm on a few artificial data sets for which the ground truth clustering can be computed. In addition, we compare their performance to k-means and hierarchical clustering.

Data. We consider six cluster problems, which are based on two-dimensional distributions cut down to $[0, 1]^2$, with different degrees of difficulty: The first distribution, see Figure 9, is a mixture of 15 Gaussian distributions and a uniform “background” distribution. In addition, 12 out of the 15 well-separated distributions have a covariance of the form λI_2, where I_2 is the 2×2-identity matrix and λ is some scaling factor, while the remaining 3 Gaussians have a different covariance matrix. Since this distribution was inspired by the S2-data set of Fränti and Virmajoki (2006) we will call it $\mathsf{S2}$ in the following. The second distribution, see Figure 10, is a modification of the first. Namely, two of the 15 clusters have been shrunken and moved towards each other, while the remaining clusters have remained unchanged. In the following we call it $\mathsf{S2-modified}$. The third distribution, called $\mathsf{toy-3G}$, is a mixture of 3 Gaussian distribution with similar but not identical covariances, see Figure 11. Unlike in the first data set, however, the Gaussians are less separated and less concentrated making this distribution slightly more difficult. The fourth distribution,

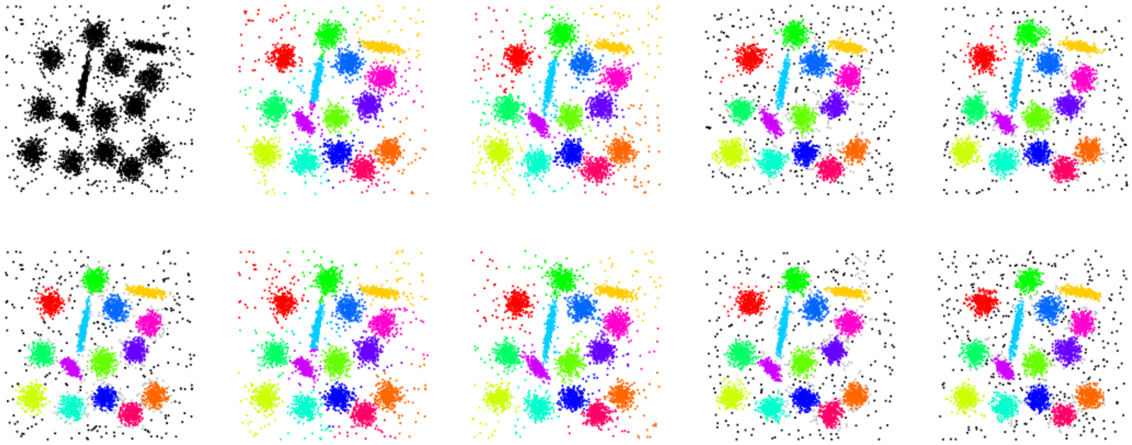

Figure 9: **Left.** A sample data set (top) of size $n = 5000$ from S2 and its ground truth clustering (bottom). **Middle-Left.** The 25^{th} (bottom) and 75^{th} (top) best run of `kmeans` on the S2 data sets. **Middle.** Corresponding 25^{th} (bottom) and 75^{th} (top) best runs of `hclust`. **Middle-Right and Right.** The 25^{th} (bottom) and 75^{th} (top) best runs of our algorithm for the moving window kernel (middle-right) and the Epanechnikov kernel (right). All algorithms perform well.

see Figure 12, is based upon the third distribution. To be more precise, two Gaussians with close-by centers and with small variances have been added. In the following we call it toy − 5G. The fifth distribution, called bananas, is a variant of the classical bananas, or two-moon, data distribution, which is often used to assess the clustering performance on non-centroid-like clusters. Unlike its usual form, however, our variant consists of two very "fuzzy" bananas, which makes even a visual inspection not immediately obvious, see Figure 13. As a result of this noise, both clusters can be almost perfectly separated by a diagonal line, and therefore, it is in principle possible to find a decent clustering with the help of two suitably chosen centroids. The sixth distribution, see Figure 14, is another mixture of Gaussians. In this distribution, however, each of its three clusters corresponds to a mixture of two Gaussians that have the same mean but different variances. In addition, the clusters are "merging" into each other, and as a result, this data set can be viewed as the most challenging one from a visual perspective. In the following, we call it crosses.

For each of the six cluster problems described above and the following sample sizes

$$n \in \{2500, 3000, 3500, 4200, 5000, 6000, 7000, 8200, 10000, 14000, 20000\} .$$

we generated 100 data sets. In addition, we also computed the true densities of the 6 distributions on a 1000×1000 grid of $[0, 1]^2$ to find a high-resolution approximation of the ground truth clustering. Applying these ground true clusters to the data sets, makes it possible to assess how well the different algorithms work.

Performance Measures. Besides visual inspection we consider two different ways of comparing clustering algorithms. To describe these comparisons, let us assume that we have a data set D, ground truth clusters $C_1, \dots, C_k \subset D$, and estimated clusters $\hat{C}_1, \dots, \hat{C}_m \subset D$.

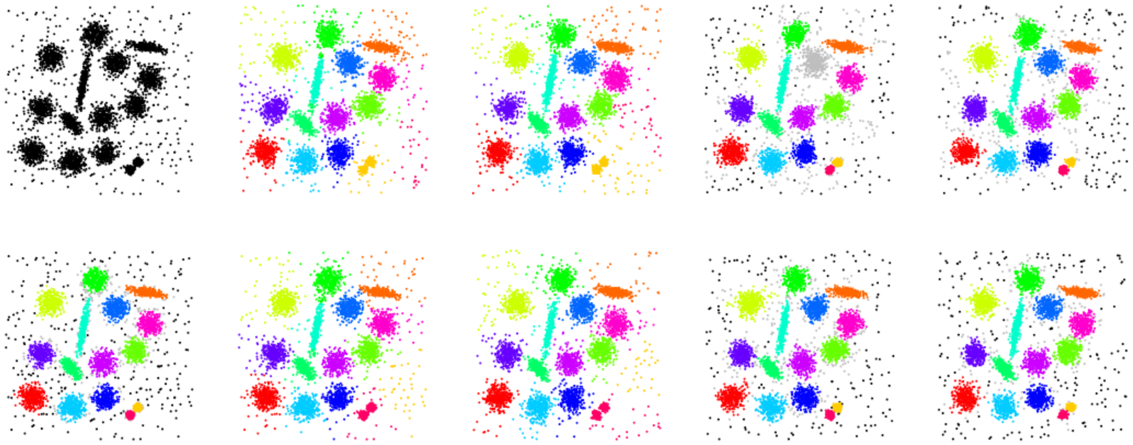

Figure 10: **Left.** A sample data set (top) of size $n = 5000$ from $\mathsf{S2-modified}$ and its ground truth clustering (bottom). **Middle-Left.** The 25^{th} (bottom) and 75^{th} (top) best run of `kmeans` on the $\mathsf{S2-modified}$ data sets. **Middle.** Corresponding 25^{th} (bottom) and 75^{th} (top) best runs of `hclust`. **Middle-Right and Right.** The 25^{th} (bottom) and 75^{th} (top) best runs of our algorithm for the moving window kernel (middle-right) and the Epanechnikov kernel (right). Even in the better runs `kmeans` and `hclust` cannot separate the two modified clusters in the bottom-right, while for our algorithms a problem occurs only for the moving window kernel in the worse run. Namely, in this case one cluster remained undetected.

In all experiments we always observed $m \leq k$. Since k-means and hierarchical clustering produce clusterings with $\hat{C}_1 \cup \cdots \cup \hat{C}_m = D$, while the ground truth clusters do not have this property for our data sets, we restrict the considered data set to the samples that occur in both a true and an estimated cluster, that is

$$D_{\mathrm{perf}} := \left\{ x \in D : x \in \bigcup_{i=1}^{k} C_i \cap \bigcup_{j=1}^{m} \hat{C}_j \right\}.$$

Since in general the jth estimated cluster does not relate to the jth ground truth cluster, we first needed to find a suitable matching, that is an injective map $\Phi : \{1, \dots, m\} \to \{1, \dots, k\}$. To this end, we define the matching error to be

$$\mathcal{E}(\Phi) := \frac{|\{x \in D_{\mathrm{perf}} : x \notin C_{\Phi(j(x))}\}|}{|D_{\mathrm{perf}}|}\,, \tag{37}$$

where for all $x \in D_{\mathrm{perf}}$, we denote by $j(x)$ the unique index j with $x \in \hat{C}_j$. In other words, $\mathcal{E}(\Phi)$ equals the fraction of samples x in D_{perf} for which Φ does not provide a correct matching of clusters $\hat{C}_{j(x)} \to C_{\Phi(j(x))}$. Note that if Φ is a perfect matching in the sense of $\hat{C}_j \cap D_{\mathrm{perf}} = C_{\Phi(j)} \cap D_{\mathrm{perf}}$ for all $j = 1, \dots, m$, then $\mathcal{E}(\Phi) = 0$. For this reason, we determined a matching Φ^* that minimizes $\mathcal{E}(\cdot)$, where we note that even in the case of

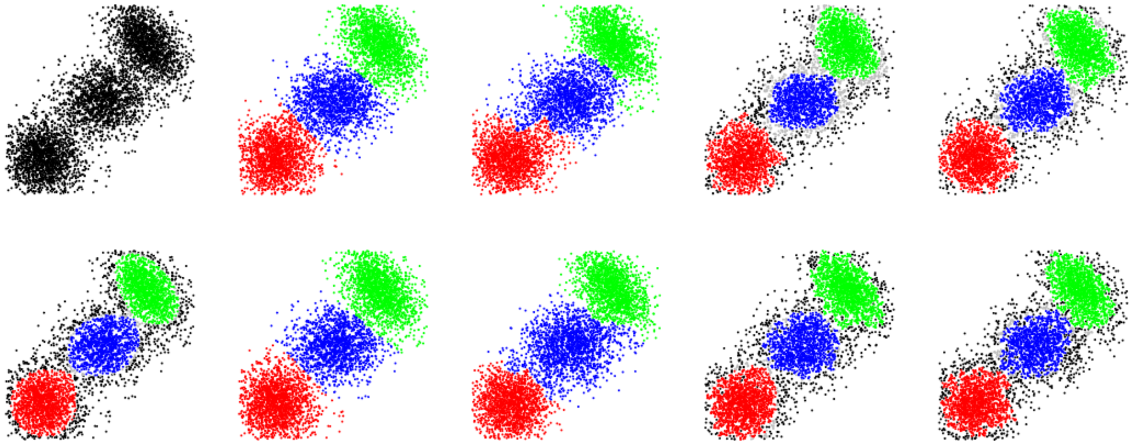

Figure 11: **Left.** A sample data set (top) of size $n = 5000$ from toy$-$3G and its ground truth clustering (bottom). **Middle-Left.** The 25^{th} (bottom) and 75^{th} (top) best run of kmeans on the toy $-$ 3G data sets. **Middle.** Corresponding 25^{th} (bottom) and 75^{th} (top) best runs of hclust. **Middle-Right and Right.** The 25^{th} (bottom) and 75^{th} (top) best runs of our algorithm for the moving window kernel (middle-right) and the Epanechnikov kernel (right). All algorithms perform well.

the first two cluster problems this was computationally feasible with the help of a greedy approach followed by a brute-force calculation over a restricted set of permutations. The corresponding optimal matching errors $\mathcal{E}(\Phi^*)$, averaged over all 100 repetitions of each cluster problem and each considered algorithm, are reported in Figure 15.

At this point let us briefly discuss alternative ways to define the (optimal) matching error. To this end, we define $C_0 := D \setminus (C_1, \ldots, C_k)$ and $\hat{C}_0 := D \setminus (\hat{C}_1, \ldots, \hat{C}_m)$, that is, we have $D_{\mathrm{perf}} = D \setminus (C_0 \cup \hat{C}_0)$. Of course, we could also consider e.g. $D_{\mathrm{perf}} := D \setminus \hat{C}_0$ in the definition of (37) by setting $\Phi(0) := 0$ for all considered maps $\Phi : \{1, \ldots, m\} \to \{1, \ldots, k\}$. In this alternative, all samples $x \in C_0 \setminus \hat{C}_0$ are additionally counted for the matching error. Now kmeans and hclust both assign all samples to some estimated cluster, that is, we have $\hat{C}_0 = \emptyset$. In other words, all samples $x \in C_0$ are automatically counted as an error for both algorithms. This effect is clearly not desirable in a fair comparison, in particular since for some distributions such as bananas and crosses the set C_0 is substantial, see Figures 13 and 14. Conversely, if we consider $D_{\mathrm{perf}} := D \setminus C_0$ then (37) additionally counts all samples $x \in \hat{C}_0 \setminus C_0$. For kmeans and hclust, this set is empty by design, while our algorithms usually satisfy $|\hat{C}_0 \setminus C_0| > 0$. Consequently, we would again have an undesirable effect. The final choice $D_{\mathrm{perf}} := D$ inherits both issues of the previously considered alternatives.

For the second numerical comparison, we again consider Φ^*. For this Φ^*, we counted the number of ground truth clusters not found by the considered algorithm, that is $k - m$, and added the number of non-covering clusters, that is, the number of clusters $\hat{C}_j$ that do not cover at least 50% of the samples in the matched ground truth cluster $C_{\Phi^*(j)} \cap D_{\mathrm{perf}}$.

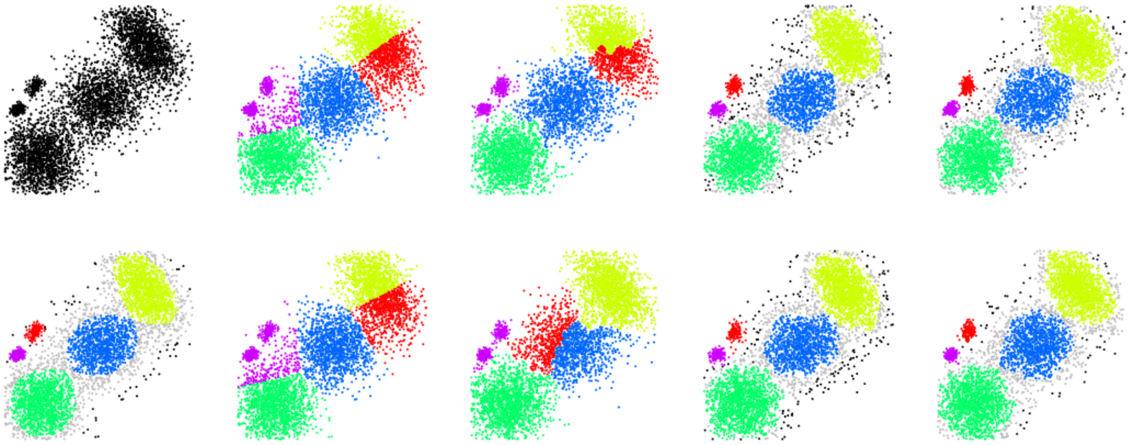

Figure 12: **Left.** A sample data set (top) of size $n = 5000$ from $\texttt{toy} - \texttt{5G}$ and its ground truth clustering (bottom). **Middle-Left.** The 25^{th} (bottom) and 75^{th} (top) best run of `kmeans` on the $\texttt{toy} - \texttt{5G}$ data sets. **Middle.** Corresponding 25^{th} (bottom) and 75^{th} (top) best runs of `hclust`. **Middle-Right and Right.** The 25^{th} (bottom) and 75^{th} (top) best runs of our algorithm for the moving window kernel (middle-right) and the Epanechnikov kernel (right). Both `kmeans` and `hclust` cannot separate the two spatially smaller clusters on the right, and as result, they are forced to cut through one of the spatially larger clusters. In contrast, our algorithms perform well.

The averages over the resulting identification errors

$$\mathcal{I}(\Phi^*) := k - m + \left| \left\{ j : \frac{|\hat{C}_j \cap C_{\Phi(j)} \cap D_{\mathrm{perf}}|}{|C_{\Phi(j)} \cap D_{\mathrm{perf}}|} \leq 1/2 \right\} \right|$$

are reported in Figure 16.

Finally, for visual inspection we ordered the 100 results of each of the algorithms described below on each of the six cluster problems with $n = 5000$ with the help of the joint error $\mathcal{E}(\Phi^*)$. We then plotted the clusterings on the 25^{th} and 75^{th} best run, see Figures 9 to 14. These visualizations help to understand the sources of the typical errors made by the individual algorithms.

Algorithms. We implemented an iterative version of Algorithm 2, in which for $k = 0, 1, \ldots$ we first computed the τ-connected components of $L_{\rho+k\varepsilon}$ that do not vanish at level $L_{\rho+(k+2)\varepsilon}$, see also Line 3 of Algorithm 1. With the help of these connected components we then generated the corresponding cluster tree estimate, where we note that we skipped the Line 7 of Algorithm 1 since this line has only been inserted into Algorithm 1 to simplify the form of the statistical guarantees. We then called the resulting cluster tree estimator for different KDE level set estimators, that is, for different values of δ. The first split in the cluster tree was then obtained by choosing the δ^* that resulted in the smallest first split level. If all split levels have the same separation exponent κ, then it asymptotically suffices to consider this δ^* for the entire cluster tree. In general, however, splits further

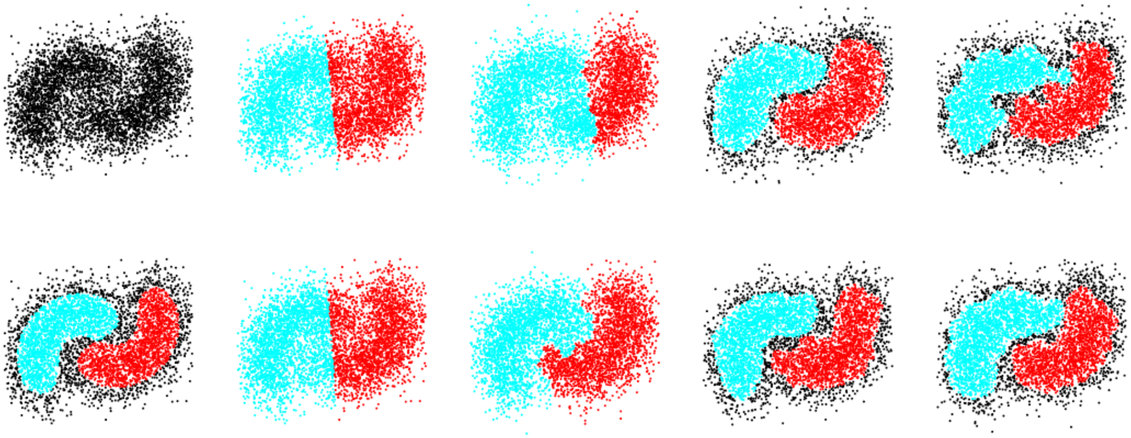

Figure 13: **Left.** A sample data set (top) of size $n = 5000$ from `bananas` and its ground truth clustering (bottom). **Middle-Left.** The 25^{th} (bottom) and 75^{th} (top) best run of `kmeans` on the `bananas` data sets. **Middle.** Corresponding 25^{th} (bottom) and 75^{th} (top) best runs of `hclust`. **Middle-Right and Right.** The 25^{th} (bottom) and 75^{th} (top) best runs of our algorithm for the moving window kernel (middle-right) and the Epanechnikov kernel (right). While the implicit bias of `kmeans` seems to consistently direct it towards imperfect centroids, the performance of `hclust` seems to be rather sensitive to the instance of the data set. In comparison, our algorithms are less sensitive and perform sufficiently well even in their worse runs.

up in the tree may have either a different κ or at least a different constant $\underline{c}_{\mathrm{sep}}$. In these cases, choosing a different δ at different splits may be beneficial. We adopted this insight by considering, after each detected split level ρ, those δ whose estimated cluster tree has another split level within one of the clusters emerging at ρ. Among those δ we then again choose the one resulting in the smallest next split level.

We considered 500 geometrically spaced candidate values of δ between $c(\ln(n)/n)^{1/d}$ and $c(\ln n)^{-1/d}$, where in the experiments, the factor c was determined by an estimate of the median mutual distance between the samples of the considered data set. Notice that modulo this factor c and some (double) logarithmic terms, this setup coincides with the theoretically derived one, see (35). Moreover, we considered both a plain moving window kernel and the Epanechnikov kernel, where in both cases the underlying norm was the Euclidean distance. Since both kernels have bounded support, we simply chose $\sigma := \delta$, see (24), and $\varepsilon := 3\sqrt{\|h_{D,\delta}\|_\infty n^{-1}\delta^{-d}}$ for each candidate value δ. Modulo some logarithmic terms, this choice for ε follows our theoretical insights, see (33). Finally, we decided to focus on thickness guarantees with the most natural choice $\gamma := 1$, see the detailed discussion in (Steinwart, 2015b, Appendix A.5), that is, we do not expect the algorithm to correctly keep clusters together that have thinner cusps or bridges. Based on this decision, we choose $\tau := (2 + \epsilon) \cdot \delta$ with $\epsilon = 0.00001$, where we note that our theoretical findings actually hold

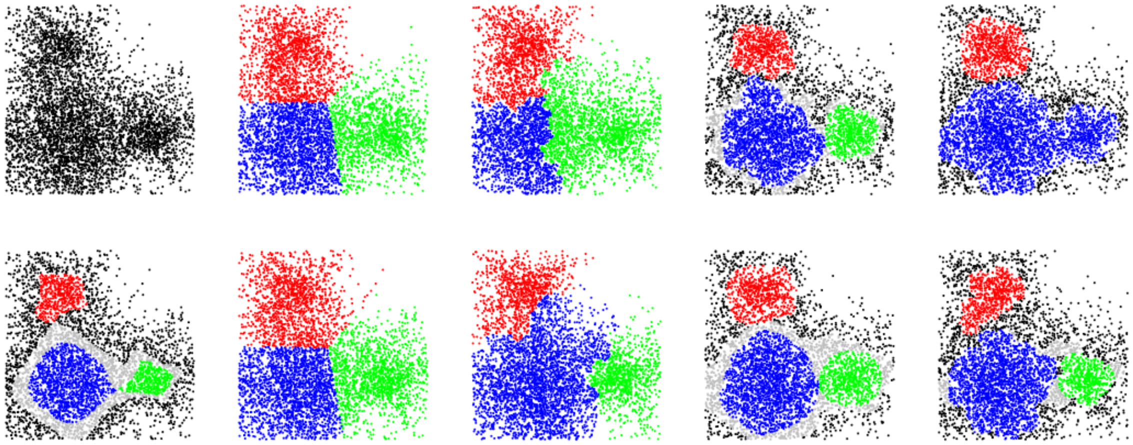

Figure 14: **Left.** A sample data set (top) of size $n = 5000$ from crosses and its ground truth clustering (bottom). **Middle-Left.** The 25^{th} (bottom) and 75^{th} (top) best run of `kmeans` on the crosses data sets. **Middle.** Corresponding 25^{th} (bottom) and 75^{th} (top) best runs of `hclust`. **Middle-Right and Right.** The 25^{th} (bottom) and 75^{th} (top) best runs of our algorithm for the moving window kernel (middle-right) and the Epanechnikov kernel (right). While the implicit bias of `kmeans` seems to consistently direct it towards centroids with sufficiently benign Voronoi partitions, the behavior of `hclust` seems to be rather sensitive to the instance of the data set. In comparison, our algorithm with the moving window kernel is less sensitive and performs sufficiently well even in its worse runs, while the Epanechnikov kernel performs less reliably.

true for each value $\tau > 2\delta$, if one carefully tracks all constants. In addition, this choice makes it possible that the estimated clusters can be as close as $\epsilon \cdot \delta$ to each other.

Besides our methods we also considered k-means and hierarchical clustering. To this end, we used the functions `kmeans`, `kmeans++`, and `hclust` of `R`. Both types of algorithms have some hyper-parameters, which ideally would be chosen in a data-driven approach. To ensure fairness for `kmeans`, `kmeans++`, and `hclust`, such data-driven approaches would need to be calibrated to at least one of our two performance measures $\mathcal{E}(\Phi^*)$ or $\mathcal{I}(\Phi^*)$. However, we are not aware of any method to reliably estimate these performance measures, or some calibrated surrogates for them, from data alone and hence we proceeded differently. Namely, `kmeans`, `kmeans++`, and `hclust` received the correct number k of ground truth clusters as an input parameter, and `kmeans` was repeated with 100 random initializations using the parameter `nstart = 100`. In addition, we first considered the 3 different "versions" of the `kmeans` function on our data sets and identified the best one with the help of our performance metrics. Here it turned out that all three "versions" led to very similar results with `Lloyd` having a marginal advantage over the over two. In addition, all three versions produced better results than `kmeans++`, probably because we used 100 random initializations for `kmeans`. Similarly, we considered the 8 different "versions" of `hclust`. It turned out

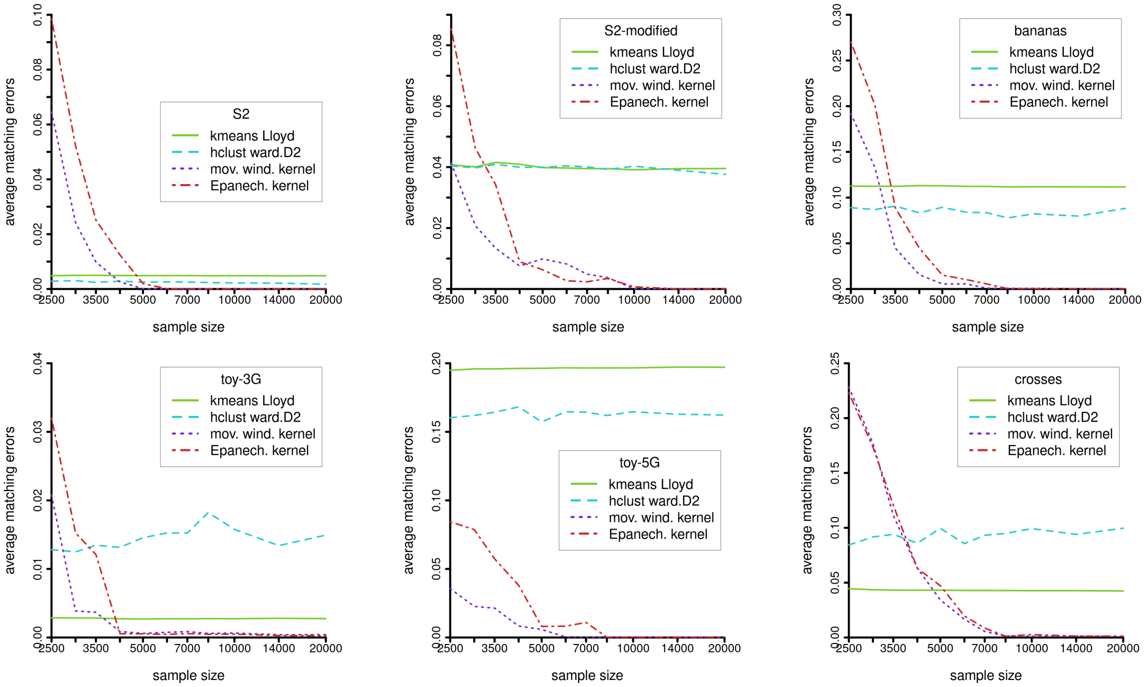

Figure 15: Average matching errors $\mathcal{E}(\Phi^*)$ for optimal matching Φ^*, see (37) for the different cluster algorithms and for different sample sizes. While `kmeans` and `hclust` are essentially not influenced by the sample size, our methods clearly become better with increasing sample size n.

that `ward.D` and `ward.D2` performed best with a slight advantage of `ward.D2`, which we then used in the comparisons.

In the experiments both `kmeans` and `hclust` were thus privileged in two ways: a) the correct number of clusters were given to them, while our algorithms had no information at all about the cluster problems at hand; and b) the best performing version of `kmeans` and `hclust` were chosen in hindsight, and with respect to our ground-truth performance measures. In contrast, our algorithms had to choose their hyper-parameters in a fully adaptive way and without access to ground truth performance measures.

Results. Figures 15 and 16 clearly show that the performance of `kmeans` and `hclust` is almost independent of the data set size n, while our two algorithms heavily depend on n. For example, on five of the six cluster problems, namely on all but $\mathtt{toy} - \mathtt{5G}$, the matching errors achieved by our algorithms for $n = 2500$ are substantially worse than those of `kmeans` and `hclust` and the same is true for the identification errors. For $n = 5000$, however, the matching errors of our algorithms are at least as good as those of `kmeans` and `hclust` and on at least four of cluster problems our algorithms clearly outperform `kmeans` and `hclust`. For the identification errors the situation is still mixed, but this is not really surprising, as the first term $k - m$ in the computation of $\mathcal{I}(\Phi^*)$ vanishes for `kmeans` and `hclust` since the privileged information given to these two algorithms always ensures $k = m$. Despite this advantage, our algorithms always achieve identification errors that are as least as good

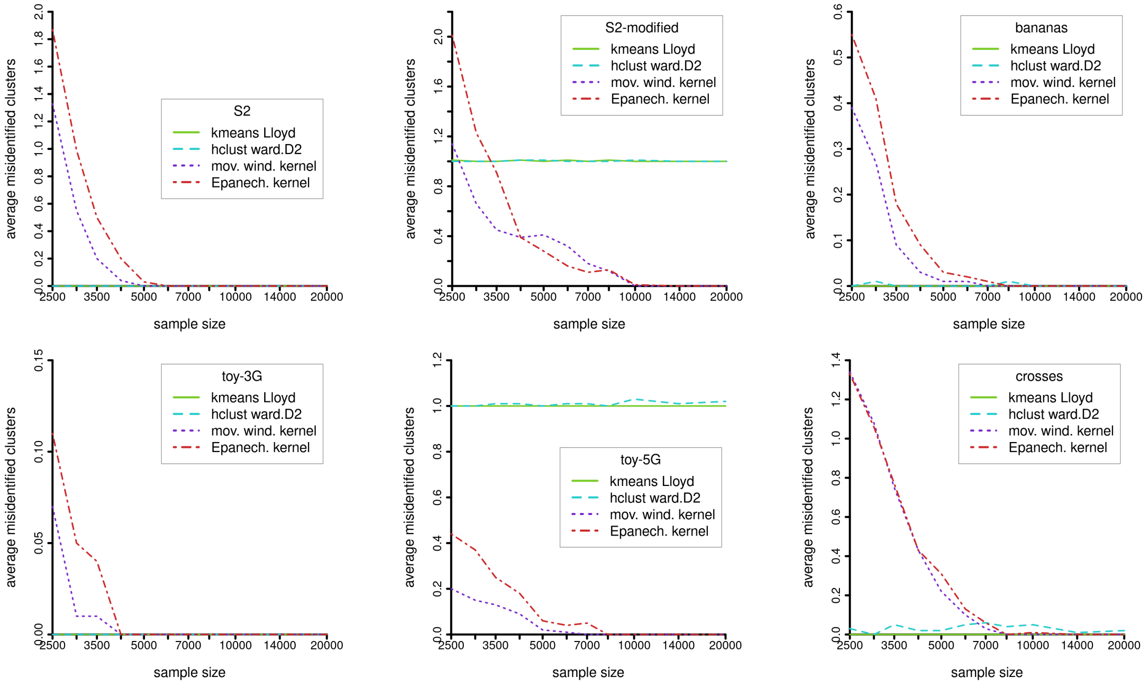

Figure 16: Average identification errors $\mathcal{I}(\Phi^*)$ for optimal matching Φ^*. Both kmeans and hclust have difficulties on the two modified data sets. Our methods become better with increasing sample sizes and in all cases, nor errors are made for $n \geq 10000$.

as those of kmeans and hclust as soon as we have $n \geq 8.200$. Finally, for $n \geq 10000$ our algorithms work almost error free with respect to both performance measures on all six cluster problems.

In terms of identification errors, both kmeans and hclust have substantial difficulties with the two modifications $\mathsf{S2-modified}$ and $\mathsf{toy-5G}$, while on the remaining four cluster problems they perform almost flawless. A closer visual inspection reveals, that both cluster algorithms cannot identify the two spatially smaller, close-by clusters in the modifications, see Figures 10 and 12 and this phenomenon occurs at least on most of the data sets of e.g., size $n = 5000$. In comparison, our algorithm with the moving window kernel has some problems identifying one cluster in $\mathsf{S2-modified}$, see Figure 10, but this phenomenon does not occur as often as the the problems of kmeans and hclust. The same observation can be made for our algorithm working with the Epanechnikov kernel on crosses.

Unsurprisingly, the poor identification error performance of kmeans and hclust on the two modified cluster problems $\mathsf{S2-modified}$ and $\mathsf{toy-5G}$ directly translate into poor matching errors. In addition, both algorithms have problems on bananas and on crosses, see Figure 15. The reasons for these problems are different: While kmeans constantly choose imperfect centroids on the bananas data sets, see Figure 13, hclust seems to be sensitive to the particular instance of the data set. The same behavior of both algorithms can be observed on crosses, see Figure 14. In comparison, a poor matching error performance of our algorithms seem to be directly related to the identification errors made in some cases.

8. Proofs

Proof of Lemma 4: Let us first assume that $A^{-\delta} \neq \emptyset$. Then there exists an $x \in A^{-\delta}$ and since $A^{-\delta}$ is open there also exists an $\varepsilon > 0$ with $B(x, \varepsilon) \subset A^{-\delta}$. Our first intermediate goal is to show

$$B(y, \delta) \subset A \,, \qquad\qquad \text{for all } y \in B(x, \varepsilon)\,. \tag{38}$$

To this end, we note that $y \in B(x, \varepsilon) \subset A^{-\delta}$ implies $y \notin (\mathbb{R}^d \setminus A)^{+\delta}$. For $z \in \mathbb{R}^d \setminus A$ we thus find $d(y, z) \geq d(y, \mathbb{R}^d \setminus A) > \delta$, which shows (38).

Now observe that for the proof of $\delta < \operatorname{inrad} A$ it clearly suffices to establish

$$B(x, \delta + \varepsilon) \subset A\,. \tag{39}$$

Let us therefore fix a $z \in B(x, \delta+\varepsilon)$. By considering $y := x$ in (38) we first note that in the case $\|x - z\| \leq \delta$ we have $z \in A$. Hence it suffices to consider the case $\delta < \|x - z\| \leq \delta + \varepsilon$. For $t := \delta/\|x - z\| \in (0, 1)$ and $y := (1 - t)z + tx$ a quick calculation then shows both $\|y - z\| = \delta$ and $\|x - y\| = \|x - z\| - \delta \leq \varepsilon$. Applying (38) then yields (39).

Let us now assume that $\delta < \operatorname{inrad} A$. Then there exists an $\varepsilon > 0$ with $\delta+\varepsilon < \operatorname{inrad} A$ and hence we find an $x \in A$ with $B(x, \delta + \varepsilon) \subset A$. Clearly, it suffice to show that $x \in A^{-\delta}$. To this end, we assume that $x \notin A^{-\delta}$, that is $x \in (\mathbb{R}^d \setminus A)^{+\delta}$. Since this means $d(x, \mathbb{R}^d \setminus A) \leq \delta$, we then find a $y \in \mathbb{R}^d \setminus A$ with $d(x, y) \leq \delta + \varepsilon$. This contradicts $B(x, \delta + \varepsilon) \subset A$. ■

8.1 Proofs for the Generic Algorithm in Section 3

Proof of Lemma 5: *i) ⇒ ii).* We choose a $\tau > 2c \operatorname{inrad} M$ and $\delta := \operatorname{inrad} M$. By Lemma 4 we then know $M^{-\delta} = \emptyset$, and our construction also gives $\tau > 2c\delta$. If $|M^{+\delta}| = 1$ we obviously have $\operatorname{diam} M^{+\delta} = 0 < \tau$. Moreover, if $|M^{+\delta}| > 1$, there exist $x, y \in M^{+\delta}$ with $\|x - y\| = \operatorname{diam} M^{+\delta}$ since $M^{+\delta}$ is compact and $\|\cdot - \cdot\| : M^{+\delta} \times M^{+\delta} \to \mathbb{R}$ is continuous. For $A := \{x, y\} \subset M^{+\delta}$ we then have $|\mathcal{C}_\tau(A)| = 1$, which shows $\|x - y\| < \tau$. In summary, we therefore always have $\operatorname{diam} M^{+\delta} < \tau$.

Now a simple calculation shows $\operatorname{diam} M^{+\delta} = 2\delta + \operatorname{diam} M$, see also (Steinwart et al., 2021, Lemma 8.2), and thus we obtain $\tau > 2 \operatorname{inrad} M + \operatorname{diam} M$. Since we initially chose $\tau > 2c \operatorname{inrad} M$ arbitrarily, we conclude that $2c \operatorname{inrad} M \geq 2 \operatorname{inrad} M + \operatorname{diam} M$.

ii) ⇒ i). Let us fix a $\delta > 0$ with $M^{-\delta} = \emptyset$. Lemma 4 then shows $\delta \geq \operatorname{inrad} M$. We know fix a non-empty $A \subset M^{+\delta}$ and some $\tau > 2c\delta$. This yields

$$\tau > 2(c - 1)\delta + 2\delta \geq 2(c - 1) \operatorname{inrad} M + 2\delta \geq \operatorname{diam} M + 2\delta = \operatorname{diam} M^{+\delta}\,.$$

For $x, y \in A$ we thus find $\|x - y\| \leq \operatorname{diam} A \leq \operatorname{diam} M^{+\delta} < \tau$, and hence A is τ-connected, that is $|\mathcal{C}_\tau(A)| = 1$. ■

Theorem 25 *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*

$$M^{-\delta}_{\rho+\varepsilon} \subset L_\rho \subset M^{+\delta}_{\rho-\varepsilon} \tag{40}$$

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

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

$$\mathcal{C}_\tau(L_\rho) = \zeta\bigl(\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon})\bigr) \cup \bigl\{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \emptyset\bigr\}\,. \tag{41}$$

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

$$\bigl|\bigl\{B \in \mathcal{C}_\tau(L_\rho) : B \cap L_{\rho+2\varepsilon} \neq \emptyset\bigr\}\bigr| \leq 1\,, \qquad\qquad \tau > 2c_{\text{thick}}\delta^\gamma\,. \tag{42}$$

Proof of Theorem 25: *i).* We first note that $M^{-\delta}_{\rho+3\varepsilon} \neq \emptyset$ implies $M^{-\delta}_{\rho+\varepsilon} \neq \emptyset$. Now, (Steinwart, 2015b, Lemma A.4.3) showed that for all bounded $A \subset \mathbb{R}^d$ with $|\mathcal{C}(A)| < \infty$, all $\delta > 0$ with $A^{-\delta} \neq \emptyset$, and all $\tau > 2\psi^*_A(\delta)$ we have $|\mathcal{C}_\tau(A^{-\delta})| \leq |\mathcal{C}(A)|$, Thus we 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 (41) we now write $A := M^{-\delta}_{\rho+\varepsilon}$ and $B := \zeta(A)$. Our first intermediate goal is to establish the following *disjoint* union:

$$\mathcal{C}_\tau(L_\rho) = \{B\} \cup \bigl\{B' \in \mathcal{C}_\tau(L_\rho) \setminus \{B\} : B' \cap L_{\rho+2\varepsilon} \neq \emptyset\bigr\} \cup \bigl\{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \emptyset\bigr\}\,. \tag{43}$$

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

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

and thus $\{B' \in \mathcal{C}_\tau(L_\rho) \setminus \{B\} : B' \cap L_{\rho+2\varepsilon} = \emptyset\} = \{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \emptyset\}$. This gives (43).

Let us now show (41). 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 (43) 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 (Steinwart, 2015b, 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 hence $\delta \leq \psi^*_{M_{\rho+\varepsilon}}(\delta)$, see (Steinwart et al., 2021, Lemma 8.5), gives

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

However, $B' \neq B$ implies $d(B', B'') \geq \tau$ by (Steinwart, 2015b, Lemma A.2.4), i.e. we have found a contradiction.

ii). 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 (42) 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 (Steinwart, 2015b, 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 thus $\zeta(A_i) \cap B_i \neq \emptyset$ 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 (Steinwart, 2015b, Lemma A.2.4) that $d(A_1, A_2) \geq \tau$, we conclude by (Steinwart, 2015b, 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 $\delta \in (0, \delta_{\text{thick}}]$ and $\tau > 2c_{\text{thick}}\delta^\gamma$ ensures $|\mathcal{C}_\tau(A)| = 1$. Since this contradicts $|\mathcal{C}_\tau(A)| = 2$ we have proven (42). ■

Proof of Theorem 7: 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 by (Steinwart et al., 2021, Lemma 8.3). For $i = 0, \dots, i^*$, part *i)* of Theorem 25 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 25 shows that Algorithm 1 identifies at most one component in Line 7. ■

Lemma 26 *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 $\delta \in (0, \delta_{\text{thick}}]$ *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 26: To adapt to the notation of Steinwart (2015a,b) 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 (Steinwart, 2015b, Proposition A.2.10).

ii). The assertion has been shown in part *iii)* of (Steinwart, 2015b, Lemma A.4.1).

iii). We first note that using part *ii)* of (Steinwart, 2015a, Theorem 2.7) with $\varepsilon^* := \varepsilon$ and $\rho = \rho^* + \varepsilon^*$ we find $|\mathcal{C}_\tau(M_\rho^{-\delta})| = |\mathcal{C}(M_\rho)| = 2$. Moreover, we have

$$d(A_{1,\rho}, A_{2,\rho}) = \tau^*_{M_\rho} = 3\tau^*(\varepsilon) \geq \tau > \psi(\delta) = 3c_{\text{thick}}\delta^\gamma > \psi^*_{M_\rho}(\delta) \geq \delta\,, \tag{44}$$

where the last inequality follows from (Steinwart et al., 2021, Lemma 8.5) since *Assumption M* implies *Assumption P*, and hence X is connected. Consequently, part *v)* of (Steinwart, 2015b, Lemma A.3.1) yields

$$M_\rho^{-\delta} = \bigl(A_{1,\rho} \cup A_{2,\rho}\bigr)^{-\delta} = A_{1,\rho}^{-\delta} \cup A_{2,\rho}^{-\delta}\,, \tag{45}$$

and we additionally note that (44) implies

$$d(A_{1,\rho}^{-\delta}, A_{2,\rho}^{-\delta}) \geq d(A_{1,\rho}, A_{2,\rho}) \geq \tau\,. \tag{46}$$

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 (45) 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 (45) 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 (46). ■

Lemma 27 *Let* Assumption M *be satisfied, and* P_1 *and* P_2 *be defined by* (5) *for some fixed* $\rho^\dagger \in (\rho^*, \rho^{**}]$. *Then for* $i = 1, 2$ *and* $\rho \geq \rho^\dagger$ *we have*

$$M_{i,\rho} = M_\rho \cap A_{i,\rho^\dagger}\,. \tag{47}$$

Proof of Lemma 27: We first note that since P is normal at all levels $\rho > 0$, we have $\lambda^d(M_\rho \vartriangle \{h \geq \rho\}) = 0$ for all λ^d-densities h of P and all $\rho > 0$. For a fixed $\rho \geq \rho^\dagger > 0$. we can thus find a λ^d-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 λ^d-density of P_i and we have

$$\{h_i \geq \rho\} = M_\rho \cap A_{i,\rho^\dagger}\,. \tag{48}$$

Moreover, we have $M_{i,\rho} = \operatorname{supp} \lambda^d\bigl(\,\cdot\, \cap \{h_i \geq \rho\}\bigr)$, 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

$$\lambda^d(U \cap M_\rho) = \lambda^d\bigl(U \cap \{h \geq \rho\}\bigr) \geq \lambda^d\bigl(U \cap \{h_i \geq \rho\}\bigr) > 0\,,$$

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

$$\lambda^d\bigl(A_{i,\rho^\dagger} \cap \{h_i \geq \rho\}\bigr) = \lambda^d(\{h_i \geq \rho\}) = \lambda^d\bigl(X \cap \{h_i \geq \rho\}\bigr)\,.$$

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 (48) 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 $\lambda^d(U \cap \{h_i \geq \rho\}) > 0$. 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 $\lambda^d(U \cap M_\rho) > 0$. 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

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

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

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

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

Lemma 28 *Let* Assumption M *be satisfied, and P_1 and P_2 be defined by* (5) *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 28: *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

$$A_k \subset \xi(A_k) \subset A_{k,\rho^\dagger}\,, \qquad\qquad k = 1, 2. \tag{49}$$

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 (Steinwart, 2015b, Lemma A.3.1). In view of (47), we consequently need to prove that

$$\big((A_1 \cup A_2) \cap A_{i,\rho^\dagger}\big)^{+\delta} = \big(A_1^{+\delta} \cup A_2^{+\delta}\big) \cap A_{i,\rho^\dagger}^{+\delta}\,. \tag{50}$$

Be begin by observing that

$$(A_1 \cup A_2) \cap A_{i,\rho^\dagger} = (A_1 \ \cap A_{i,\rho^\dagger}) \cup (A_2 \ \cap A_{i,\rho^\dagger}) = A_i\,, \tag{51}$$

where we used (49) and $A_{1,\rho^\dagger} \cap A_{2,\rho^\dagger} = \emptyset$. Similarly, the right-hand side of (50) can be written as

$$\big(A_1^{+\delta} \cup A_2^{+\delta}\big) \cap A_{i,\rho^\dagger}^{+\delta} = \big(A_1^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta}\big) \cup \big(A_2^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta}\big)\,. \tag{52}$$

In addition, (49) ensures $A_i^{+\delta} \subset A_{i,\rho^\dagger}^{+\delta}$, and by continuing (52) we thus find

$$\bigl(A_1^{+\delta} \cup A_2^{+\delta}\bigr) \cap A_{i,\rho^\dagger}^{+\delta} = A_i^{+\delta} \cup \bigl(A_k^{+\delta} \cap A_{i,\rho^\dagger}^{+\delta}\bigr)\,, \qquad\qquad k \in \{1,2\} \setminus \{i\}\,. \tag{53}$$

Moreover, by part *ii)* of Lemma 26 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 $\delta < \tau < \tau^*(\varepsilon^\dagger)$. Thus, we have $A_{1,\rho^\dagger}^{+\delta} \cap A_{2,\rho^\dagger}^{+\delta} = \emptyset$, and since (49) 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 (53) gives

$$\bigl(A_1^{+\delta} \cup A_2^{+\delta}\bigr) \cap A_{i,\rho^\dagger}^{+\delta} = A_i^{+\delta}\,. \tag{54}$$

Now, (50) follows from combining (51) with (54).

ii). This directly follows from combining (47) with (Steinwart et al., 2021, Lemma 8.7). ■

Proof of Theorem 9: *i).* We first note that $\varepsilon^* \leq \varepsilon^{**} := \rho^{**} - \rho^*$ implies $\tau^*(\varepsilon^*) \leq \tau^*(\varepsilon^{**})$. Consequently, part *iii)* of Lemma 26 applied for $\rho := \rho^{**}$ gives the assertion.

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

$$\begin{tikzcd}
\mathcal{C}_\tau(M_{\rho^{**}}^{-\delta}) \arrow[rr, "\xi_{\rho_{\mathrm{out}}+\varepsilon}"] \arrow[dr, "\xi_{\rho+\varepsilon}"'] & & \mathcal{C}_\tau(M_{\rho_{\mathrm{out}}+\varepsilon}^{-\delta}) \\
& \mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}) \arrow[ur, "\xi"'] &
\end{tikzcd}$$

Now, part *iv)* of (Steinwart, 2015b, Theorem A.6.2) shows that $\xi_{\rho_{\mathrm{out}}+\varepsilon}$ is bijective, and consequently, $\xi_{\rho+\varepsilon}$ is injective. Moreover, part *i)* of (Steinwart, 2015a, Theorem 2.7) shows that $1 \leq |\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta})| \leq 2$, and since we already know that $|\mathcal{C}_\tau(M_{\rho^{**}}^{-\delta})| = 2$ and that $\xi_{\rho+\varepsilon}$ is injective, we conclude that $|\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta})| = 2$ 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 $\widehat{\xi}_\rho : \mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}) \to \widehat{\mathcal{C}}_\tau(L_\rho)$ is well-defined and bijective. To this end, we first recall that our assumption $\rho \leq \rho^{**} - 3\varepsilon$ together with (Steinwart, 2015a, Theorem 2.8) gives the following *disjoint* union:

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

Consequently, we have $\widehat{\xi}_\rho\bigl(\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta})\bigr) = \widehat{\mathcal{C}}_\tau(L_\rho)$, that is, we can view $\widehat{\xi}_\rho$ as a *surjective CRM* $\widehat{\xi}_\rho : \mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}) \to \widehat{\mathcal{C}}_\tau(L_\rho)$. Similarly, part *i)* of Theorem 6 ensures

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

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

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

i.e. $\breve{\xi}(B) \in \widehat{\mathcal{C}}_\tau(L_{\rho_{\mathrm{out}}})$. 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. In summary, we obtain the following commutative diagram of CRMs

$$\begin{tikzcd}
\mathcal{C}_\tau(M^{-\delta}_{\rho_{\mathrm{out}}+\varepsilon}) \arrow[r, "\widehat{\xi}_{\rho_{\mathrm{out}}}"] & \widehat{\mathcal{C}}_\tau(L_{\rho_{\mathrm{out}}}) \\
\mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon}) \arrow[u, "\xi"] \arrow[r, "\widehat{\xi}_{\rho}"'] & \widehat{\mathcal{C}}_\tau(L_\rho) \arrow[u, "\breve{\xi}_{|\widehat{\mathcal{C}}_\tau(L_\rho)}"']
\end{tikzcd}$$

Now, we have already seen that ξ is bijective, and in addition, part *ii)* of (Steinwart, 2015b, Theorem A.6.2) shows that $\widehat{\xi}_{\rho_{\mathrm{out}}}$ is injective. Moreover, our considerations above showed that $\widehat{\xi}_{\rho_{\mathrm{out}}}$ is surjective, and hence it 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

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

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

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

where the latter identity follows from part *iii)* of (Steinwart, 2015b, Theorem A.6.2).

iii). We first observe that $\varepsilon^\dagger := \rho^\dagger - \rho^*$ satisfies $\varepsilon^\dagger \geq \varepsilon^*$ and by (Steinwart et al., 2021, Lemma 8.5) we hence find $\delta \leq \psi^*(\delta) < \tau^*(\varepsilon^*) \leq \tau^*(\varepsilon^\dagger)$. Lemma 28 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

$$M^{-\delta}_{\rho+\varepsilon} \cap A^{-\delta}_{i,\rho^\dagger} \subset L_\rho \cap B_i\,, \tag{55}$$

$$L_\rho \cap B_i \subset M^{+\delta}_{\rho-\varepsilon} \cap A^{+\delta}_{i,\rho^\dagger}\,. \tag{56}$$

We begin by proving (55). To this end, we first observe that (3) 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 26 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 6 shows $\rho_{\mathrm{out}} \leq \rho^* + \varepsilon^* + 5\varepsilon \leq \rho^\dagger - \varepsilon$, and hence we have $\rho^\dagger - \varepsilon \in [\rho_{\mathrm{out}}, \rho^{**} - 3\varepsilon]$. Applying (Steinwart, 2015a, Theorem 2.8) and the already established part *ii)* to the level $\rho^\dagger - \varepsilon$ we then obtain $A^{-\delta}_{i,\rho^\dagger} \subset B_i^{\rho^\dagger - \varepsilon} \subset B_i$ for $i \in \{1, 2\}$.

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 $\rho^\dagger + 2\varepsilon \in [\rho_{\mathrm{out}}, \rho^{**} - 3\varepsilon]$ 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 (56). To this end, we note that (3) ensures $L_\rho \subset M^{+\delta}_{\rho-\varepsilon}$, and hence it suffices to prove $L_\rho \cap B_i \subset A^{+\delta}_{i,\rho^\dagger}$. Moreover, we have already shown that $L_\rho \cap B_i \subset B_i^{\rho^\dagger+2\varepsilon}$, and therefore, it suffices to establish

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

To this end, recall that we have already observed $\tau \leq \tau^*(\varepsilon^*) \leq \tau^*(\varepsilon^\dagger)$. Part *ii)* of Lemma 26 thus shows that $A^{+\delta}_{1,\rho^\dagger}$ and $A^{+\delta}_{2,\rho^\dagger}$ are the two τ-connected components of $M^{+\delta}_{\rho^\dagger}$. Now consider the CRM $\xi : \mathcal{C}_\tau(L_{\rho^\dagger+2\varepsilon}) \to \mathcal{C}_\tau(M^{+\delta}_{\rho^\dagger})$. 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

$$B_1^{\rho^\dagger+2\varepsilon} \subset A^{+\delta}_{1,\rho^\dagger}\,, \tag{57}$$

$$B_1^{\rho^\dagger+2\varepsilon} \subset A^{+\delta}_{2,\rho^\dagger}\,. \tag{58}$$

However, our construction ensures $V_1 \subset A^{+\delta}_{1,\rho^\dagger}$, and part *ii)* gives $V_1 \subset B_1^{\rho^\dagger+2\varepsilon}$. This gives $\emptyset \neq V_1 \subset B_1^{\rho^\dagger+2\varepsilon} \cap A^{+\delta}_{1,\rho^\dagger}$, and therefore we can exclude (58). Consequently (57) is true. The inclusion $B_2^{\rho^\dagger+2\varepsilon} \subset A^{+\delta}_{2,\rho^\dagger}$ can be shown analogously. ■

8.2 Proofs for Section 4

Lemma 29 *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*

$$h_{P,\delta}(x) \geq \rho - \rho\kappa_1(\tfrac{\sigma}{\delta}) \tag{59}$$

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

$$h_{P,\delta}(x) < \rho + \delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta})\,. \tag{60}$$

Finally, if P has a bounded density h, then the inequality (59) can be replaced by

$$h_{P,\delta}(x) \geq \rho - \kappa_1(\tfrac{\sigma}{\delta}) \cdot \|h\|_\infty \tag{61}$$

whenever $0 \leq \rho \leq \|h\|_\infty$ and (60) can be replaced, for all $\rho \geq 0$, by

$$h_{P,\delta}(x) < \rho + \kappa_1(\tfrac{\sigma}{\delta}) \cdot \|h\|_\infty\,. \tag{62}$$

Proof of Lemma 29: Let h be a λ^d-density of P. For the proof of (59), 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

$$\begin{aligned}\int_{B(x,\sigma)} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y) &= \int_{B(x,\sigma)\cap\{h\geq\rho\}} K_\delta(x-y)\, h(y)\ \mathrm{d}\lambda^d(y)\\ &\geq \rho \int_{B(x,\sigma)\cap\{h\geq\rho\}} K_\delta(x-y)\ \mathrm{d}\lambda^d(y)\\ &= \rho \int_{B(x,\sigma)} K_\delta(x-y)\ \mathrm{d}\lambda^d(y)\,, \end{aligned} \tag{63}$$

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 (10). Now, the assertion follows from (11), and for a bounded density h and $\rho \leq \|h\|_\infty$, the inequality (61) is a direct consequence of (59).

To show (60) we first note that (1) yields

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

Analogously to (63) 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 (59) 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, for bounded h inequality (62) follows from (11), 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 (60). ■

Proof of Theorem 11: To prove the first inclusion, we fix an $x \in M^{-2\sigma}_{\rho+\varepsilon+\epsilon}$ 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

$$B(x, 2\sigma) \subset M_{\rho+\varepsilon+\epsilon}\,. \tag{64}$$

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

$$h_{P,\delta}(x_i) < \rho + \varepsilon\,. \tag{65}$$

On the other hand, $\|x-x_i\| \leq \sigma$ together with the already shown (64) implies $B(x_i,\sigma) \subset M_{\rho+\varepsilon+\epsilon}$ by a simple application of the triangle inequality. Consequently, (59) together with $\epsilon \geq \rho\kappa_1(\frac{\sigma}{\delta})$ gives $h_{P,\delta}(x_i) \geq \rho+\varepsilon$, which contradicts (65). 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

$$h_{D,\delta}(x) = \frac{1}{n}\sum_{i=1}^n \delta^{-d} K\Big(\frac{x-x_i}{\delta}\Big) \leq \frac{1}{n}\sum_{i=1}^n \delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta}) = \delta^{-d}\kappa_\infty(\tfrac{\sigma}{\delta}) \leq \rho\,. \tag{66}$$

On the other hand, we have $B(x,\sigma) \subset B(x,2\sigma) \subset M_{\rho+\varepsilon+\epsilon}$ and therefore (59) 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 $h_{D,\delta}(x) > \rho$, which contradicts (66). Therefore there does exist a sample $x_i \in D$ with $\|x - x_i\| \leq \sigma$. Using the inclusion (64) together with the triangle inequality we then again find $B(x_i,\delta) \subset M_{\rho+\varepsilon+\epsilon}$, and hence (59) 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 (59) with (61) in the proof above.

To prove the second inclusion, 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 (60) 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 (60) with (62) in the proof above. ■

For the proof of Lemma 14 we need to recall the following classical result, which is a reformulation of (van der Vaart and Wellner, 1996, Theorem 2.6.4).

Theorem 30 *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* (17) *only depend on V.*

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

Lemma 31 *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 31: 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 14: The set $\mathcal{A} := \{x + B_{\|\cdot\|} : x \in \mathbb{R}^d\}$ of has finite VC-dimension by (Devroye and Lugosi, 2001, Corollary 4.2) or (Devroye and Lugosi, 2001, Lemma 4.1), respectively. In both cases, Theorem 30 thus shows that for

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

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 $\mathbb{R}^d$. Moreover, 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 31 leads to

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

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

Proof of Lemma 15: We first recall that if $A \subset E$ 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 then we have

$$\mathcal{N}\bigl(T(A), \|\cdot\|_F, |T|_\alpha \epsilon^\alpha\bigr) \leq \mathcal{N}\bigl(A, \|\cdot\|_E, \epsilon\bigr)\,, \qquad \epsilon > 0\,,$$

where $|T|_\alpha$ is the α-Hölder constant of T. We now fix a $0 < \delta \leq \bigl(\frac{|K|_\alpha}{\|K\|_\infty}\bigr)^{1/\alpha} \operatorname{diam}_{\|\cdot\|}(X)$ 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

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

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 $T(x) := k_{x,\delta}$. Moreover, T is α-Hölder continuous, since for $x, x' \in X$, we have

$$\|T(x) - T(x')\|_\infty = \sup_{y \in \mathbb{R}^d} \left|\delta^{-d} K\Bigl(\frac{x - y}{\delta}\Bigr) - \delta^{-d} K\Bigl(\frac{x' - y}{\delta}\Bigr)\right| \leq \delta^{-(\alpha+d)} |K|_\alpha\, \|x - x'\|^\alpha\,,$$

i.e. we have shown $|T|_\alpha \leq \delta^{-(\alpha+d)} |K|_\alpha$. By our initial observation and (19) we then conclude that

$$\mathcal{N}\bigl(\mathcal{K}_\delta, \|\cdot\|_{L_2(P)}, |T|_\alpha \epsilon^\alpha\bigr) = \mathcal{N}\bigl(T(X), \|\cdot\|_{L_2(P)}, |T|_\alpha \epsilon^\alpha\bigr) \leq \mathcal{N}\bigl(X, \|\cdot\|_\infty, \epsilon\bigr) \leq C_{\|\cdot\|}(X) \epsilon^{-d}$$

for all $\epsilon \in (0, \mathrm{diam}_{\|\cdot\|}(X)]$. A variable transformation together with our bound on $|T|_\alpha$ thus yields

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

for all $0 < \epsilon \leq \delta^{-(\alpha+d)}|K|_\alpha \,\mathrm{diam}_{\|\cdot\|}(X)$. Since the assumed $\delta \leq \big(\frac{|K|_\alpha}{\|K\|_\infty}\big)^{1/\alpha}\mathrm{diam}_{\|\cdot\|}(X)$ implies

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

we then see that (20) does hold for all $0 < \epsilon \leq \delta^{-d}\|K\|_\infty$. ■

For the proof of Theorem 16 we quote a version of Talagrand's inequality due to Bousquet (2002) from (Steinwart and Christmann, 2008, Theorem 7.5).

Theorem 32 *Let (Z, P) be a probability space and $\mathcal{G}$ be a set of measurable functions from Z to $\mathbb{R}$. Furthermore, let $B \geq 0$ and $\sigma \geq 0$ be constants such that $\mathbb{E}_P g = 0$, $\mathbb{E}_P g^2 \leq \sigma^2$, and $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}$. For $n \geq 1$, define $G : Z^n \to \mathbb{R}$ by*

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

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

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

For the proof of Theorem 16 we also need (Giné and Guillou, 2001, 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 33 *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 with $\sigma \leq B$ 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 in Theorem 32 satisfies*

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

We are now able to establish the following generalization of Theorem 16.

Proposition 34 *Let $X \subset \mathbb{R}^d$ and P be a probability measure on X with 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 be a symmetric kernel. Suppose further that the set $\mathcal{K}_\delta$ defined in (18) satisfies (21) 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*

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

Proof of Proposition 34: We define $\theta := \frac{1}{2} + \frac{1}{2p}$. Then $K \in L_1(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ leads to

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

We further define $k_{x,\delta} := \delta^{-d} K\bigl(\delta^{-1}(x - \cdot)\bigr)$ and $f_{x,\delta} := k_{x,\delta} - \mathbb{E}_P k_{x,\delta}$. Then we have $\mathbb{E}_P f_{x,\delta} = 0$ and $\|f_{x,\delta}\|_\infty \le 2\|K\|_\infty \delta^{-d}$ for all $x \in X$ and $\delta > 0$. Moreover, we have $\mathbb{E}_P f_{x,\delta}^2 \le \mathbb{E}_P k_{x,\delta}^2$ and thus

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

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

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

Applying Theorem 32 to $\mathcal{G} := \{f_{x,\delta} : x \in X\}$, we thus see, for all $\delta > 0$, $\varsigma > 0$, and $n \ge 1$, that

$$\|h_{D,\delta} - h_{P,\delta}\|_{\ell_\infty(X)} < 4\mathbb{E}_{D' \sim P^n} \|h_{D',\delta} - h_{P,\delta}\|_{\ell_\infty(X)} + \frac{2\varsigma}{n\delta^d} + \sqrt{\frac{2\varsigma \|h\|_p \|K\|_\infty^{2\theta}}{n\delta^{d(1+1/p)}}} \tag{67}$$

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 $|\mathbb{E}_P k_{x,\delta}| \le \|k_{x,\delta}\|_\infty = \delta^{-d}\|K\|_\infty =: B_\delta$. Consequently, we have

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

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

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

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

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

Theorem 33 together with $2\theta = 1 + 1/p$ thus yields

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

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

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

Combining this with (67) gives

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

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

Proof of Theorem 16: By Proposition 34, it suffices to find a constant C' such that

$$\frac{C\varsigma}{n\delta^d}\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}} + \sqrt{\frac{C\|h\|_p\varsigma}{\delta^{d(1+1/p)}n}\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}}} \leq C'\sqrt{\frac{\|h\|_p\,|\log\delta|\,\varsigma}{n\delta^{d(1+1/p)}}}\,. \tag{68}$$

To this end, we first observe that $\delta^{a+dq} \leq \|h\|_p^{1/2}C^{-1}$ implies $C\delta^{-a-dq}\|h\|_p^{-1/2} \leq \delta^{-2a-2dq}$, and thus we obtain $\log\frac{C}{\delta^{a+dq}\|h\|_p^{1/2}} \leq (2a+2dq)\log\delta^{-1}$. For $C'' := (2a+2dq)C$ we therefore find

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

Moreover, it is easy to check that the assumption $\frac{|\log\delta|}{n\delta^{d/p'}} \leq \frac{\|h\|_p}{C''\varsigma}$ ensures that

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

and from the latter we conclude that (68) holds for $C' := 2\sqrt{C''}$. The assertion now follows for the constant $C''' := \max\{C, C', C''\}$. ■

8.3 Proofs for the KDE-Based Clustering in Section 5

Proof of Theorem 17: Let us fix a $D \in X^n$ with $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon/2$. By (23) 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 $\epsilon = 0\ \delta^{-d}\kappa_\infty(\frac{\sigma}{\delta}) = 0 \le \rho_0$ as noted after Theorem 11. Furthermore, in the case of (9), (Steinwart et al., 2021, Lemma 4.2) shows

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

where in the third to last step we used $0 < \delta \le 1$ and in the second to last step we used (Steinwart et al., 2021, Lemma 8.17). In addition, we have $\delta^{-d}\kappa_\infty(\frac{\sigma}{\delta}) \le c\delta^{-d}e^{-|\log\delta|^2} = c\delta^{|\log\delta|-d} \le \varepsilon \le \rho_0$. Consequently, Theorem 11 shows, for all $\rho \ge \rho_0$, that

$$M_{\rho+\varepsilon}^{-2\sigma} \subset L_{D,\rho} \subset M_{\rho-\varepsilon}^{+2\sigma}\,. \tag{69}$$

i). The assertion follows from Theorem 7 applied in the case $\tilde{d} := 2\sigma$. Indeed, we have just seen that (40) holds for all $\rho \ge \rho_0$, if we replace δ by $\tilde{\delta}$, and our assumptions guarantee $\tilde{\delta} \in (0, \delta_{\mathrm{thick}}]$, $\rho_0 \ge \rho_*$, and $\tau > \psi(\tilde{\delta}) = 3c_{\mathrm{thick}}\tilde{\delta}^\gamma > 2c_{\mathrm{thick}}\tilde{\delta}^\gamma$. Moreover, (27) follows from (69).

ii). We check that the assumptions of Theorem 6 are satisfied for $\tilde{\delta} := 2\sigma$, if $\varepsilon^* \le (\rho^{**} - \rho^*)/9$. Clearly, we have $\tilde{\delta} \in (0, \delta_{\mathrm{thick}}]$, $\varepsilon \in (0, \varepsilon^*]$, and $\psi(\tilde{\delta}) < \tau$. To show $\tau \le \tau^*(\varepsilon^*)$ we write

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

Since we assumed $\varepsilon^* < \infty$, we obtain $E \neq \emptyset$ by the definition of ε^*. There thus exists an $\varepsilon' \in E$ with $\varepsilon' \le \inf E + \varepsilon \le \varepsilon^*$. Using the monotonicity of τ^* established in (Steinwart, 2015a, Theorem A.4.2) we then conclude that $\tau \le \tau^*(\varepsilon') \le \tau^*(\varepsilon^*)$, and hence all assumptions of (Steinwart, 2015a, Theorem 2.9) are indeed satisfied with δ replaced by $\tilde{\delta}$. The assertions now immediately follow from this theorem. ■

Proof of Corollary 18: Using Theorem 17 the proof of *ii)* is a literal copy of the proof of (Steinwart, 2015a, Theorem 4.1) and the proof of *i)* is an easy adaptation of this proof. ■

Proof of Corollary 20: Using Theorem 17 the proof is a simple combination and adaptation of the proofs of (Steinwart, 2015a, Theorem 4.3) and (Steinwart, 2015a, Corollary 4.4). ■

Proof of Corollary 23: Using Theorem 17 the proof is a simple combination and adaptation of the proofs of (Steinwart, 2015a, Theorem 4.7) and (Steinwart, 2015a, Corollary 4.8). ■

Proof of Theorem 24: The definition of $\varepsilon_{\delta,n}$ in (33) together with $4C_u^2 \log\log n \ge C\|h\|_\infty$ ensures that (25) and (26) are satisfied for all $\delta \in \Delta$. In addition, the assumptions of Theorem 24 ensure that the remaining conditions of Theorem 17 are satisfied. Now the assertion follows by some union bound arguments, which are analogous to those of the proof of (Steinwart, 2015a, Theorem 5.1). ■

Acknowledgment

The authors profusely thank the action editor and the reviewers for their comments and suggestions that significantly improved the readability of the paper. The work of I. Steinwart and P. Thomann was funded by the DFG Grant STE 1074/2-1. I. Steinwart was also funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016. B. K. Sriperumbudur thanks the National Science Foundation for support under the grant NSF-DMS-1713011.

References

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

K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 23*, pages 343–351. 2010.

K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. von Luxburg. Consistent procedures for cluster tree estimation and pruning. *IEEE Trans. Inf. Theory*, 60:7900–7912, 2014.

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

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

P. Fränti and O. Virmajoki. Iterative shrinking method for clustering problems. *Pattern Recognition*, 39:761–765, 2006.

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.

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.

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

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

S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. In *Proceedings of the 28th International Conference on Machine Learning*, pages 225–232. 2011.

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.

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

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

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

B.K. Sriperumbudur and I. Steinwart. Consistency and rates for clustering with DBSCAN. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics 2012*, pages 1090–1098, 2012.

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

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

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

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

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

I. Steinwart, B.K. Sriperumbudur, and P. Thomann. Adaptive clustering using kernel density estimators. Technical report, 2021. http://arxiv.org/abs/1708.05254v3.

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

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.

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

D. Wang, X. Lu, and A. Rinaldo. DBSCAN: Optimal rates for density-based cluster estimation. *J. Mach. Learn. Res.*, 20:1–50, 2019.