Uniform Convergence Rate of the Kernel Density Estimator Adaptive to Intrinsic Volume Dimension

Jisu Kim [1] **Jaehyeok Shin** [2] **Alessandro Rinaldo** [2] **Larry Wasserman** [2]

Abstract

We derive concentration inequalities for the supremum norm of the difference between a kernel density estimator (KDE) and its point-wise expectation that hold uniformly over the selection of the bandwidth and under weaker conditions on the kernel and the data generating distribution than previously used in the literature. We first propose a novel concept, called the volume dimension, to measure the intrinsic dimension of the support of a probability distribution based on the rates of decay of the probability of vanishing Euclidean balls. Our bounds depend on the volume dimension and generalize the existing bounds derived in the literature. In particular, when the data-generating distribution has a bounded Lebesgue density or is supported on a sufficiently well-behaved lower-dimensional manifold, our bound recovers the same convergence rate depending on the intrinsic dimension of the support as ones known in the literature. At the same time, our results apply to more general cases, such as the ones of distribution with unbounded densities or supported on a mixture of manifolds with different dimensions. Analogous bounds are derived for the derivative of the KDE, of any order. Our results are generally applicable but are especially useful for problems in geometric inference and topological data analysis, including level set estimation, density-based clustering, modal clustering and mode hunting, ridge estimation and persistent homology.

1. Introduction

Density estimation (see, e.g. Rao, 1983) is a classic and fundamental problem in non-parametric statistics that, especially in recent years, has also become a key step in many geometric inferential tasks. Among the numerous existing methods for density estimation, kernel density estimators (KDEs) are especially popular because of their conceptual simplicity and nice theoretical properties. A KDE is simply the Lebesgue density of the probability distribution obtained by convolving the empirical measure induced by the sample with an appropriate function, called kernel, (Parzen, 1962; Wand & Jones, 1994). Formally, let $X_1, \ldots, X_n$ be an independent and identically distributed sample from an unknown Borel probability distribution P in $\mathbb{R}^d$. For a given kernel K, where K is an appropriate function on $\mathbb{R}^d$ (often a density), and bandwidth $h > 0$, the corresponding KDE is the random Lebesgue density function defined as

$$x \in \mathbb{R}^d \mapsto \hat{p}_h(x) := \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \qquad (1)$$

The point-wise expectation of the KDE is the function

$$x \in \mathbb{R}^d \mapsto p_h(x) := \mathbb{E}[\hat{p}_h(x)],$$

and can be regarded as a smoothed version of the density of P, if such a density exists. In fact, interestingly, both $\hat{p}_h$ and p_h are Lebesgue probability densities for any choice of $h > 0$, regardless of whether P admits a Lebesgue density. What is more, p_h is often times able to capture important topological properties of the underlying distribution P or of its support (see, e.g. Fasy et al., 2014, Section 4.4). For instance, if a data-generating distribution consists of two point masses, it has no Lebesgue density but the pointwise mean of the KDE with Gaussian kernel is a density of mixtures of two Gaussian distributions whose mean parameters are the two point masses. Although P is quite different from the distribution corresponding to p_h, for practical purposes, one may in fact rely on p_h.

Though seemingly contrived, the previous example illustrates a general phenomenon encountered in many geometrical inference problems, namely that using p_h as a target for inference leads to not only well-defined statistical tasks but also to faster or even dimension independent rates. Results of this form, which require a uniform control over $\|\hat{p}_h - p_h\|_\infty := \sup_{x\in\mathbb{R}^d} \|\hat{p}_h(x) - p_h(x)\|$ are plentiful in the literature on density-based clustering (Rinaldo & Wasserman, 2010; Wang et al., 2017), modal clustering and mode

[1]Inria Saclay – Île-de-France, Palaiseau, France [2]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Jisu Kim <jisu.kim@inria.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

hunting (Chacón et al., 2015; Azizyan et al., 2015), mean-shift clustering (Arias-Castro et al., 2016), ridge estimation (Chen et al., 2015a;b) and inference for density level sets (Chen et al., 2017), cluster density trees (Balakrishnan et al., 2013; Kim et al., 2016) and persistent diagrams (Fasy et al., 2014; Chazal et al., 2014).

Asymptotic and finite-sample bounds on $\|\hat{p}_h - p_h\|_\infty$ under the existence of Lebesgue density have been well-studied for fixed bandwidth cases (Rao, 1983; Giné & Guillou, 2002; Sriperumbudur & Steinwart, 2012; Steinwart et al., 2017).

Bounds for KDEs not only uniform in $x \in \mathbb{R}^d$ but also with respect the choice of the bandwidth h have received relatively less attentions, although such bounds are important to analyze the consistency of KDEs with adaptive bandwidth, which may depend on the location x. Einmahl et al. (2005) showed that,

$$\limsup_{n\to\infty} \sup_{(c\log n)/n \le h \le 1} \frac{\sqrt{nh^d}\|\hat{p}_h - p_h\|_\infty}{\sqrt{\log(1/h) \vee \log\log n}} < \infty,$$

for regular kernels and bounded Lebesgue densities. Jiang (2017) provided a finite-sample bound on $\|\hat{p}_h - p_h\|_\infty$ that holds uniformly on h and under appropriate assumptions on K, and extended it to case of densities over well-behaved manifolds.

The main goal of this paper is to extend existing uniform bounds on KDEs by weakening the conditions on the kernel and making it adaptive to the intrinsic dimension of the underlying distribution. We first propose a novel concept, called the *volume dimension,* to characterize the intrinsic dimension of the underlying distribution. In detail, the volume dimension d_{vol} is the rate of decay of the probability of vanishing Euclidean balls, i.e. fix a subset $\mathbb{X} \subset \mathbb{R}^d$, then

$$d_{\mathrm{vol}} = \sup\left\{\nu \in \mathbb{R} :\ \limsup_{r\to 0} \sup_{x\in\mathbb{X}} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu} < \infty\right\}.$$

We show that, if K satisfies mild regularity conditions, with probability at least $1-\delta$,

$$\sup_{h\ge l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)| \le C\sqrt{\frac{(\log(1/l_n))_+ + \log(2/\delta)}{n l_n^{2d-d_{\mathrm{vol}}+\epsilon}}}, \qquad (2)$$

for any $\epsilon \in (0, d_{\mathrm{vol}})$, $\{l_n\}$ a positive sequence approaching 0 and C is a constant that does not depend on n nor l_n. Under additional, weak regularity conditions on P, the quantity ϵ can be taken to be 0 in (2). If the distribution has a bounded Lebesgue density, $d_{\mathrm{vol}} = d$ so our result recovers existing results in literature in terms of rates of convergence. For a bounded density on a d_M-dimensional manifold we obtain, under appropriate conditions, that $d_{\mathrm{vol}} = d_M$. Thus, if KDEs are defined with a correct normalizing factor h^{d_M} instead of h^d, our rate also recovers the ones in the literature on density estimation over manifolds. At the same time, our bounds apply to more general cases, such as a distribution with an unbounded density or supported on a mixture of manifolds with different dimensions. We have also shown the optimality of (2) up to log terms by showing that under the mild regularity conditions on K and P,

$$\sup_{h\ge l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)| \ge C'\sqrt{\frac{1}{n l_n^{2d-d_{\mathrm{vol}}}}}. \qquad (3)$$

We make the following contributions:

1. We propose a novel concept, called the volume dimension, to characterize the convergence rate of the KDE on arbitrary distributions.
2. We derive high probability finite sample bounds for $\|\hat{p} - p_h\|_\infty$, uniformly over the choice of $h \ge l_n$, for a given l_n depending on n.
3. We derive rates of consistency in the $Ł_\infty$ norm that are adaptive to the volume dimension of the distribution under conditions on the kernel that, to the best of our knowledge, are weaker than the ones existing in the literature, and without assumptions on the distribution. Hence, our bounds recover known previous results, and apply to more general cases such as a distribution with unbounded density or supported on a mixture of manifolds with different dimensions.
4. We show that our bound is optimal up to log terms under weak conditions on the kernel and the distribution.
5. We also obtain analogous bounds for all higher order derivatives of $\hat{p}_h$ and p_h.

The closest results to the ones we present are by Jiang (2017), who relies on relative VC bounds to derive finite sample bounds on $\|\hat{p}_h - p_h\|_\infty$ for a special class of kernels and assuming P to have a well-behaved support. Our analysis relies instead on more sophisticated techniques rooted in the theory of empirical process theory as outlined in (Sriperumbudur & Steinwart, 2012) and are applicable to a broader class of kernels. In addition, we do not assume any condition on the underlying distribution.

2. Notation

Below, we recap basic concepts and establish some notation that are used throughout the paper. For more detailed definitions, see Appendix A.

We let $\|\cdot\|$ be the Euclidean 2-norm. For $x \in \mathbb{R}^d$ and $r > 0$, we use the notation $\mathbb{B}_{\mathbb{R}^d}(x, r)$ for the open Euclidean ball centered at x and radius r, i.e. $\mathbb{B}_{\mathbb{R}^d}(x,r) = \{y \in \mathbb{R}^d : \|y-x\| < r\}$. We fix a subset $\mathbb{X} \subset \mathbb{R}^d$ on which we are considering the uniform convergence of the KDE.

The Hausdorff measure is a generalization of the Lebesgue measure to lower dimensional subsets of $\mathbb{R}^d$. The Hausdorff dimension is a generalization of the intrinsic dimension of a manifold to general sets. For $\nu \in \{1,\ldots,d\}$, let λ_ν be a normalized ν-dimensional Hausdorff measure on $\mathbb{R}^d$ satisfying that its measure on any ν-dimensional unit cube is 1. We use the notation $\omega_\nu := \lambda_\nu(\mathbb{B}_{\mathbb{R}^\nu}(0,1)) = \frac{\pi^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2}+1)}$ for the volume of the unit ball in $\mathbb{R}^\nu$ for $\nu = 1,\ldots,d$.

First introduced by (Federer, 1959), the reach has been the minimal regularity assumption in the geometric measure theory. A manifold with positive reach means that the projection to the manifold is well defined in a small neighborhood of the manifold.

3. Volume Dimension

We first characterize the intrinsic dimension of a probability distribution in terms of the rate of decay of the probability of Euclidean balls of vanishing volumes. When a probability distribution P has a bounded density p with respect to a well-behaved manifold M of dimension d_M, it is known that, for any point $x \in M$, the measure on the ball $\mathbb{B}_{\mathbb{R}^d}(x,r)$ centered at x and radius r decays as

$$P\left(\mathbb{B}_{\mathbb{R}^d}(x,r)\right) \sim r^{d_M},$$

when r is small enough. From this, we define the volume dimension to be the maximum possible exponent rate that can dominate the probability volume decay on balls.

Definition 1 (Volume Dimension)**.** *Let P be a probability distribution on $\mathbb{R}^d$. The* volume dimension *of P is a non-negative real number defined as*

$$d_{\mathrm{vol}}(P) := \sup\left\{\nu \ge 0 : \limsup_{r\to 0}\sup_{x\in\mathbb{X}} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu} < \infty\right\}. \tag{4}$$

We will use the notation d_{vol} when P is clearly specified by the context.

The volume dimension has a connection with the Hausdorff dimension. If a probability distribution has a positive measure on a set, then the volume dimension is between 0 and the Hausdorff dimension of the set. So, if that set is a manifold, then the volume dimension is always between 0 and the dimension of the manifold. In particular, the volume dimension of any probability distribution is between 0 and the ambient dimension d.

Proposition 1. *Let P be a probability distribution on $\mathbb{R}^d$, and d_{vol} be its volume dimension. Suppose there exists a set A satisfying $P(A\cap\mathbb{X}) > 0$ and with Hausdorff dimension d_H. Then $0 \le d_{\mathrm{vol}} \le d_H$. Hence if A is a d_M-dimensional manifold, then $0 \le d_{\mathrm{vol}} \le d_M$. In particular, for any probability distribution P on $\mathbb{R}^d$, $0 \le d_{\mathrm{vol}} \le d$. Also, if P has a point mass, i.e. there exists $x \in \mathbb{X}$ with $P(\{x\}) > 0$, then $d_{\mathrm{vol}} = 0$.*

The volume dimension is well defined with mixtures of distributions. Specifically, the volume dimension of the mixture is the minimum of the volume dimensions of the component distributions.

Proposition 2. *Let $P_1,\ldots,P_m$ be probability distributions on $\mathbb{R}^d$, and $\lambda_1,\ldots,\lambda_m \in (0,1)$ with $\sum_{i=1}^m \lambda_i = 1$. Then*

$$d_{\mathrm{vol}}\left(\sum_{i=1}^m \lambda_i P_i\right) = \min\left\{d_{\mathrm{vol}}(P_i) : 1 \le i \le m\right\}.$$

In particular, when d_{vol} is understood as a real-valued function on the space of probability distributions, both its sublevel sets and superlevel sets are convex.

The name "volume dimension" suggests that the volume dimension of a probability distribution has a connection with the dimension of the support. The two dimensions are indeed equal when the support is a manifold with positive reach and the probability distribution has a bounded density with respect to the uniform measure on the manifold (e.g. the Hausdorff measure). In particular when the probability distribution has a bounded density with respect to the d-dimensional Lebesgue measure, the volume dimension equals the ambient dimension d.

Proposition 3. *Let P be a probability distribution on $\mathbb{R}^d$, and d_{vol} be its volume dimension. Suppose there exists a d_M-dimensional manifold M with positive reach satisfying $P(M\cap\mathbb{X}) > 0$ and $\mathrm{supp}(P) \subset M$. If P has a bounded density p with respect to the normalized d_M-dimensional Hausdorff measure λ_{d_M}, then $d_{\mathrm{vol}} = d_M$. In particular, when P has a bounded density p with respect to the d-dimensional Lebesgue measure λ_d, then $d_{\mathrm{vol}} = d$.*

See Section C for a comparison of the volume dimension with the Hausdorff dimension and other notions of the dimension.

Even though, as we will soon show, our bounds for KDEs hold without any assumptions on the probability distribution and lead to convergence rates arbitrary close to the optimal minimax rates, in order to actually achieve such exact optimal rate, we require weak additional conditions on the probability distributions. Note that, from the definition of the volume dimension, the ratio $\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}$ is uniformly bounded for ν smaller than the volume dimension.

Lemma 4. *Let P be a probability distribution on $\mathbb{R}^d$, and d_{vol} be its volume dimension. Then for any $\nu \in [0, d_{\mathrm{vol}})$,* *there exists a constant $C_{\nu,P}$ depending only on P and ν such that for all $x \in \mathbb{X}$ and $r > 0$,*

$$\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{\nu}} \leq C_{\nu,P}. \tag{5}$$

For the exact optimal rate, we impose conditions on how the probability volume decay in (5) behaves with respect to the volume dimension.

Assumption 1. *Let P be a probability distribution P on $\mathbb{R}^d$, and $d_{\rm vol}$ be its volume dimension. We assume that*

$$\limsup_{r\to 0}\sup_{x\in\mathbb{X}} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\rm vol}}} < \infty. \tag{6}$$

Assumption 2. *Let P be a probability distribution on $\mathbb{R}^d$, and $d_{\rm vol}$ be its volume dimension. We assume that*

$$\sup_{x\in\mathbb{X}}\liminf_{r\to 0} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\rm vol}}} > 0. \tag{7}$$

These assumptions are in fact weak and hold for common probability distributions. For example, if a probability distribution is supported on a manifold, Assumption 1 and 2 hold under the same condition as in Proposition 3. In particular, Assumption 1 and 2 hold when the probability distribution has a bounded density with respect to the d-dimensional Lebesgue measure.

Proposition 5. *Under the same condition as in Proposition 3, Assumption 1 and 2 hold.*

Also, the Assumption 1 and 2 is closed under the convex combination. In other words, a mixture of probability distributions satisfy Assumption 1 and 2 if all its component satisfy those assumptions.

Proposition 6. *The set of probability distributions satisfying Assumption 1 is convex. And so is the set of probability distributions satisfying Assumption 2.*

We end this section with an example of an unbounded density. In this case, the volume dimension is strictly smaller than the dimension of the support which illustrates why the dimension of the support is not enough to characterize the dimensionality of a distribution.

Example 7. *Let P be a distribution on $\mathbb{R}^d$ having a density p with respect to the d-dimensional Lebesgue measure. Fix $\beta < d$, and suppose $p : \mathbb{R}^d \to \mathbb{R}$ is defined as*

$$p(x) = \frac{(d-\beta)\Gamma\left(\frac{d}{2}\right)}{2\pi^{\frac{d}{2}}} \|x\|^{-\beta} I(\|x\| \leq 1).$$

Then, for each fixed $r \in [0,1]$,

$$\sup_{x\in\mathbb{R}^d} P(\mathbb{B}_{\mathbb{R}^d}(x,r)) = P(\mathbb{B}_{\mathbb{R}^d}(0,r)) = r^{d-\beta}.$$

Hence from Definition 1, the volume dimension is

$$d_{\rm vol}(P) = d - \beta,$$

and from (6) and (7), Assumption 1 and 2 are satisfied.

4. Uniform convergence of the Kernel Density Estimator

To derive a bound on the performance of a kernel density estimator that is valid uniformly in h and $x \in \mathbb{X}$, we first rewrite

$$\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)|$$

as a supremum over a function class. Formally, for $x \in \mathbb{X}$ and $h \geq l_n > 0$, let $K_{x,h}(\cdot) := K\left(\frac{x-\cdot}{h}\right)$ and consider the following class of normalized kernel functions centered around each point in $\mathbb{X}$ and with bandwidth greater than or equal to $l_n > 0$:

$$\tilde{\mathcal{F}}_{K,[l_n,\infty)} := \left\{(1/h^d)K_{x,h} : x \in \mathbb{X},\, h \geq l_n\right\}.$$

Then $\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)|$ can be rewritten as a supremum of an empirical process indexed by $\tilde{\mathcal{F}}$, that is,

$$\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)| = \sup_{f\in\tilde{\mathcal{F}}_{K,[l_n,\infty)}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right|. \tag{8}$$

We combine Talagrand's inequality and a VC type bound to bound (8), following the approach of Sriperumbudur & Steinwart (2012, Theorem 3.1). The following version of Talagrand's inequality is from Bousquet (2002, Theorem 2.3) and simplified in Steinwart & Christmann (2008, Theorem 7.5).

Proposition 8. *(Bousquet, 2002, Theorem 2.3), (Steinwart & Christmann, 2008, Theorem 7.5, Theorem A.9.1)*

Let $(\mathbb{R}^d, P)$ be a probability space and let $X_1, \ldots, X_n$ be i.i.d. from P. Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that is separable in $L_\infty(\mathbb{R}^d)$. Suppose all functions $f \in \mathcal{F}$ are P-measurable, and there exists $B, \sigma > 0$ such that $\mathbb{E}_P f = 0$, $\mathbb{E}_P f^2 \leq \sigma^2$, and $\|f\|_\infty \leq B$, for all $f \in \mathcal{F}$. Let

$$Z := \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i)\right|,$$

Then for any $\delta > 0$,

$$P\left(Z \geq \mathbb{E}_P[Z] + \sqrt{\left(\frac{2}{n}\log\frac{1}{\delta}\right)(\sigma^2 + 2B\mathbb{E}_P[Z])} + \frac{2B\log\frac{1}{\delta}}{3n}\right) \leq \delta.$$

By applying Talagrand's inequality to (8), $\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)|$ can be upper bounded in terms of n, $\|K_{x,h}\|_\infty$, $\mathbb{E}_P[K_{x,h}^2]$, and

$$\mathbb{E}_P\left[\sup_{f\in\tilde{\mathcal{F}}_{K,[l_n,\infty)}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right|\right]. \tag{9}$$

To bound the last term, we use the uniformly bounded VC class assumption on the kernel. The following bound on the expected suprema of empirical processes of VC classes of functions is from Giné & Guillou (2001, Proposition 2.1).

Proposition 9. *(Giné & Guillou (2001, Proposition 2.1), (Sriperumbudur & Steinwart, 2012, Theorem A.2))*

Let $(\mathbb{R}^d, P)$ be a probability space and let $X_1, \ldots, X_n$ be i.i.d. from P. Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that is uniformly bounded VC-class with dimension ν, i.e. there exists positive numbers A,B such that, for all $f \in \mathcal{F}$, $\|f\|_\infty \leq B$, and the covering number $\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon)$ satisfies

$$\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon) \leq \left(\frac{AB}{\epsilon}\right)^\nu .$$

for every probability measure Q on $\mathbb{R}^d$ and for every $\epsilon \in (0, B)$. Let $\sigma > 0$ be a positive number such that $\mathbb{E}_P f^2 \leq \sigma^2$ for all $f \in \mathcal{F}$. Then there exists a universal constant C not depending on any parameters such that

$$\mathbb{E}_P\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)\right|\right] \leq C\left(\frac{\nu B}{n}\log\left(\frac{AB}{\sigma}\right) + \sqrt{\frac{\nu\sigma^2}{n}\log\left(\frac{AB}{\sigma}\right)}\right).$$

By applying Proposition 8 and Proposition 9 to $\tilde{\mathcal{F}}_{K,[l_n,\infty)}$, it can be shown that the upper bound of

$$\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)|$$

can be written as a function of $\|K_{x,h}\|_\infty$ and $\mathbb{E}_P[K_{x,h}^2]$. When the lower bound on the interval l_n is not too small, the terms relating to $\mathbb{E}_P[K_{x,h}^2]$ are more dominant. Hence, to get a good upper bound with respect to both n and h, it is important to get a tight upper bound for $\mathbb{E}_P[K_{x,h}^2]$. Under the existence of the Lebesgue density of P, it can be shown that

$$\mathbb{E}_P[K_{x,h}^2] \leq \|K\|_2 \|p\|_\infty h^d,$$

by change of variables. (see, e.g. the proof of Proposition A.5. in Sriperumbudur & Steinwart (2012).)

For general distributions (such as the ones supported on a lower-dimensional manifold), the change of variables argument is no longer directly applicable. However, under an integrability condition on the kernel, detailed below, we can provide a bound based on the volume dimension.

Assumption 3. *Let $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel function with $\|K\|_\infty < \infty$, and fix $k > 0$. We impose an integrability condition: either $d_{\mathrm{vol}} = 0$ or*

$$\int_0^\infty t^{d_{\mathrm{vol}}-1} \sup_{\|x\|\geq t} |K(x)|^k dt < \infty. \tag{10}$$

We set $k = 2$ by default unless it is specified in otherwise.

Remark 10. *It is important to emphasize that Assumption 3 is weak, as it is satisfied by commonly used kernels. For instance, if the kernel function $K(x)$ decays at a polynomial rate strictly faster than d_{vol}/k (which is at most d/k) as $x \to \infty$, that is, if*

$$\limsup_{x\to\infty} \|x\|^{d_{\mathrm{vol}}/k+\epsilon} K(x) < \infty,$$

for any $\epsilon > 0$, the integrability condition (10) is satisfied. Also, if the kernel function $K(x)$ is spherically symmetric, that is, if there exists $\tilde{K} : [0, \infty) \to \mathbb{R}$ with $K(x) = \tilde{K}(\|x\|)$, then the integrability condition (10) is satisfied provided $\|K\|_k < \infty$. Kernels with bounded support also satisfy the condition (10). Thus, most of the commonly used kernels including Uniform, Epanechnikov, and Gaussian kernels satisfy the above integrability condition.

By combining Assumption 3 and Lemma 4, we can bound $\mathbb{E}_P[K_{x,h}^2]$ in terms of the volume dimension d_{vol}.

Lemma 11. *Let $(\mathbb{R}^d, P)$ be a probability space and let $X \sim P$. For any kernel K satisfying Assumption 3 with $k > 0$, the expectation of the k-moment of the kernel is upper bounded as*

$$\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right] \leq C_{k,P,K,\epsilon} h^{d_{\mathrm{vol}}-\epsilon}, \tag{11}$$

for any $\epsilon \in (0, d_{\mathrm{vol}})$, where $C_{k,P,K,\epsilon}$ is a constant depending only on k, P, K, and ϵ. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0 in (11).

4.1. Uniformity on a ray of bandwidths

In this subsection, we demonstrate an L_∞ convergence rate for kernel density estimators, that is valid is uniformly on a ray of bandwidths $[l_n, \infty)$.

To apply the VC type bound from Proposition 9, the function class,

$$\mathcal{F}_{K,[l_n,\infty)} := \{K_{x,h} : x \in \mathbb{X}, h \geq l_n\},$$

should be not too complex. One common approach is to assume that $\mathcal{F}_{K,[l_n,\infty)}$ is a uniformly bounded VC-class, which is defined imposing appropriate bounds on the metric entropy of the function class (Giné & Guillou, 1999; Sriperumbudur & Steinwart, 2012).

Assumption 4. *Let $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel function with $\|K\|_\infty, \|K\|_2 < \infty$. We assume that,*

$$\mathcal{F}_{K,[l_n,\infty)} := \{K_{x,h} : x \in \mathbb{X}, h \geq l_n\}$$

is a uniformly bounded VC-class with dimension ν, i.e., there exists positive numbers A and ν such that, for every *probability measure Q on $\mathbb{R}^d$ and for every $\epsilon \in (0, \|K\|_\infty)$, the covering numbers $\mathcal{N}(\mathcal{F}_{K,[l_n,\infty)}, L_2(Q), \epsilon)$ satisfies*

$$\mathcal{N}(\mathcal{F}_{K,[l_n,\infty)}, L_2(Q), \epsilon) \leq \left(\frac{A\|K\|_\infty}{\epsilon} \right)^\nu,$$

where the covering number is defined as the minimal number of open balls of radius ϵ with respect to $L_2(Q)$ distance whose centers are in $\mathcal{F}_{K,[l_n,\infty)}$ to cover $\mathcal{F}_{K,[l_n,\infty)}$.

Since $[l_n, \infty) \subset (0, \infty)$, one sufficient condition for Assumption 4 is to impose uniformly bounded VC class condition on a larger function class,

$$\mathcal{F}_{K,(0,\infty)} = \{K_{x,h} : x \in \mathbb{X}, h > 0\}.$$

This is implied by condition (K) in Giné et al. (2004) or condition (K_1) in Giné & Guillou (2001), which are standard conditions to assume for the uniform bound on the KDE. In particular, the condition is satisfied when $K(x) = \phi(p(x))$, where p is a polynomial and ϕ is a bounded real function of bounded variation as in Nolan & Pollard (1987), which covers commonly used kernels, such as Gaussian, Epanechnikov, Uniform, etc.

Under Assumption 3 and 4, we derive our main concentration inequality for $\sup_{h \geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)|$.

Theorem 12. *Let P be a probability distribution and let K be a kernel function satisfying Assumption 3 and 4. Then, with probability at least $1-\delta$,*

$$\begin{aligned} &\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)| \\ &\leq C \left(\frac{(\log(1/l_n))_+}{n l_n^d} + \sqrt{\frac{(\log(1/l_n))_+}{n l_n^{2d-d_{\mathrm{vol}}+\epsilon}}} \right. \\ &\qquad \left. + \sqrt{\frac{\log(2/\delta)}{n l_n^{2d-d_{\mathrm{vol}}+\epsilon}}} + \frac{\log(2/\delta)}{n l_n^d} \right), \end{aligned} \tag{12}$$

for any $\epsilon \in (0, d_{\mathrm{vol}})$, where C is a constant depending only on A, $\|K\|_\infty$, d, ν, d_{vol}, $C_{k=2,P,K,\epsilon}$, ϵ. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0 in (12).

When δ is fixed and $l_n < 1$, the dominating terms in (12) are $\frac{\log(1/l_n)}{n l_n^d}$ and $\sqrt{\frac{\log(1/l_n)}{n l_n^{2d-d_{\mathrm{vol}}}}}$. If l_n does not vanish too rapidly, then the second term dominates the upper bound in (12) as in the following corollary.

Corollary 13. *Let P be a probability distribution and let K be a kernel function satisfying Assumption 3 and 4. Fix $\epsilon \in (0, d_{\mathrm{vol}})$. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0. Suppose*

$$\limsup_n \frac{(\log(1/\ell_n))_+ + \log(2/\delta)}{n \ell_n^{d_{\mathrm{vol}}-\epsilon}} < \infty.$$

Then, with probability at least $1-\delta$,

$$\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)| \leq C' \sqrt{\frac{(\log(\frac{1}{l_n}))_+ + \log(\frac{2}{\delta})}{n l_n^{2d-d_{\mathrm{vol}}+\epsilon}}}, \tag{13}$$

where C' depending only on A, $\|K\|_\infty$, d, ν, d_{vol}, $C_{k=2,P,K,\epsilon}$, ϵ.

4.2. Fixed bandwidth

In this subsection, we prove a finite-sample uniform convergence bound on kernel density estimators for one *fixed* choice $h_n > 0$ of the bandwidth (we leave the dependence on n explicit in our notation to emphasize that the choice of the bandwidth may still depend on n). We are interested in a high probability bound on

$$\sup_{x\in\mathbb{X}} |\hat{p}_{h_n}(x) - p_{h_n}(x)|.$$

Of course, the above quantity can be bounded by the results in the previous subsection because

$$\sup_{x\in\mathbb{X}} |\hat{p}_{h_n}(x) - p_{h_n}(x)| \leq \sup_{h\geq h_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)|, \tag{14}$$

Therefore, the convergence bound uniform on a ray of bandwidths in Theorem 12 and Corollary 13 is applicable to fixed bandwidth cases.

However, if the set $\mathbb{X}$ is bounded, that is, if there exists $R > 0$ such that $\mathbb{X} \subset \mathbb{B}_{\mathbb{R}^d}(0, R)$, then, for the kernel density estimator with a M_K-Lipschitz continuous kernel and fixed bandwidth, we can derive a uniform convergence bound without the finite VC condition of (Giné & Guillou, 2001; Giné et al., 2004) based on the following lemma.

Lemma 14. *Suppose there exists $R > 0$ with $\mathbb{X} \subset \mathbb{B}_{\mathbb{R}^d}(0, R)$. Let the kernel K is M_K-Lipschitz continuous. Then for all $\eta \in (0, \|K\|_\infty)$, the supremum of the η-covering number $\mathcal{N}(\mathcal{F}_{K,h}, L_2(Q), \eta)$ over all measure Q is upper bounded as*

$$\sup_Q \mathcal{N}(\mathcal{F}_{K,h}, L_2(Q), \eta) \leq \left(\frac{2RM_K h^{-1} + \|K\|_\infty}{\eta} \right)^d.$$

Corollary 15. *Suppose there exists $R > 0$ with $\mathbb{X} \subset \mathbb{B}_{\mathbb{R}^d}(0, R)$. Let K be a M_K-Lipschitz continuous kernel function satisfying Assumption 3. Fix $\epsilon \in (0, d_{\mathrm{vol}})$. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0. Suppose*

$$\limsup_n \frac{(\log(1/h_n))_+ + \log(2/\delta)}{n h_n^{d_{\mathrm{vol}}-\epsilon}} < \infty.$$

Then with probability at least $1-\delta$,

$$\sup_{x\in\mathbb{X}} |\hat{p}_{h_n}(x) - p_{h_n}(x)| \leq C'' \sqrt{\frac{(\log(\frac{1}{h_n}))_+ + \log(\frac{2}{\delta})}{n h_n^{2d-d_{\mathrm{vol}}+\epsilon}}}, \tag{15}$$

where C'' is a constant depending only on R, M_K, $\|K\|_\infty$, d, ν, d_{vol}, $C_{k=2,P,K,\epsilon}$, ϵ.

5. Lower bound for the convergence of the Kernel Density Estimator

Consider the fixed bandwidth case. In Corollary 15, it was shown that, with probability $1-\delta$,

$$\sup_{x\in\mathbb{X}}|\hat{p}_{h_n}(x)-p_{h_n}(x)|\leq C''_{\delta}\sqrt{\frac{(\log(1/h_n))_+}{nh_n^{2d-d_{\rm vol}}}},$$

where C''_{δ} might depend on δ but not on n or h_n. In this Section, we show that this upper bound is not improvable and is therefore optimal up to a $\log(1/h_n)$ term, by showing that there exists a high probability lower bound of order $1/\sqrt{nh_n^{2d-d_{\rm vol}}}$.

Proposition 16. *Suppose P is a distribution satisfying Assumption 2 and with positive volume dimension $d_{\rm vol}>0$. Let K be a kernel function satisfying Assumption 3 with $k=1$ and $\lim_{t\to0}\inf_{\|x\|\leq t}K(x)>0$. Suppose $\lim_n nh_n^{d_{\rm vol}}=\infty$. Then, with probability $1-\delta$, the following holds for all large enough n and small enough h_n:*

$$\sup_{x\in\mathbb{X}}|\hat{p}_{h_n}(x)-p_{h_n}(x)|\geq C_{P,K,\delta}\sqrt{\frac{1}{nh_n^{2d-d_{\rm vol}}}}.$$

where $C_{P,K,\delta}$ is a constant depending only on P, K,and δ.

This gives an immediate corollary for a ray of bandwidths.

Corollary 17. *Assume the same condition as in Proposition 16, and suppose $l_n\to0$ with $nl_n^{d_{\rm vol}}\to\infty$. Then, with probability $1-\delta$, the following holds for all large n:*

$$\sup_{h\geq l_n,x\in\mathbb{X}}|\hat{p}_h(x)-p_h(x)|\geq C_{P,K,\delta}\sqrt{\frac{1}{nl_n^{2d-d_{\rm vol}}}}.$$

By combining the lower and upper bounds together, we conclude that, with high probability,

$$\sqrt{\frac{1}{nh_n^{2d-d_{\rm vol}}}}\lesssim\sup_{x\in\mathbb{X}}|\hat{p}_{h_n}(x)-p_{h_n}(x)|\lesssim\sqrt{\frac{(\log(\frac{1}{h_n}))_+}{nh_n^{2d-d_{\rm vol}}}},$$

for all large enough n. Similar holds for a ray of bandwidths as well. They imply that the uniform convergence KDE bounds in our paper are optimal up to $\log(1/h_n)$ terms for both the fixed bandwidth and the ray on bandwidths cases.

Example 18 (Example 7, revisited)**.** *Let P be as in Example 7 and let K be any Lipschitz continuous kernel function with $K(0)>0$ and compact support. It can be easily checked that the conditions in Corollary 15 are satisfied with $R=2$, $d_{\rm vol}=d-\beta$ and the kernel satisfies the integrability Assumption 3 with $k=1,2$. It can be also shown that $\lim_{t\to0}\inf_{\|x\|\leq t}K(x)>0$. Therefore, for small enough h_n, Corollary 15 and Proposition 16 imply*

$$C'\sqrt{\frac{1}{nh_n^{d+\beta}}}\leq\sup_{x\in\mathbb{X}}|\hat{p}_{h_n}(x)-p_{h_n}(x)|\leq C''\sqrt{\frac{\log(\frac{1}{h_n})}{nh_n^{d+\beta}}},$$

with high probability for all large enough n. That is, the L_∞ convergence rate of the KDE is of order $\sqrt{\frac{1}{nh_n^{d+\beta}}}$ (up to a $\log(1/h_n)$ term). Hence, although it has a Lebesgue density, its convergence rate is different from $\sqrt{\frac{1}{nh_n^{d}}}$, which is the usual rate for probability distributions with bounded Lebesgue density.

6. Uniform convergence of the Derivatives of the Kernel Density Estimator

In this final section, we provide analogous finite-sample uniform convergence bound on the derivatives of the kernel density estimator. For a nonnegative integer vector $s=(s_1,\ldots,s_d)\in(\{0\}\cup\mathbb{N})^d$, define $|s|=s_1+\cdots+s_d$ and

$$D^s:=\frac{\partial^{|s|}}{\partial x_1^{s_1}\cdots\partial x_d^{s_d}}.$$

For D^s operator to be well defined and interchange with integration, we need the following smoothness condition on the kernel K.

Assumption 5. *For given $s\in(\{0\}\cup\mathbb{N})^d$, let $K:\mathbb{R}^d\to\mathbb{R}$ be a kernel function satisfying such that the partial derivative $D^sK:\mathbb{R}^d\to\mathbb{R}$ exists and $\|D^sK\|_\infty<\infty$.*

Under Assumption 5, Leibniz's rule is applicable and, for each $x\in\mathbb{X}$, $D^s\hat{p}_h(x)-D^sp_h(x)$ can be written as

$$\begin{aligned}&D^s\hat{p}_h(x)-D^sp_h(x)\\&=\frac{1}{n}\sum_{i=1}^n\frac{1}{h^{d+|s|}}D^sK_{x,h}(X_i)-\mathbb{E}_P\left[\frac{1}{h^{d+|s|}}D^sK_{x,h}\right],\end{aligned}$$

where $K_{x,h}(\cdot)=K\left(\frac{x-\cdot}{h}\right)$, as defined it in Section 4. Following the arguments from Section 4, let

$$\mathcal{F}^s_{K,[l_n,\infty)}:=\{D^sK_{x,h}:x\in\mathbb{X},h\geq l_n\}$$

be a class of unnormalized kernel functions centered on $\mathbb{X}$ and bandwidth greater than or equal to l_n, and let

$$\tilde{\mathcal{F}}^s_{K,[l_n,\infty)}:=\left\{\frac{1}{h^{d+|s|}}D^sK_{x,h}:x\in\mathbb{X},\,h\geq l_n\right\}$$

be a class of normalized kernel functions. Then $\sup_{h\geq l_n,x\in\mathbb{X}}|D^s\hat{p}_h(x)-D^sp_h(x)|$ can be rewritten as

$$\begin{aligned}&\sup_{h\geq l_n,x\in\mathbb{X}}|D^s\hat{p}_h(x)-D^sp_h(x)|\\&=\sup_{f\in\tilde{\mathcal{F}}^s_{K,[l_n,\infty)}}\left|\frac{1}{n}\sum_{i=1}^nf(X_i)-\mathbb{E}[f(X)]\right|.\end{aligned}\qquad(16)$$

To derive a good upper bound on $\sup_{h\geq l_n,x\in\mathbb{X}}|D^s\hat{p}_h(x)-D^sp_h(x)|$, it is important to first show a tight upper bound for $\mathbb{E}_P[(D^sK_{x,h})^2]$. Towards that end, we impose the following integrability condition.

Assumption 6. *The derivative of kernel is such that*

$$\int_0^\infty t^{d_{\mathrm{vol}}-1} \sup_{\|x\|\geq t} (D^s K)^2(x) dt < \infty. \tag{17}$$

Under Assumption 6, we can bound $\mathbb{E}_P[D^s K^2_{x,h}]$ in terms of the volume dimension d_{vol} as follows.

Lemma 19. *Let $(\mathbb{R}^d, P)$ be a probability space and let $X \sim P$. For any kernel K satisfying Assumption 6, the expectation of the square of the derivative of the kernel is upper bounded as*

$$\mathbb{E}_P\left[\left(D^s K\left(\frac{x-X}{h}\right)\right)^2\right] \leq C_{s,P,K,\epsilon} h^{d_{\mathrm{vol}}-\epsilon}, \tag{18}$$

for any $\epsilon \in (0, d_{\mathrm{vol}})$, where $C_{s,P,K,\epsilon}$ is a constant depending only on s, P, K, ϵ. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0 in (18).

To apply the VC type bound on (16), the function class $\mathcal{F}^s_{K,[l_n,\infty)}$ should be not too complex. Like in Section 4, we assume that $\mathcal{F}^s_{K,[l_n,\infty)}$ is a uniformly bounded VC-class.

Assumption 7. *Let $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel function with $\|D^s K\|_\infty, \|D^s K\|_2 < \infty$. We assume that*

$$\mathcal{F}^s_{K,[l_n,\infty)} := \{D^s K_{x,h} : x \in \mathbb{X}, h \geq l_n\}$$

is a uniformly bounded VC-class with dimension ν, i.e. there exists positive numbers A and ν such that, for every probability measure Q on $\mathbb{R}^d$ and for every $\epsilon \in (0, \|D^s K\|_\infty)$, the covering numbers $\mathcal{N}(\mathcal{F}^s_{K,[l_n,\infty)}, L_2(Q), \epsilon)$ satisfies

$$\mathcal{N}(\mathcal{F}^s_{K,[l_n,\infty)}, L_2(Q), \epsilon) \leq \left(\frac{A\|D^s K\|_\infty}{\epsilon}\right)^\nu.$$

Finally, to bound $\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)|$ with high probability, we combine the Talagrand inequality and VC type bound with Lemma 19. The following theorem provides a high probability upper bound for (16), and is analogous to Theorem 12.

Theorem 20. *Let P be a distribution and K be a kernel function satisfying Assumption 5, 6, and 7. Then, with probability at least $1-\delta$,*

$$\begin{aligned}&\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)| \\ &\leq C\left(\frac{(\log(1/l_n))_+}{n l_n^{d+|s|}} + \sqrt{\frac{(\log(1/l_n))_+}{n l_n^{2d+2|s|-d_{\mathrm{vol}}+\epsilon}}} \right. \\ &\quad \left. + \sqrt{\frac{\log(2/\delta)}{n l_n^{2d+2|s|-d_{\mathrm{vol}}+\epsilon}}} + \frac{\log(2/\delta)}{n l_n^{d+|s|}}\right),\end{aligned} \tag{19}$$

for any $\epsilon \in (0, d_{\mathrm{vol}})$, where C is a constant depending only on A, $\|D^s K\|_\infty$, d, ν, d_{vol}, $C_{s,P,K,\epsilon}$, ϵ. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0 in (19).

When l_n is not going to 0 too fast, then $\sqrt{\frac{\log(1/l_n)}{n l_n^{2d+2|s|-d_{\mathrm{vol}}}}}$ term dominates the upper bound in (19) as follows.

Corollary 21. *Let P be a distribution and K be a kernel function satisfying Assumption 5, 6, and 7. Suppose*

$$\limsup_n \frac{(\log(1/l_n))_+ + \log(2/\delta)}{n l_n^{d_{\mathrm{vol}}-\epsilon}} < \infty,$$

for fixed $\epsilon \in (0, d_{\mathrm{vol}})$. Then, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)| \\ &\leq C'\sqrt{\frac{(\log(1/l_n))_+ + \log(2/\delta)}{n l_n^{2d+2|s|-d_{\mathrm{vol}}+\epsilon}}},\end{aligned} \tag{20}$$

where C' is a constant depending only on A, $\|D^s K\|_\infty$, d, ν, d_{vol}, $C_{s,P,K,\epsilon}$, ϵ. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0.

We now turn to the case of a fixed bandwidth $h_n > 0$. We are interested in a high probability bound on

$$\sup_{x\in\mathbb{X}} |D^s\hat{p}_{h_n}(x) - D^s p_{h_n}(x)|.$$

Of course, Theorem 20 and Corollary 21 are applicable to the fixed bandwidth case.

But if the support of P is bounded, then, for a M_K-Lipschitz continuous derivative of kernel density estimator and fixed bandwidth, we can again derive a uniform convergence bound without the finite VC condition of (Giné & Guillou, 2001; Giné et al., 2004).

Lemma 22. *Suppose there exists $R > 0$ with $\mathbb{X} \subset \mathbb{B}_{\mathbb{R}^d}(0, R)$. Also, suppose that $D^s K$ is M_K-Lipschitz, i.e.*

$$\|D^s K(x) - D^s K(y)\|_2 \leq M_K \|x-y\|_2.$$

Then for all $\eta \in (0, \|D^s K\|_\infty)$, the supremum of the η-covering number $\mathcal{N}(\mathcal{F}^s_{K,h}, L_2(Q), \eta)$ over all measure Q is upper bounded as

$$\sup_Q \mathcal{N}(\mathcal{F}^s_{K,h}, L_2(Q), \eta) \leq \left(\frac{2RM_K h^{-1} + \|D^s K\|_\infty}{\eta}\right)^d.$$

Corollary 23. *Suppose there exists $R > 0$ with $\mathrm{supp}(P) = \mathbb{X} \subset \mathbb{B}_{\mathbb{R}^d}(0, R)$. Let K be a kernel function with M_K-Lipschitz continuous derivative satisfying Assumption 6. If*

$$\limsup_n \frac{(\log(1/h_n))_+ + \log(2/\delta)}{n h_n^{d_{\mathrm{vol}}-\epsilon}} < \infty,$$

for fixed $\epsilon \in (0, d_{\mathrm{vol}})$. Then, with probability at least $1-\delta$,

$$\sup_{x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)| \leq C''\sqrt{\frac{(\log(\frac{1}{h_n}))_+ + \log(\frac{2}{\delta})}{n h_n^{2d+2|s|-d_{\mathrm{vol}}+\epsilon}}}, \tag{21}$$

where C'' is a constant depending only on A, $\|D^s K\|_\infty$, d, M_k, d_{vol}, $C_{s,P,K,\epsilon}$, ϵ. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ϵ can be 0.

References

Ambrosio, L., Fusco, N., and Pallara, D. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000. ISBN 0-19-850245-1.

Arias-Castro, E., Mason, D., and Pelletier, B. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, 17(1):1487–1514, 2016.

Azizyan, M., Chen, Y.-C., Singh, A., and Wasserman, L. Risk bounds for mode clustering. *arXiv preprint arXiv:1505.00482*, 2015.

Balakrishnan, S., Narayanan, S., Rinaldo, A., Singh, A., and Wasserman, L. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, pp. 2679–2687, 2013.

Bousquet, O. A bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris, Ser. I*, 334:495–500, 2002.

Chacón, J. E. et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30 (4):518–532, 2015.

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.

Chen, Y.-C., Genovese, C. R., Ho, S., and Wasserman, L. Optimal ridge detection using coverage risk. In *Advances in Neural Information Processing Systems*, pp. 316–324, 2015a.

Chen, Y.-C., Genovese, C. R., Wasserman, L., et al. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5):1896–1928, 2015b.

Chen, Y.-C., Genovese, C. R., and Wasserman, L. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112(520): 1684–1696, 2017.

Einmahl, U., Mason, D. M., et al. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

Falconer, K. *Fractal geometry*. John Wiley & Sons, Ltd., Chichester, third edition, 2014. ISBN 978-1-119-94239-9. Mathematical foundations and applications.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., Singh, A., et al. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

Federer, H. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959. ISSN 0002-9947.

Giné, E. and Guillou, A. Laws of the iterated logarithm for censored data. *Ann. Probab.*, 27(4):2042–2067, 10 1999. doi: 10.1214/aop/1022874828. URL https://doi.org/10.1214/aop/1022874828.

Giné, E. and Guillou, A. On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 37(4):503 – 522, 2001. ISSN 0246-0203. doi: https://doi.org/10.1016/S0246-0203(01)01081-0. URL http://www.sciencedirect.com/science/article/pii/S0246020301010810.

Giné, E. and Guillou, A. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pp. 907–921. Elsevier, 2002.

Giné, E., Koltchinskii, V., and Zinn, J. Weighted uniform consistency of kernel density estimators. *Ann. Probab.*, 32(3B):2570–2605, 07 2004. doi: 10.1214/009117904000000063. URL https://doi.org/10.1214/009117904000000063.

Jiang, H. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pp. 1694–1703, 2017.

Kim, J., Chen, Y.-C., Balakrishnan, S., Rinaldo, A., and Wasserman, L. Statistical inference for cluster trees. In *Advances in Neural Information Processing Systems 29*, pp. 1839–1847. 2016.

Kim, J., Rinaldo, A., and Wasserman, L. A. Minimax rates for estimating the dimension of a manifold. *JoCG*, 10 (1):42–95, 2019. URL http://jocg.org/index.php/jocg/article/view/278.

Lee, J. A. and Verleysen, M. *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 978-0-3873-9350-6. URL https://books.google.com/books?id=o_TIoyeO7AsC&dq=isbn:038739351X&source=gbs_navlinks_s.

Mattila, P. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995. doi: 10.1017/CBO9780511623813.

Mattila, P., Morán, M., and Rey, J.-M. Dimension of a measure. *Studia Math.*, 142(3):219–233, 2000. ISSN 0039-3223. doi: 10.4064/sm-142-3-219-233. URL https://doi.org/10.4064/sm-142-3-219-233.

Niyogi, P., Smale, S., and Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3): 419–441, 2008. ISSN 0179-5376. doi: 10.1007/s00454-008-9053-2. URL http://dx.doi.org/10.1007/s00454-008-9053-2.

Nolan, D. and Pollard, D. u-processes: Rates of convergence. *Ann. Statist.*, 15(2):780–799, 06 1987. doi: 10.1214/aos/1176350374. URL https://doi.org/10.1214/aos/1176350374.

Parzen, E. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3): 1065–1076, 1962.

Pesin, Y. B. *Dimension theory in dynamical systems*. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 1997. ISBN 0-226-66221-7; 0-226-66222-5. doi: 10.7208/chicago/9780226662237.001.0001. URL https://doi.org/10.7208/chicago/9780226662237.001.0001. Contemporary views and applications.

Rao, B. P. *Nonparametric functional estimation*. Academic press, 1983.

Rinaldo, A. and Wasserman, L. Generalized density clustering. *Ann. Statist.*, 38(5):2678–2722, 2010. ISSN 0090-5364. doi: 10.1214/10-AOS797. URL https://doi.org/10.1214/10-AOS797.

Sriperumbudur, B. and Steinwart, I. Consistency and rates for clustering with dbscan. In Lawrence, N. D. and Girolami, M. (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 1090–1098, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL http://proceedings.mlr.press/v22/sriperumbudur12.html.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Steinwart, I., Sriperumbudur, B. K., and Thomann, P. Adaptive Clustering Using Kernel Density Estimators. *arXiv e-prints*, art. arXiv:1708.05254, August 2017.

Wand, M. P. and Jones, M. C. *Kernel smoothing*. Chapman and Hall/CRC, 1994.

Wang, D., Lu, X., and Rinaldo, A. Optimal rates for cluster tree estimation using kernel density estimators. *arXiv preprint arXiv:1706.03113*, 2017.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv e-prints*, art. arXiv:1707.00087, Jun 2017.