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

Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, Larry Wasserman

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

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

[1]Inria Saclay – Île-de-France
[2]Department of Statistics and Data Science, Carnegie Mellon University

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 and 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

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

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 and Wasserman, 2010, Wang et al., 2017], modal clustering and mode 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 and Guillou, 2002, Sriperumbudur and 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,

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

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 $1-\delta$,

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

for any $\varepsilon \in (0, d_{\mathrm{vol}})$, $\{l_n\}$ 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,

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

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 $h \geq l_n$, 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 and 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*

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

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 $0 \leq d_{\mathrm{vol}} \leq d_H$. Hence if A is a d_M-dimensional manifold, then $0 \leq d_{\mathrm{vol}} \leq d_M$. In particular, for any probability distribution P on $\mathbb{R}^d$, $0 \leq d_{\mathrm{vol}} \leq d$. 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*

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

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_{vol} be its volume dimension. We assume that*

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

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

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

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

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

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 $l_n > 0$:

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

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 $\tilde{\mathscr{F}}$, that is,

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

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

Proposition 8. *[Bousquet, 2002, Theorem 2.3], [Steinwart and 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 $\mathscr{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that is separable in $L_\infty(\mathbb{R}^d)$. Suppose all functions $f \in \mathscr{F}$ 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 $f \in \mathscr{F}$. Let

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

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, $\left\|K_{x,h}\right\|_\infty$, $\mathbb{E}_P[K_{x,h}^2]$, and

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

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 and Guillou [2001, Proposition 2.1].

Proposition 9. *(Giné and Guillou [2001, Proposition 2.1], [Sriperumbudur and 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 $\mathscr{F}$ 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 $f \in \mathscr{F}$, $\|f\|_\infty \leq B$, and the covering number $\mathscr{N}(\mathscr{F}, L_2(Q), \varepsilon)$ satisfies

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

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

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

By applying Proposition 8 and Proposition 9 to $\tilde{\mathscr{F}}_{K,[l_n,\infty)}$, 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 and 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*

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

for any $\varepsilon > 0$, 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*

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

for any $\varepsilon \in (0, d_{\mathrm{vol}})$, where $C_{k,P,K,\varepsilon}$ 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,

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

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

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

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 $\varepsilon \in (0, \|K\|_\infty)$, the covering numbers $\mathcal{N}(\mathscr{F}_{K,[l_n,\infty)}, L_2(Q), \varepsilon)$ satisfies

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

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 $\mathscr{F}_{K,[l_n,\infty)}$ to cover $\mathscr{F}_{K,[l_n,\infty)}$.

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,

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

This is implied by condition (K) in Giné et al. [2004] or condition (K_1) in Giné and 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 and 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 $\sup_{h \geq l_n, x \in \mathbb{X}} \left| \hat{p}_h(x) - p_h(x) \right|$.

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 $1 - \delta$,*

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

for any $\varepsilon \in (0, d_{\rm vol})$, where C is a constant depending only on A, $\|K\|_\infty$, d, ν, $d_{\rm vol}$, $C_{k=2,P,K,\varepsilon}$, ε. Further, if $d_{\rm vol} = 0$ 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 $\sqrt{\frac{\log(1/l_n)}{n l_n^{2d - d_{\rm vol}}}}$. 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 $\varepsilon \in (0, d_{\rm vol})$. Further, if $d_{\rm vol} = 0$ or under Assumption 1, ε can be 0. Suppose*

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

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

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

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

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

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

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

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 and 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 $\varepsilon \in (0, d_{\mathrm{vol}})$. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ε can be 0. Suppose*

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

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

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

where C'' is a constant depending only on R, M_K, $\|K\|_\infty$, d, ν, d_{vol}, $C_{k=2,P,K,\varepsilon}$, 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 $1-\delta$,

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

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 $1/\sqrt{nh_n^{2d-d_{\mathrm{vol}}}}$.

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

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

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 $l_n \to 0$ with $nl_n^{d_{\mathrm{vol}}} \to \infty$. Then, with probability $1-\delta$, the following holds for all large n:*

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

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

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

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 $R = 2$, $d_{\mathrm{vol}} = d - \beta$ and the kernel satisfies the integrability Assumption 3 with $k = 1,2$. It can be also shown that $\lim_{t\to 0}\inf_{\|x\|\le t} K(x) > 0$. Therefore, for small enough h_n, Corollary 15 and Proposition 16 imply*

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

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

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

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

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

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

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

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

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

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*

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

Under Assumption 6, we can bound $\mathbb{E}_P[D^sK^2_{x,h}]$ 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*

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

for any $\varepsilon \in (0, d_{\mathrm{vol}})$, where $C_{s,P,K,\varepsilon}$ 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 $\mathscr{F}^s_{K,[l_n,\infty)}$ should be not too complex. Like in Section 4, we assume that $\mathscr{F}^s_{K,[l_n,\infty)}$ 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*

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

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 $\varepsilon \in (0, \|D^s K\|_\infty)$, the covering numbers $\mathscr{N}(\mathscr{F}^s_{K,[l_n,\infty)}, L_2(Q), \varepsilon)$ satisfies

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

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 $1-\delta$,*

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

for any $\varepsilon \in (0, d_{\mathrm{vol}})$, where C is a constant depending only on A, $\|D^s K\|_\infty$, d, ν, d_{vol}, $C_{s,P,K,\varepsilon}$, 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*

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

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

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

where C' is a constant depending only on A, $\|D^s K\|_\infty$, d, ν, d_{vol}, $C_{s,P,K,\varepsilon}$, 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 and 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*

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

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

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

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

References

Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *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.

Ery Arias-Castro, David Mason, and Bruno Pelletier. 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.

Martin Azizyan, Yen-Chi Chen, Aarti Singh, and Larry Wasserman. Risk bounds for mode clustering. *arXiv preprint arXiv:1505.00482*, 2015.

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

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

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

Frédéric Chazal, Brittany T Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.

Yen-Chi Chen, Christopher R Genovese, Shirley Ho, and Larry Wasserman. Optimal ridge detection using coverage risk. In *Advances in Neural Information Processing Systems*, pages 316–324, 2015a.

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

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

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

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

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, Aarti Singh, et al. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42 (6):2301–2339, 2014.

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

Evarist Giné and Armelle Guillou. 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`.

Evarist Giné and Armelle Guillou. 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`.

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

Evarist Giné, Vladimir Koltchinskii, and Joel Zinn. 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`.

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

Jisu Kim, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, and Larry Wasserman. Statistical inference for cluster trees. In *Advances in Neural Information Processing Systems 29*, pages 1839–1847. 2016.

Jisu Kim, Alessandro Rinaldo, and Larry A. Wasserman. 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`.

John A. Lee and Michel Verleysen. *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`.

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

Pertti Mattila, Manuel Morán, and José-Manuel Rey. 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`.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. 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.

Deborah Nolan and David Pollard. *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.

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

Yakov B. Pesin. *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.

BLS Prakasa Rao. *Nonparametric functional estimation*. Academic press, 1983.

Alessandro Rinaldo and Larry Wasserman. 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.

Bharath Sriperumbudur and Ingo Steinwart. Consistency and rates for clustering with dbscan. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1090–1098, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL http://proceedings.mlr.press/v22/sriperumbudur12.html.

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

Ingo Steinwart, Bharath K. Sriperumbudur, and Philipp Thomann. Adaptive Clustering Using Kernel Density Estimators. *arXiv e-prints*, art. arXiv:1708.05254, August 2017.

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

Daren Wang, Xinyang Lu, and Alessandro Rinaldo. Optimal rates for cluster tree estimation using kernel density estimators. *arXiv preprint arXiv:1706.03113*, 2017.

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

SUPPLEMENTARY MATERIAL

A Backgrounds and Basic Definitions

First, we define the Hausdorff measure ([Pesin, 1997, Section 6], [Falconer, 2014, Section 2.2]), which is a generalization of the Lebesgue measure to lower dimensional subsets of $\mathbb{R}^d$. For a subset $A \subset \mathbb{R}^d$, we let $\text{diam}(A)$ be its diameter, that is

$$\text{diam}(A) = \sup\{\|x-y\| : x, y \in A\}.$$

Definition 2. *Fix $\nu > 0$ and $\delta > 0$. For any set $A \subset \mathbb{R}^d$, define H_δ^ν be*

$$H_\delta^\nu(A) := \inf\left\{\sum_{i=1}^{\infty}(\text{diam} U_i)^\nu : A \subset \bigcup_{i=1}^{\infty} U_i \text{ and } \text{diam}(U_i) < \delta\right\},$$

where the infimum is over all countable covers of A by sets $U_i \subset \mathbb{R}^d$ satisfying $\text{diam}(U_i) < \delta$. *Then, let the ν-*dimensional Hausdorff measure *H^ν be*

$$H^\nu(A) := \lim_{\delta \to 0} H_\delta^\nu(A).$$

Then, the Hausdorff dimension of a set is the infimum over dimensions that make the Hausdorff measure on that set to be 0.

Definition 3. *For any set $A \subset \mathbb{R}^d$, its* Hausdorff dimension *$d_H(A)$ is*

$$d_H(A) := \inf\{\nu : H^\nu(A) = 0\}.$$

We use the normalized ν-dimensional Hausdorff measure so that when ν is an integer, its measure on ν-dimensional unit cube is 1. This can be done by defining the normalized ν-dimensional Hausdorff measure λ_ν as

$$\lambda_\nu = \frac{\pi^{\frac{\nu}{2}}}{2^\nu \Gamma(\frac{\nu}{2}+1)} H^\nu.$$

Now, we define the reach, which is a regularity parameter in geometric measure theory. Given a closed subset $A \subset \mathbb{R}^d$, the medial axis of A, denoted by $\text{Med}(A)$, is the subset of $\mathbb{R}^d$ composed of the points that have at least two nearest neighbors on A. Namely, denoting by $d(x,A) = \inf_{q \in A} ||q - x||$ the distance function of a generic point x to A,

$$\text{Med}(A) = \left\{x \in \mathbb{R}^d \setminus A \,|\, \exists q_1 \neq q_2 \in A, ||q_1 - x|| = ||q_2 - x|| = d(x,A)\right\}. \tag{22}$$

The reach of A is then defined as the minimal distance from A to $\text{Med}(A)$.

Definition 4. *The* reach *of a closed subset $A \subset \mathbb{R}^d$ is defined as*

$$\tau_A = \inf_{q \in A} d\left(q, \text{Med}(A)\right) = \inf_{q \in A, x \in \text{Med}(A)} ||q - x||. \tag{23}$$

B Proof for Section 3

We show Lemma 4 first, which is a simple argument from the definition of d_{vol} in (4) in Definition 1.

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 $r > 0$,*

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

Proof of Lemma 4. From the definition of d_{vol} in (4) in Definition 1, $\nu \in [0, d_{\mathrm{vol}})$ implies that

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

Then there exist $r_0 > 0$ and $C'_{\nu,P} > 0$ such that for all $r \leq r_0$ and for all $x \in \mathbb{X}$,

$$\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu} \leq C'_{\nu,P}. \tag{24}$$

And for all $r \geq r_0$ and for all $x \in \mathbb{X}$,

$$\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu} \leq \frac{1}{r_0^\nu}. \tag{25}$$

Hence combining (24) and (25) gives that for all $r > 0$ and for all $x \in \mathbb{X}$,

$$\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu} \leq \max\left\{C'_{\nu,P}, \frac{1}{r_0^\nu}\right\}.$$

□

Then we can show Proposition 1 by using Lemma 4 and the definition of Hausdorff dimension in Definition 3.

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 $0 \leq d_{\mathrm{vol}} \leq d_H$. Hence if A is a d_M-dimensional manifold, then $0 \leq d_{\mathrm{vol}} \leq d_M$. In particular, for any probability distribution P on $\mathbb{R}^d$, $0 \leq d_{\mathrm{vol}} \leq d$. Also, if P has a point mass, i.e. there exists $x \in \mathbb{X}$ with $P(\{x\}) > 0$, then $d_{\mathrm{vol}} = 0$.*

Proof of Proposition 1. We first show $d_{\mathrm{vol}} \geq 0$. For any $x \in \mathbb{X}$ and $r \geq 0$,

$$\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^0} \leq 1 < \infty.$$

Hence $d_{\mathrm{vol}} \geq 0$ holds.

Now we show $d_{\text{vol}} \leq d_H = d_H(A)$. Fix any $\nu < d_{\text{vol}}$, and we will show that $H^{\nu}(A \cap \mathbb{X}) > 0$. Let $\{U_i\}$ be a countable cover of $A \cap \mathbb{X}$, i.e. $A \cap \mathbb{X} \subset \bigcup_{i=1}^{\infty} U_i$, and let $r_i = \text{diam}(U_i)$. For each i, we can assume that $U_i \cap (A \cap \mathbb{X}) \neq \emptyset$ and choose $x_i \in U_i \cap (A \cap \mathbb{X})$. Then $U_i \subset \overline{\mathbb{B}_{\mathbb{R}^d}(x_i, r_i)} \subset \mathbb{B}_{\mathbb{R}^d}(x_i, 2r_i)$, and hence

$$A \cap \mathbb{X} \subset \bigcup_{i=1}^{\infty} \mathbb{B}_{\mathbb{R}^d}(x_i, 2r_i).$$

Then with $x_i \in \mathbb{X}$, applying (5) from Lemma 4 gives

$$\begin{aligned} P(A \cap \mathbb{X}) < P\left(\bigcup_{i=1}^{\infty} \mathbb{B}_{\mathbb{R}^d}(x_i, 2r_i)\right) &= \sum_{i=1}^{\infty} P(\mathbb{B}_{\mathbb{R}^d}(x_i, 2r_i)) \\ &\leq \sum_{i=1}^{\infty} 2^{\nu} C_{\nu,P} r_i^{\nu}. \end{aligned}$$

Hence

$$\sum_{i=1}^{\infty} r_i^{\nu} \geq \frac{P(A \cap \mathbb{X})}{2^{\nu} C_{\nu,P}} > 0.$$

Since this holds for arbitrary covers of $A \cap \mathbb{X}$, $H_{\delta}^{\nu}(A \cap \mathbb{X}) \geq \frac{P(A \cap \mathbb{X})}{2^{\nu} C_{\nu,P}}$ for all $\delta > 0$. And $A \cap \mathbb{X} \subset A$ implies

$$H^{\nu}(A) \geq H^{\nu}(A \cap \mathbb{X}) = \lim_{\delta \to 0} H_{\delta}^{\nu}(A \cap \mathbb{X}) \geq \frac{P(A \cap \mathbb{X})}{2^{\nu} C_{\nu,P}} > 0.$$

Since this holds for arbitrary $\nu < d_{\text{vol}}$, the definition of Hausdorff dimension in Definition 3 gives that

$$d_H = \inf\{\nu : H^{\nu}(A) = 0\} \geq d_{\text{vol}}.$$

Now, if A is a d_M-dimensional manifold, then the Hausdorff dimension of A is d_M, and hence $0 \leq d_{\text{vol}} \leq d_M$ holds. In particular, setting $A = \mathbb{R}^d$ gives $0 \leq d_{\text{vol}} \leq d$ for all probability distributions. Also, if there exists $x \in \mathbb{X}$ with $P(\{x\}) > 0$, then setting $A = \{x\}$ gives $d_{\text{vol}} = 0$. □

Proposition 2 is again a simple argument from the definition of d_{vol} in (4) in Definition 1.

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*

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

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.

Proof of Proposition 2. It is enough to show for the case $m = 2$. Let $P := \lambda_1 P_1 + \lambda_2 P_2$.

We first show $d_{\text{vol}}(P) \geq \min\{d_{\text{vol}}(P_1), d_{\text{vol}}(P_2)\}$. Fix $\nu < \min\{d_{\text{vol}}(P_1), d_{\text{vol}}(P_2)\}$, then Definition 1 gives that

$$\limsup_{r \to 0} \sup_{x \in \mathbb{X}} \frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x, r))}{r^{\nu}}, \limsup_{r \to 0} \sup_{x \in \mathbb{X}} \frac{P_2(\mathbb{B}_{\mathbb{R}^d}(x, r))}{r^{\nu}} < \infty.$$

And hence

$$\limsup_{r\to 0}\sup_{x\in\mathbb{X}}\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}=\limsup_{r\to 0}\sup_{x\in\mathbb{X}}\left\{\frac{\lambda_1 P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}+\frac{\lambda_2 P_2(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}\right\}$$
$$\leq \lambda_1\limsup_{r\to 0}\sup_{x\in\mathbb{X}}\frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}+\lambda_2\limsup_{r\to 0}\sup_{x\in\mathbb{X}}\frac{P_2(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}<\infty.$$

And hence $d_{\mathrm{vol}}(P)\geq\min\{d_{\mathrm{vol}}(P_1),d_{\mathrm{vol}}(P_2)\}$ holds.

Next, we show $d_{\mathrm{vol}}(P)\leq\min\{d_{\mathrm{vol}}(P_1),d_{\mathrm{vol}}(P_2)\}$. Without loss of generality, suppose $d_{\mathrm{vol}}(P_1)\leq d_{\mathrm{vol}}(P_2)$, and fix $\nu>d_{\mathrm{vol}}(P_1)$. Then Definition 1 gives that

$$\limsup_{r\to 0}\sup_{x\in\mathbb{X}}\frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}=\infty.$$

Then from $P\geq\lambda_1 P_1$,

$$\limsup_{r\to 0}\sup_{x\in\mathbb{X}}\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}\geq\limsup_{r\to 0}\sup_{x\in\mathbb{X}}\frac{\lambda_1 P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^\nu}=\infty.$$

And hence $d_{\mathrm{vol}}(P)\leq d_{\mathrm{vol}}(P_1)=\min\{d_{\mathrm{vol}}(P_1),d_{\mathrm{vol}}(P_2)\}$ holds.

□

For Proposition 3 and 5, we need to bound the volume of the ball on the manifold. The following is rephrased from Lemma 3 in Kim et al. [2019].

Lemma 24. *Let $M\subset\mathbb{R}^d$ be a d_M-dimensional submanifold with reach τ_M. For a subset $U\subset M$ and $r<\tau_M$, let $U_r:=\{x\in\mathbb{R}^d:\mathrm{dist}(x,U)<r\}$ be an r-neighborhood of U in $\mathbb{R}^d$. Then*

$$\lambda_{d_M}(U)\leq\frac{d!}{d_M!}r^{d_M-d}\lambda_d(U_r).$$

Then, the following Lemma is by combining Lemma 5.3 in Niyogi et al. [2008] and Lemma 24.

Lemma 25. *Let $M\subset\mathbb{R}^d$ be a d_M-dimensional submanifold with reach τ_M. Then, for $x\in M$ and $r<\tau_M$,*

$$\left(1-\frac{r^2}{4\tau_M^2}\right)^{\frac{d_M}{2}}r^{d_M}\omega_d\leq\lambda_{d_M}(M\cap\mathbb{B}_{\mathbb{R}^d}(x,r))\leq\frac{d!}{d_M!}2^d r^{d_M}\omega_d.$$

Proof of Lemma 25. The LHS inequality is from Lemma 5.3 in Niyogi et al. [2008]. The RHS inequality is applying $U=M\cap\mathbb{B}_{\mathbb{R}^d}(x,r)$ to Lemma 24 and $\lambda_d(U_r)\leq\lambda_d(\mathbb{B}_{\mathbb{R}^d}(x,2r))=(2r)^d\omega_d$.

□

Now, we show Proposition 3 and 5 simultaneously via the following Proposition:

Proposition 26. *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 $d_{\mathrm{vol}} = d_M$, and Assumption 1 and 2 are satisfied. In particular, when P has a bounded density p with respect to the d-dimensional Lebesgue measure λ_d, then $d_{\mathrm{vol}} = d$, and Assumption 1 and 2 are satisfied.*

Proof for Proposition 26. Let τ_M be the reach of M.

We first show $d_{\mathrm{vol}} = d_M$ and Assumption 1. Since the density p is bounded, for all $x \in \mathbb{X}$ and $r > 0$, the probability on the ball $\mathbb{B}_{\mathbb{R}^d}(x,r)$ is bounded as

$$P(\mathbb{B}_{\mathbb{R}^d}(x,r)) \leq \|p\|_\infty \lambda_{d_M}(M \cap \mathbb{B}_{\mathbb{R}^d}(0,r)). \tag{26}$$

Then for $r < \tau_M$, Lemma 25 implies $\lambda_{d_M}(M \cap \mathbb{B}_{\mathbb{R}^d}(x,r)) \leq \frac{d!}{d_M!} 2^d r^{d_M} \omega_d$, and hence

$$\limsup_{r \to 0} \sup_{x \in \mathbb{X}} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_M}} \leq \|p\|_\infty \frac{d!}{d_M!} 2^d \omega_d < \infty, \tag{27}$$

which implies

$$d_{\mathrm{vol}} \geq d_M.$$

Then from Proposition 1,

$$d_{\mathrm{vol}} = d_M.$$

Now, (27) shows that Assumption 1 is satisfied.

For Assumption 2, define a density $q : \mathbb{R}^d \to \mathbb{R}$ as

$$q(x) = \lim_{r \to 0} \frac{\Gamma(\frac{d_M}{2}+1)}{\pi^{\frac{d_M}{2}}} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_M}}.$$

Since M is a submanifold with positive reach, P is λ_{d_M}-rectifiable. This imply that such limit $q(x)$ exists a.e. $[\lambda_{d_M}]$, and for any measurable set A,

$$P(A) = \int_{A \cap M} q(x) d\lambda_{d_M}(x).$$

See, for instance, Rinaldo and Wasserman [2010, Appendix], Mattila [1995, Corollary 17.9], or Ambrosio et al. [2000, Theorem 2.83]. Then from

$$P(M \cap \mathbb{X}) = \int_{M \cap \mathbb{X}} q(x) d\lambda_{d_M}(x) > 0,$$

there exists $x_0 \in M \cap \mathbb{X}$ with $q(x_0) > 0$. And hence

$$\sup_{x \in \mathbb{X}} \liminf_{r \to 0} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_M}} \geq q(x_0) > 0,$$

and hence Assumption 2 is satisfied.

□

The proof of Proposition 6 is simply checking the convexities for Assumption 1 and Assumption 2.

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

Proof of Proposition 6. Suppose P_1, P_2 are two probability distributions and $\lambda \in (0,1)$. Let $P := \lambda P_1 + (1-\lambda)P_2$. Then Proposition 2 implies that

$$d_{\mathrm{vol}}(P) = \min\{d_{\mathrm{vol}}(P_1), d_{\mathrm{vol}}(P_2)\}.$$

Consider Assumption 1 first. Suppose P_1 and P_2 satisfies Assumption 1. Then for all $x \in \mathbb{X}$ and $r \leq 1$, applying $d_{\mathrm{vol}}(P_1) \leq d_{\mathrm{vol}}(P_1), d_{\mathrm{vol}}(P_2)$ gives

$$\begin{aligned}\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P)}} &= \lambda \frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P)}} + (1-\lambda)\frac{P_2(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P)}} \\ &\leq \lambda \frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P_1)}} + (1-\lambda)\frac{P_2(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P_2)}}.\end{aligned}$$

Hence,

$$\begin{aligned}\limsup_{r\to 0}\sup_{x\in\mathbb{X}} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P)}} &\leq \limsup_{r\to 0}\sup_{x\in\mathbb{X}} \left\{\lambda \frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P_1)}} + (1-\lambda)\frac{P_2(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P_2)}}\right\} \\ &\leq \lambda \limsup_{r\to 0}\sup_{x\in\mathbb{X}} \frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P_1)}} + (1-\lambda)\limsup_{r\to 0}\sup_{x\in\mathbb{X}} \frac{P_2(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}(P_2)}} \\ &< \infty,\end{aligned}$$

and Assumption 1 is satisfied for $P = \lambda P_1 + (1-\lambda)P_2$.

Now, consider Assumption 2. Suppose P_1 and P_2 satisfies Assumption 1, and without loss of generality, assume $d_{\mathrm{vol}}(P_1) \leq d_{\mathrm{vol}}(P_2)$. Then there exists $x_0 \in \mathbb{X}$ such that

$$\liminf_{r\to 0} \frac{P_1(\mathbb{B}_{\mathbb{R}^d}(x_0,r))}{r^{d_{\mathrm{vol}}(P_1)}} > 0.$$

Then $P \geq \lambda P_1$ and $d_{\mathrm{vol}}(P) = d_{\mathrm{vol}}(P_1)$ give

$$\liminf_{r\to 0} \frac{P(\mathbb{B}_{\mathbb{R}^d}(x_0,r))}{r^{d_{\mathrm{vol}}(P)}} \geq \liminf_{r\to 0} \frac{\lambda P_1(\mathbb{B}_{\mathbb{R}^d}(x_0,r))}{r^{d_{\mathrm{vol}}(P_1)}} > \infty.$$

Hence

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

and Assumption 2 is satisfied for $P = \lambda P_1 + (1-\lambda)P_2$.

□

C Volume Dimension and Other Dimensions

In this section, we compare the volume dimension with other various dimensions.

For a set, one commonly used dimension other than the Hausdorff dimension is the box dimension ([Pesin, 1997, Section 6], [Falconer, 2014, Section 3.1]). This has various names as Kolmogorov entropy, entropy dimension, capacity dimension, metric dimension, logarithmic density or Minkowski dimension.

Definition 5. *For any set $A \subset \mathbb{R}^d$ and $\delta > 0$, let $N(A,\delta)$ be the smallest number of balls of radius δ to cover A. Then the* lower box dimension *of A is defined as*

$$d_B^-(A) := \liminf_{\delta \to 0} \frac{\log N(A,\delta)}{-\log \delta},$$

and the upper box dimension *of A is defined as*

$$d_B^+(A) := \limsup_{\delta \to 0} \frac{\log N(A,\delta)}{-\log \delta}.$$

The Hausdorff dimension and the lower and upper box dimensions are related as [Pesin, 1997, Theorem 6.2 (2)]:

$$\text{forall } A \subset \mathbb{R}^d,\, d_H(A) \le d_B^-(A) \le d_B^+(A). \tag{28}$$

So far the Hausdorff dimension in Section A and the box dimension is defined for a set. For a probability distribution, there are two ways for natural extension. One way is to take the infimum of the set dimensions over all sets with positive probabilities ([Mattila et al., 2000, Section 2], [Falconer, 2014, Section 13.7]). We will use this as the definition of the Hausdorff dimension and the box dimension.

Definition 6. *Let P be a probability distribution on $\mathbb{R}^d$. Its* Hausdorff dimension *$d_H(P)$ is the infimum of the Hausdorff dimensions over a set with positive probability, i.e.,*

$$d_H(P) := \inf_{A:P(A)>0} d_H(A).$$

Similarly, the lower box dimension *$d_B^-(P)$ and the* upper box dimension *$d_B^+(P)$ is the infimum of the lower box dimensions and the upper box dimensions, respectively, over a set with positive probability, i.e.*

$$d_B^-(P) := \inf_{A:P(A)>0} d_B^-(A),$$
$$d_B^+(P) := \inf_{A:P(A)>0} d_B^+(A).$$

Another way is to take the infimum of the set dimensions over all sets with probabilities 1 [Pesin, 1997, Section 6]. We will denote these dimensions as Hausdorff support dimension and the box support dimension to differentiate from the previous dimensions.

Definition 7. *Let P be a probability distribution on $\mathbb{R}^d$. Its* Hausdorff support dimension $d_{HS}(P)$ *is the infimum of the Hausdorff dimensions over a set with probability* 1*, i.e.,*

$$d_{HS}(P) := \inf_{A:P(A)=1} d_H(A).$$

Similarly, the lower box dimension $d_{BS}^-(P)$ *and the* upper box dimension $d_{BS}^+(P)$ *is the infimum of the lower box dimensions and the upper box dimensions, respectively, over a set with positive probability, i.e.*

$$d_{BS}^-(P) := \inf_{A:P(A)=1} d_B^-(A),$$
$$d_{BS}^+(P) := \inf_{A:P(A)=1} d_B^+(A).$$

The volume dimension, the Hausdorff dimension, and the lower and upper box dimensions have the following relations.

Proposition 27. *Let P be a probability distribution on $\mathbb{R}^d$ with $P(\mathbb{X}) > 0$. Then its volume dimension, Hausdorff dimension, lower and upper box dimension, Hausdorff support dimension, and lower and upper box support dimension satisfy the following inequality:*

$$d_{\mathrm{vol}}(P) \le d_H(P) \le d_B^-(P) \le d_B^+(P),$$

and

$$d_{\mathrm{vol}}(P) \le d_{HS}(P) \le d_{BS}^-(P) \le d_{BS}^+(P).$$

Proof. Since $P(\mathrm{supp}(P) \cap \mathbb{X}) = P(\mathbb{X}) > 0$, $d_{\mathrm{vol}}(P) \le d_H(P)$ is direct from Proposition 1. Now, combining this with $d_H(P) \le d_B^-(P) \le d_B^+(P)$ and $d_{HS}(P) \le d_{BS}^-(P) \le d_{BS}^+(P)$ from (28) and that $d_H(P) \le d_{HS}(P)$ gives the statement.

□

Now, we introduce the q-dimension, which generalizes the box support dimension [Lee and Verleysen, 2007, Section 3.2.1].

Definition 8. *Let P be a probability distribution on $\mathbb{R}^d$. For $q \ge 0$ and $\delta > 0$, define $C_q(P, \delta)$ as*

$$C_q(P, \delta) := \int [P(\overline{\mathbb{B}_{\mathbb{R}^d}(x, \delta)})]^{q-1} dP(x).$$

Now for $q \ge 0$ and $q \ne 1$, the lower q-dimension *of P is*

$$d_q^-(P) := \liminf_{\delta \to 0} \frac{\log C_q(P, \delta)}{(q-1)\log \delta},$$

and the upper q-dimension *of P is*

$$d_q^+(P) := \limsup_{\delta \to 0} \frac{\log C_q(P, \delta)}{(q-1)\log \delta}.$$

For $q = 1$, we understand in the limit sense, i.e., $d_1^-(P) = \lim_{q \to 1} d_q^-(P)$ and $d_1^+(P) = \lim_{q \to 1} d_q^+(P)$.

This q-dimension is a generalization of the box support dimension in the sense that when $q = 0$, the lower and upper q-dimensions reduce to the lower and upper box support dimensions, respectively, i.e. $d_0^-(P) = d_{BS}^-(P)$ and $d_0^+(P) = d_{BS}^+(P)$ Pesin [1997, Section 8]. When $q = 1$, the q-dimension is called the information dimension, and when $q = 2$, the q-dimension is called the correlation dimension.

The volume dimension and the q-dimension have the following relation.

Proposition 28. *Let P be a probability distribution on $\mathbb{R}^d$ with $P(\mathbb{X}) = 1$. Then for any $q \geq 0$, the volume dimension and the q-dimension has the following inequality:*

$$d_{\mathrm{vol}}(P) \leq d_q^-(P) \leq d_q^+(P).$$

Proof. Since $d_q^-(P) \leq d_q^+(P)$ is obvious, we only need to show $d_{\mathrm{vol}}(P) \leq d_q^-(P)$.

Fix any $\nu < d_{\mathrm{vol}}(P)$. Then from $P(\mathbb{X}) = 1$, $C_q(P, \delta)$ can be expressed as taking an integration over $\mathbb{X}$. Hence applying (5) from Lemma 4 gives

$$\begin{aligned} C_q(P, \delta) &= \int_{\mathbb{X}} [P(\overline{B_{\mathbb{R}^d}(x, \delta)})]^{q-1} dP(x) \\ &\leq \int_{\mathbb{X}} [P(B_{\mathbb{R}^d}(x, 2\delta))]^{q-1} dP(x) \\ &\leq (2^\nu C_{\nu,P} \delta^\nu)^{q-1}. \end{aligned}$$

And hence $d_q^-(P)$ is lower bounded as

$$\begin{aligned} d_q^-(P) &= \liminf_{\delta \to 0} \frac{\log C_q(P, \delta)}{(q-1)\log \delta} \geq \liminf_{\delta \to 0} \frac{\log(2^\nu C_{\nu,P} \delta^\nu)}{\log \delta} \\ &= \nu + \liminf_{\delta \to 0} \frac{\log(2^\nu C_{\nu,P})}{\log \delta} = \nu. \end{aligned}$$

Since this holds for arbitrary $\nu < d_{\mathrm{vol}}(P)$, we have

$$d_{\mathrm{vol}}(P) \leq d_q^-(P).$$

□

We end this section by comparing the volume dimension and the Wasserstein dimension [Weed and Bach, 2017, Definition 4].

Definition 9. *Let P be a probability distribution on $\mathbb{R}^d$. For any $\delta > 0$ and $\tau \in [0, 1]$, let the (δ, τ)-covering number of P be*

$$N(P, \delta, \tau) := \inf\{N(A, \delta) : P(A) \geq 1 - \tau\},$$

and let the (δ, τ)-dimension be

$$d_\delta(P, \tau) := \frac{\log N(P, \delta, \tau)}{-\log \delta}.$$

Then for a fixed $p>0$, the lower and upper Wasserstein dimensions are respectively,

$$d_*(P)=\lim_{\tau\to 0}\liminf_{\delta\to 0} d_\delta(P,\tau)$$
$$d_p^*(P)=\inf\{s\in(2p,\infty):\limsup_{\delta\to 0} d_\varepsilon(P,\delta^{\frac{sp}{s-2p}})\le s\}.$$

Proposition 29. *Let P be a probability distribution on $\mathbb{R}^d$ with $P(\mathbb{X})>0$. Then its volume dimension and lower and upper Wasserstein dimensions satisfy the following inequality:*

$$d_{\mathrm{vol}}(P)\le d_{HS}(P)\le d_*(P)\le d_p^*(P).$$

Proof. Since $P(\mathrm{supp}(P)\cap\mathbb{X})=P(\mathbb{X})>0$, $d_{\mathrm{vol}}(P)\le d_H(P)$ is direct from Proposition 1. The inequality $d_H(P)\le d_*(P)\le d_p^*(P)$ is from Weed and Bach [2017, Proposition 2]. □

D Uniform convergence on a function class

As we have seen in (8) in Section 4, uniform bound on the kernel density estimator $\sup_{h\ge l_n,x\in\mathbb{X}}\left|\hat{p}_h(x)-p_h(x)\right|$ boils down to uniformly bounding on the function class $\sup_{f\in\tilde{\mathscr{F}}_{K,[l_n,\infty)}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X)]\right|$. In this section, we derive a uniform convergence for a more general class of functions. Let $\mathscr{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$, and consider a random variable

$$\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X)]\right|. \tag{29}$$

As discussed in Section 4, we combine the Talagrand inequality (Theorem 8) and VC type bound (Theorem 9) to bound (29), which is generalizing the approach in Sriperumbudur and Steinwart [2012, Theorem 3.1].

Theorem 30. *Let $(\mathbb{R}^d,P)$ be a probability space and let $X_1,\dots,X_n$ be i.i.d. from P. Let $\mathscr{F}$ 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 $f\in\mathscr{F}$, $\|f\|_\infty\le B$, and for every probability measure Q on $\mathbb{R}^d$ and for every $\varepsilon\in(0,B)$, the covering number $\mathscr{N}(\mathscr{F},L_2(Q),\varepsilon)$ satisfies*

$$\mathscr{N}(\mathscr{F},L_2(Q),\varepsilon)\le\left(\frac{AB}{\varepsilon}\right)^\nu.$$

Let $\sigma>0$ with $\mathbb{E}_P f^2\le\sigma^2$ for all $f\in\mathscr{F}$. Then there exists a universal constant C not depending on any parameters such that $\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X)]\right|$ is upper bounded with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X)]\right|\\ &\le C\left(\frac{\nu B}{n}\log\left(\frac{2AB}{\sigma}\right)+\sqrt{\frac{\nu\sigma^2}{n}\log\left(\frac{2AB}{\sigma}\right)}+\sqrt{\frac{\sigma^2\log(\frac{1}{\delta})}{n}}+\frac{B\log(\frac{1}{\delta})}{n}\right).\end{aligned}$$

Proof of Theorem 30. Let $\mathscr{G} := \{f - \mathbb{E}_P[f] : f \in \mathscr{F}\}$. Then it is immediate to check that for all $g \in \mathscr{G}$,

$$\begin{aligned}
\mathbb{E}_P g &= \mathbb{E}_P f - \mathbb{E}_P f = 0, \\
\mathbb{E}_P g^2 &= \mathbb{E}_P (f - \mathbb{E}_P f)^2 \le \mathbb{E}_P f^2 \le \sigma^2, \\
\|g\|_\infty &\le \|f\|_\infty + \mathbb{E}_P f \le 2B.
\end{aligned} \tag{30}$$

Now, $\sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right|$ is expanded as

$$\sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right| = \sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right|.$$

Hence from (30), applying Proposition 8 to above gives the probabilistic bound on $\sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right|$ as

$$P\left(\sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right| < 4\mathbb{E}_P \sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right| + \sqrt{\frac{2\sigma^2 \log(\frac{1}{\delta})}{n}} + \frac{2B\log(\frac{1}{\delta})}{n}\right) \ge 1-\delta. \tag{31}$$

It thus remains to bound the term $\mathbb{E}_P \sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right|$. Let $\tilde{\mathscr{F}} := \{f - a : f \in \mathscr{F}, a \in [-B,B]\}$. Then $\mathscr{F}$ being a uniform VC-class with dimension ν implies that for all $\varepsilon \in (0,B)$,

$$\begin{aligned}
\sup_P \mathscr{N}\left(\tilde{\mathscr{F}}, L_2(P), \varepsilon\right) &\le \sup_P \mathscr{N}\left(\mathscr{F}, L_2(P), \frac{\varepsilon}{2}\right) \sup_P \mathscr{N}\left([-B,B], |\cdot|, \frac{\varepsilon}{2}\right) \\
&\le \left(\frac{2AB}{\varepsilon}\right)^{\nu+1}.
\end{aligned}$$

Hence from (30), applying Proposition 9 yields the upper bound for $\mathbb{E}_P \sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right|$ as

$$\mathbb{E}_{P^n} \sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_i)\right| \le 2C\left(\frac{2(\nu+1)B}{n}\log\left(\frac{2AB}{\sigma}\right) + \sqrt{\frac{(\nu+1)\sigma^2}{n}\log\left(\frac{2AB}{\sigma}\right)}\right). \tag{32}$$

Hence applying (32) to (31) yields that, $\sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right|$ is upper bounded with probability at least $1-\delta$ as

$$\begin{aligned}
&\sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right| \\
&\le 4C\left(\frac{2(\nu+1)B}{n}\log\left(\frac{2AB}{\sigma}\right) + \sqrt{\frac{(\nu+1)\sigma^2}{n}\log\left(\frac{2AB}{\sigma}\right)}\right) \\
&\quad + \sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{n}} + \frac{2B\log(\frac{1}{\delta})}{n} \\
&\le 16C\left(\frac{\nu B}{n}\log\left(\frac{2AB}{\sigma}\right) + \sqrt{\frac{\nu\sigma^2}{n}\log\left(\frac{2AB}{\sigma}\right)} + \sqrt{\frac{\sigma^2\log(\frac{1}{\delta})}{n}} + \frac{B\log(\frac{1}{\delta})}{n}\right).
\end{aligned}$$

□

E Proof for Section 4

Lemma 11 is shown by the calculation using integral by parts and change of variables.

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*

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

for any $\varepsilon \in (0, d_{\mathrm{vol}})$, where $C_{k,P,K,\varepsilon}$ 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).

Proof of Lemma 11. We first consider the case when $d_{\mathrm{vol}} = 0$. Then $\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right]$ is simply bounded as

$$\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right] \leq \|K\|_\infty^k h^0.$$

Now, we consider the case when $d_{\mathrm{vol}} > 0$. Fix $\varepsilon \in (0, d_{\mathrm{vol}})$. Under Assumption 1, ε can be chosen to be 0.

Let $C_{k,K,d_{\mathrm{vol}},\varepsilon} := \int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t} |K(x)|^k dt$, then it is finite from (10) and $\|K\|_\infty < \infty$ in Assumption 4 as

$$\begin{aligned}\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t} |K(x)|^k dt &\leq \int_0^1 t^{d_{\mathrm{vol}}-\varepsilon-1} \|K\|_\infty dt + \int_1^\infty t^{d_{\mathrm{vol}}-1} \sup_{\|x\|\leq t} |K(x)|^k dt \\ &\leq \frac{\|K\|_\infty}{d_{\mathrm{vol}}-\varepsilon} + \int_0^\infty t^{d_{\mathrm{vol}}-1} \sup_{\|x\|\leq t} |K(x)|^k dt < \infty.\end{aligned}$$

Fix $\eta > 0$, and let $\tilde{K}_\eta : [0,\infty) \to \mathbb{R}$ be a continuous and strictly decreasing function satisfying $\tilde{K}_\eta(t) > \sup_{\|x\|\geq t} |K(x)|^k$ for all $t \geq 0$ and $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}(\tilde{K}_\eta(t) - \sup_{\|x\|\geq t} |K(x)|^k) dt = \eta$. Such existence is possible since $t \mapsto \sup_{\|x\|\geq t} |K(x)|^k$ is nonincreasing function, so have at most countable discontinuous points, and $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t} |K(x)|^k dt < \infty$. Then it is immediate to check that

$$|K(x)|^k < \tilde{K}_\eta(\|x\|) \text{ for all } x \in \mathbb{R}. \tag{33}$$

Then $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \tilde{K}(t) dt$ can be expanded as

$$\begin{aligned}\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \tilde{K}_\eta(t) dt &= \int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t} |K(x)|^k dt + \int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} (\tilde{K}_\eta(t) - \sup_{\|x\|\geq t} |K(x)|^k) dt \\ &= C_{k,K,d_{\mathrm{vol}},\varepsilon} + \eta < \infty.\end{aligned} \tag{34}$$

Now since $\tilde{K}_\eta$ is continuous and strictly decreasing, change of variables $t = \tilde{K}_\eta(u)$ is applicable, and then $\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right]$ can be expanded as

$$\begin{aligned}\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right] &= \int_0^\infty P\left(\left|K\left(\frac{x-X}{h}\right)\right|^k > t\right) dt \\ &= \int_\infty^0 P\left(\left|K\left(\frac{x-X}{h}\right)\right|^k > \tilde{K}_\eta(u)\right) d\tilde{K}_\eta(u).\end{aligned}$$

Now, from (33) and $\tilde{K}_\eta$ being a strictly decreasing, we can upper bound $\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right]$ as

$$\begin{aligned}\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right] &\leq \int_\infty^0 P\left(\tilde{K}_\eta\left(\frac{\|x-X\|}{h}\right) > \tilde{K}_\eta(u)\right) d\tilde{K}_\eta(u) \\ &= \int_\infty^0 P\left(\frac{\|x-X\|}{h} < u\right) d\tilde{K}_\eta(u) \\ &= \int_\infty^0 P\left(\mathbb{B}_{\mathbb{R}^d}(x, hu)\right) d\tilde{K}_\eta(u).\end{aligned}$$

Now, from Lemma 4 (and (6) for Assumption 1 case), there exists $C_{d_{\mathrm{vol}}-\varepsilon,P} < \infty$ with $P\left(\mathbb{B}_{\mathbb{R}^d}(x,r)\right) \leq C_{d_{\mathrm{vol}}-\varepsilon,P}\, r^{d_{\mathrm{vol}}-\varepsilon}$ for all $x \in \mathbb{X}$ and $r > 0$. Then $\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right]$ is further upper bounded as

$$\begin{aligned}\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right] &\leq \int_\infty^0 C_{d_{\mathrm{vol}}-\varepsilon,P}(hu)^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u) \\ &= C_{d_{\mathrm{vol}}-\varepsilon,P}\, h^{d_{\mathrm{vol}}-\varepsilon} \int_\infty^0 u^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u). \qquad (35)\end{aligned}$$

Now, $\int_\infty^0 u^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u)$ can be computed using integration by part. Note first that $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(t)dt < \infty$ implies

$$\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t) = 0.$$

To see this, note that $t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)$ is expanded as

$$t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t) = \int_0^t u^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u) + \int_0^t (d_{\mathrm{vol}}-\varepsilon) u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u) du,$$

then $\int_0^\infty (d_{\mathrm{vol}}-\varepsilon) u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u) du < \infty$ and $\int_0^t u^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u)$ being monotone function of t imply that $\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)$ exists. Now, suppose $\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t) = a > 0$, then we can choose $t_0 > 0$ such that $t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t) > \frac{a}{2}$ for all $t \geq t_0$, and then

$$\infty > \int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(t)dt \geq \int_{t_0}^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(t)dt \geq \frac{a}{2}\int_{t_0}^\infty t^{-1}dt = \infty,$$

which is a contradiction. Hence $\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)=0$. Now, applying integration by part to $\int_{\infty}^{0} u^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u)$ with $d_{\mathrm{vol}}-\varepsilon>0$ gives

$$\begin{aligned}\int_{\infty}^{0} u^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u) &= \left[u^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(u)\right]_{\infty}^{0} - \int_{\infty}^{0} (d_{\mathrm{vol}}-\varepsilon) u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u)du \\ &= \int_{0}^{\infty} (d_{\mathrm{vol}}-\varepsilon) u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u)du. \end{aligned} \tag{36}$$

Then applying (34) and (36) to (35) gives an upper bound for $\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right]$ as

$$\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right] \leq C_{d_{\mathrm{vol}}-\varepsilon,P}(d_{\mathrm{vol}}-\varepsilon)h^{d_{\mathrm{vol}}-\varepsilon}(C_{k,K,d_{\mathrm{vol}},\varepsilon}+\eta). \tag{37}$$

And then note that RHS of (37) holds for any $\eta>0$, and hence $\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right]$ is further upper bounded as

$$\begin{aligned}\mathbb{E}_P\left[\left|K\left(\frac{x-X}{h}\right)\right|^k\right] &\leq \inf_{\eta>0}\left\{C_{d_{\mathrm{vol}}-\varepsilon,P}(d_{\mathrm{vol}}-\varepsilon)h^{d_{\mathrm{vol}}-\varepsilon}(C_{k,K,d_{\mathrm{vol}},\varepsilon}+\eta)\right\} \\ &= C_{d_{\mathrm{vol}}-\varepsilon,P}(d_{\mathrm{vol}}-\varepsilon)C_{k,K,d_{\mathrm{vol}},\varepsilon}h^{d_{\mathrm{vol}}-\varepsilon} \\ &= C_{k,P,K,\varepsilon}h^{d_{\mathrm{vol}}-\varepsilon},\end{aligned}$$

where $C_{k,P,K,\varepsilon}=C_{d_{\mathrm{vol}}-\varepsilon,P}(d_{\mathrm{vol}}-\varepsilon)C_{k,K,d_{\mathrm{vol}},\varepsilon}$.

□

E.1 Proof for Section 4.1

Theorem 12 follows from applying Theorem 30.

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* $1-\delta$,

$$\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x)-p_h(x)| \leq C\left(\frac{(\log(1/l_n))_+}{nl_n^d}+\sqrt{\frac{(\log(1/l_n))_+}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\sqrt{\frac{\log(2/\delta)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\log(2/\delta)}{nl_n^d}\right),$$

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

Proof of Theorem 12. For $x\in\mathbb{X}$ and $h\geq l_n$, let $K_{x,h}:\mathbb{R}^d\to\mathbb{R}$ be $K_{x,h}(\cdot)=K\left(\frac{x-\cdot}{h}\right)$, and let $\tilde{\mathscr{F}}_{K,[l_n,\infty)}:=\left\{\frac{1}{h^d}K_{x,h}: x\in\mathbb{X}, h\geq l_n\right\}$ be a class of normalized kernel functions centered on $\mathbb{X}$ and bandwidth in $[l_n,\infty)$. Note that $\hat{p}_h(x)-p_h(x)$ can be expanded as

$$\hat{p}_h(x)-p_h(x)=\frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)-\mathbb{E}_P\left[\frac{1}{h^d}K\left(\frac{x-X_i}{h}\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_{x,h}(X_i)-\mathbb{E}_P\left[\frac{1}{h^d}K_{x,h}\right].$$

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

$$\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)| = \sup_{f\in\tilde{\mathscr{F}}_{K,[l_n,\infty)}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_P[f(X)] \right|. \tag{38}$$

Now, it is immediate to check that

$$\|f\|_\infty \leq l_n^{-d} \|K\|_\infty. \tag{39}$$

For bounding the VC dimension of $\tilde{\mathscr{F}}_{K,[l_n,\infty)}$, consider $\mathscr{F}_{K,[l_n,\infty)} := \left\{K_{x,h} : x\in\mathbb{X}, h\geq l_n\right\}$ be a class of unnormalized kernel functions centered on $\mathbb{X}$ and bandwidth in $[l_n,\infty)$. Fix $\eta < l_n^{-d}\|K\|_\infty$ and a probability measure Q on $\mathbb{R}^d$. Suppose $\left[l_n, \left(\frac{\eta}{2\|K\|_\infty}\right)^{-1/d}\right]$ is covered by balls $\left\{\left(h_i - \frac{l_n^{d+1}\eta}{2d\|K\|_\infty}, h_i + \frac{l_n^{d+1}\eta}{2d\|K\|_\infty}\right) : 1\leq i\leq N_1\right\}$ and $(\mathscr{F}_{K,[l_n,\infty)}, L_2(Q))$ is covered by balls $\left\{\mathbb{B}_{L_2(Q)}\left(f_j, \frac{l_n^d\eta}{2}\right) : 1\leq j\leq N_2\right\}$, and let $f_{i,j} := h_i^{-d} f_j$ for $1\leq i\leq N_1$ and $1\leq j\leq N_2$. Also, choose $h_0 > \left(\frac{\eta}{2\|K\|_\infty}\right)^{-1/d}$, $x_0\in\mathbb{X}$, and let $f_0 = \frac{1}{h_0^d} K_{x_0,h_0}$. We will show that

$$\left\{\mathbb{B}_{L_2(Q)}\left(f_{i,j},\eta\right) : 1\leq i\leq N_1, 1\leq j\leq N_2\right\} \cup \left\{\mathbb{B}_{L_2(Q)}\left(f_0,\eta\right)\right\} \text{ covers } \tilde{\mathscr{F}}_{K,[l_n,\infty)}. \tag{40}$$

For the first case when $h \leq \left(\frac{\eta}{\|K\|_\infty}\right)^{-1/d}$, find h_i and f_j with $h\in\left(h_i - \frac{l_n^{d+1}\eta}{2d\|K\|_\infty}, h_i + \frac{l_n^{d+1}\eta}{2d\|K\|_\infty}\right)$ and $K_{x,h}\in\mathbb{B}_{L_2(Q)}\left(f_j, \frac{l_n^d\eta}{2}\right)$. Then the distance between $\frac{1}{h^d}K_{x,h}$ and $\frac{1}{h_i^d}f_j$ is upper bounded as

$$\left\|\frac{1}{h^d}K_{x,h} - \frac{1}{h_i^d}f_j\right\|_{L_2(Q)} \leq \left\|\frac{1}{h^d}K_{x,h} - \frac{1}{h_i^d}K_{x,h}\right\|_{L_2(Q)} + \left\|\frac{1}{h_i^d}K_{x,h} - \frac{1}{h_i^d}f_j\right\|_{L_2(Q)}. \tag{41}$$

Now, the first term of (41) is upper bounded as

$$\begin{aligned}
\left\|\frac{1}{h^d}K_{x,h} - \frac{1}{h_i^d}K_{x,h}\right\|_{L_2(Q)} &= \left|\frac{1}{h^d} - \frac{1}{h_i^d}\right| \left\|K_{x,h}\right\|_{L_2(Q)} \\
&= |h_i - h| \sum_{k=0}^{d-1} h_i^{k-d} h^{-1-k} \left\|K_{x,h}\right\|_{L_2(Q)} \\
&\leq |h_i - h|\, d l_n^{-d-1} \|K\|_\infty < \frac{\eta}{2}.
\end{aligned} \tag{42}$$

Also, the second term of (41) is upper bounded as

$$\begin{aligned}
\left\|\frac{1}{h_i^d}K_{x,h} - \frac{1}{h_i^d}f_j\right\|_{L_2(Q)} &= \frac{1}{h_i^d}\left\|K_{x,h} - f_j\right\|_{L_2(Q)} \\
&\leq l_n^{-d}\left\|K_{x,h} - f_j\right\|_{L_2(Q)} < \frac{\eta}{2}.
\end{aligned} \tag{43}$$

Hence applying (42) and (43) to (41) gives

$$\left\| \frac{1}{h^d} K_{x,h} - \frac{1}{h_i^d} f_j \right\|_{L_2(Q)} < \eta.$$

For the second case when $h > \left(\frac{\eta}{2\|K\|_\infty}\right)^{-1/d}$, $\left\|\frac{1}{h^d}K_{x,h}\right\|_{L_2(Q)} \leq \left\|\frac{1}{h^d}K_{x,h}\right\|_\infty < \frac{\eta}{2}$ holds, and hence

$$\left\| \frac{1}{h^d} K_{x,h} - f_0 \right\|_{L_2(Q)} \leq \left\| \frac{1}{h^d} K_{x,h} \right\|_{L_2(Q)} + \|f_0\|_{L_2(Q)} < \eta.$$

Therefore, (40) is shown. Hence combined with Assumption 4 gives that for every probability measure Q on $\mathbb{R}^d$ and for every $\eta \in (0, h^{-d}\|K\|_\infty)$, the covering number $\mathcal{N}(\tilde{\mathscr{F}}_{K,[l_n,\infty)}, L_2(Q), \eta)$ is upper bounded as

$$\begin{aligned}
&\sup_Q \mathcal{N}(\tilde{\mathscr{F}}_{K,[l_n,\infty)}, L_2(Q), \eta) \\
&\leq \mathcal{N}\left(\left[l_n, \left(\frac{\eta}{2\|K\|_\infty}\right)^{-1/d}\right], |\cdot|, \frac{l_n^{d+1}\eta}{2d\|K\|_\infty}\right) \sup_Q \mathcal{N}\left(\mathscr{F}_{K,[l_n,\infty)}, L_2(Q), \frac{l_n^d \eta}{2}\right) + 1 \\
&\leq \frac{2d\|K\|_\infty}{l_n^{d+1}\eta} \left(\frac{2\|K\|_\infty}{\eta}\right)^{1/d} \left(\frac{2A\|K\|_\infty}{l_n^d \eta}\right)^{\nu} + 1 \\
&\leq \left(\frac{2Ad\|K\|_\infty}{l_n^d \eta}\right)^{\nu+2}. \qquad (44)
\end{aligned}$$

Also, Lemma 11 implies that under Assumption 3, for any $\varepsilon \in (0, d_{\mathrm{vol}})$ (and ε can be 0 if $d_{\mathrm{vol}} = 0$ or under Assumption 1),

$$\mathbb{E}_P\left[\left(\frac{1}{h^d} K_{x,h}\right)^2\right] \leq C_{k=2,P,K,\varepsilon}\, l_n^{-2d+d_{\mathrm{vol}}-\varepsilon}. \qquad (45)$$

Hence from (39), (44), and (45), applying Theorem 30 to (38) gives that $\sup_{h \geq l_n, x \in \mathbb{X}} |\hat{p}_h(x) - p_h(x)|$

is upper bounded with probability at least $1-\delta$ as

$$
\begin{aligned}
&\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x)-p_h(x)| \\
&\leq C\left(\frac{2(\nu+2)\|K\|_\infty \log\left(\frac{2Ad\|K\|_\infty}{\sqrt{C_{k=2,P,K,\varepsilon}}l_n^{(d_{\rm vol}-\varepsilon)/2}}\right)}{nl_n^d} + \sqrt{\frac{2(\nu+2)C_{k=2,P,K,\varepsilon}\log\left(\frac{2Ad\|K\|_\infty}{\sqrt{C_{k=2,P,K,\varepsilon}}l_n^{(d_{\rm vol}-\varepsilon)/2}}\right)}{nl_n^{2d-d_{\rm vol}+\varepsilon}}}\right. \\
&\qquad \left. + \sqrt{\frac{C_{k=2,P,K,\varepsilon}\log(\frac{1}{\delta})}{nl_n^{2d-d_{\rm vol}+\varepsilon}}} + \frac{\|K\|_\infty \log(\frac{1}{\delta})}{nl_n^d}\right) \\
&\leq C_{A,\|K\|_\infty,d,\nu,d_{\rm vol},C_{k=2,P,K,\varepsilon}}\left(\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^d} + \sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d-d_{\rm vol}+\varepsilon}}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{2d-d_{\rm vol}+\varepsilon}}} + \frac{\log\left(\frac{2}{\delta}\right)}{nl_n^d}\right),
\end{aligned}
$$

where $C_{A,\|K\|_\infty,d,\nu,d_{\rm vol},C_{k=2,P,K,\varepsilon},\varepsilon}$ depends only on A, $\|K\|_\infty$, d, ν, $d_{\rm vol}$, $C_{k=2,P,K,\varepsilon}$, ε.

□

Then Corollary 13 is just simplifying the result in Theorem 12.

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

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

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

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

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

Proof of Corollary 13. From (12) in Theorem 12, $\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x)-p_h(x)|$ is upper bounded with probability at least $1-\delta$ as

$$
\begin{aligned}
&\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x)-p_h(x)| \\
&\leq C_{A,\|K\|_\infty,d,\nu,d_{\mathrm{vol}},C_{k=2,P,K,\varepsilon},\varepsilon}\left(\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^d}+\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{2d-d_{\mathrm{vol}+\varepsilon}+\varepsilon}}}+\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^d}\right) \\
&= C_{A,\|K\|_\infty,d,\nu,d_{\mathrm{vol}},C_{k=2,P,K,\varepsilon},\varepsilon} \\
&\quad\times\left(\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}\left(\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{d_{\mathrm{vol}}-\varepsilon}}}+1\right)+\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}\left(\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{d_{\mathrm{vol}}-\varepsilon}}}+1\right)\right).
\end{aligned}
$$

Then from $\limsup_n \frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+ + \log\left(\frac{2}{\delta}\right)}{nl_n^{d_{\mathrm{vol}}-\varepsilon}}<\infty$, there exists some constant C' with $\left(\log\left(\frac{1}{l_n}\right)\right)_+ + \log\left(\frac{2}{\delta}\right)\leq C'nl_n^{d_{\mathrm{vol}}+\varepsilon}$. And hence $\sup_{h\geq l_n,x\in\mathbb{X}}|\hat{p}_h(x)-p_h(x)|$ is upper bounded with probability $1-\delta$ as

$$
\begin{aligned}
&\sup_{h\geq l_n, x\in\mathbb{X}} |\hat{p}_h(x)-p_h(x)| \\
&\leq C_{A,\|K\|_\infty,d,\nu,d_{\mathrm{vol}},C_{k=2,P,K,\varepsilon},\varepsilon}\left(\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}\left(\sqrt{C'}+1\right)+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}\left(\sqrt{C'}+1\right)\right) \\
&\leq C'_{A,\|K\|_\infty,d,\nu,d_{\mathrm{vol}},C_{k=2,P,K,\varepsilon},\varepsilon}\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+ + \log\left(\frac{1}{\delta}\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}},
\end{aligned}
$$

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

□

E.2 Proof for Section 4.2

Lemma 14 is by covering $\mathbb{X}$ and then using the Lipschitz property of the kernel function K.

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*

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

Proof of Lemma 14. For fixed $\eta > 0$, let $x_1, \dots, x_M$ be the maximal η-covering of $\mathbb{B}_{\mathbb{R}^d}(0,R)$, with $M = \mathscr{M}\left(\mathbb{B}_{\mathbb{R}^d}(0,R), \|\cdot\|_2, \eta\right)$ being the packing number of $\mathbb{B}_{\mathbb{R}^d}(0,R)$. Then $\mathbb{B}_{\mathbb{R}^d}(x_i,\eta)$ and $\mathbb{B}_{\mathbb{R}^d}(x_j,\eta)$ do not intersect for any i, j and $\bigcup_{i=1}^{M} \mathbb{B}_{\mathbb{R}^d}(x_i,\eta) \subset \mathbb{B}_{\mathbb{R}^d}(x_i, R+\eta)$, and hence

$$\sum_{i=1}^{M} \lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(x_i,\eta)\right) \leq \lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(x_i,R+\eta)\right). \tag{46}$$

Then $\lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(x,r)\right) = r^d \lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(0,1)\right)$ gives the upper bound on $\mathscr{M}\left(\mathbb{B}_{\mathbb{R}^d}(0,R), \|\cdot\|_2, \eta\right)$ as

$$\mathscr{M}\left(\mathbb{B}_{\mathbb{R}^d}(0,R), \|\cdot\|_2, \eta\right) \leq \left(1 + \frac{R}{\eta}\right)^d.$$

Then $\mathbb{X} \subset \mathbb{B}_{\mathbb{R}^d}(0,R)$ and the relationship between covering number and packing number gives the upper bound on the covering number $\mathscr{N}\left(\mathbb{X}, \|\cdot\|_2, \eta\right)$ as

$$\mathscr{N}\left(\mathbb{X}, \|\cdot\|_2, \eta\right) \leq \mathscr{N}\left(\mathbb{B}_{\mathbb{R}^d}(0,R), \|\cdot\|_2, \eta\right) \leq \mathscr{M}\left(\mathbb{B}_{\mathbb{R}^d}(0,R), \|\cdot\|_2, \frac{\eta}{2}\right) \leq \left(1 + \frac{2R}{\eta}\right)^d. \tag{47}$$

Now, note that for all $x, y \in \mathbb{X}$ and for all $z \in \mathbb{R}^d$, $\left|K_{x,h}(z) - K_{y,h}(z)\right|$ is upper bounded as

$$\left|K_{x,h}(z) - K_{y,h}(z)\right| = \left|K\left(\frac{x-z}{h}\right) - K\left(\frac{y-z}{h}\right)\right| \leq \frac{M_K}{h}\|(x-z)-(y-z)\|_2 = \frac{M_K}{h}\|x-y\|_2.$$

Hence for any measure Q on $\mathbb{R}^d$, $\left\|K_{x,h} - K_{y,h}\right\|_{L_2(Q)}$ is upper bounded as

$$\left\|K_{x,h} - K_{y,h}\right\|_{L_2(Q)} = \sqrt{\int (K_{x,h}(z) - K_{y,h}(z))^2 dQ(z)} \leq \frac{M_K}{h}\|x-y\|_2.$$

Hence applying this to (47) implies that for all $\eta > 0$, the supremum of the covering number $\mathscr{N}(\mathscr{F}_{K,h}, L_2(Q), \eta)$ over all measure Q is upper bounded as

$$\sup_Q \mathscr{N}(\mathscr{F}_{K,h}, L_2(Q), \eta) \leq \mathscr{N}\left(\mathbb{X}, \|\cdot\|_2, \frac{h\eta}{M_K}\right) \leq \left(1 + \frac{2RM_K}{h\eta}\right)^d.$$

Hence for all $\eta \in (0, \|K\|_\infty)$,

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

□

Then Corollary 15 follows from applying Theorem 30 with bounding the covering number from Lemma 14.

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 $\varepsilon \in (0, d_{\mathrm{vol}})$. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, ε can be 0. Suppose*

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

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

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

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

Proof of Corollary 15. For $x \in \mathbb{X}$, let $K_{x,h} : \mathbb{R}^d \to \mathbb{R}$ be $K_{x,h}(\cdot) = K\left(\frac{x-\cdot}{h}\right)$, and let $\tilde{\mathscr{F}}_{K,h} := \left\{\frac{1}{h^d}K_{x,h} : x \in \mathbb{X}\right\}$ be a class of normalized kernel functions centered on $\mathbb{X}$ and bandwidth h. Note that $\hat{p}_h(x) - p_h(x)$ can be expanded as

$$\hat{p}_h(x) - p_h(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) - \mathbb{E}_P\left[\frac{1}{h^d}K\left(\frac{x-X_i}{h}\right)\right] = \frac{1}{n}\sum_{i=1}^n \frac{1}{h^d}K_{x,h}(X_i) - \mathbb{E}_P\left[\frac{1}{h^d}K_{x,h}\right].$$

Hence $\sup_{x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)|$ can be expanded as

$$\sup_{x\in\mathbb{X}} |\hat{p}_h(x) - p_h(x)| = \sup_{f\in\tilde{\mathscr{F}}_{K,h}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_P[f(X)]\right|. \tag{48}$$

Now, it is immediate to check that

$$\|f\|_\infty \le h^{-d}\|K\|_\infty. \tag{49}$$

Also, Since $\tilde{\mathscr{F}}_{K,h} = h^{-d}\mathscr{F}_{K,h}$, VC dimension is uniformly bounded as Lemma 14 gives that for every probability measure Q on $\mathbb{R}^d$ and for every $\eta \in (0, h^{-d}\|K\|_\infty)$, the covering number $\mathcal{N}(\tilde{\mathscr{F}}_{K,h}, L_2(Q), \eta)$ is upper bounded as

$$\begin{aligned}
\sup_Q \mathcal{N}(\tilde{\mathscr{F}}_{K,h}, L_2(Q), \eta) &= \sup_Q \mathcal{N}(\mathscr{F}_{K,h}, L_2(Q), h^d\eta) \\
&\le \left(\frac{2RM_K h^{-1} + \|K\|_\infty}{h^d\eta}\right)^d \\
&\le \left(\frac{2RM_K\|K\|_\infty}{h^{d+1}\eta}\right)^d.
\end{aligned} \tag{50}$$

Also, Lemma 11 implies that under Assumption 3, for any $\varepsilon \in (0, d_{\mathrm{vol}})$ (and ε can be 0 if $d_{\mathrm{vol}} = 0$ or under Assumption 1),

$$\mathbb{E}_P\left[\left(\frac{1}{h^d}K_{x,h}\right)^2\right] \le C_{k=2,P,K,\varepsilon}h^{-2d+d_{\mathrm{vol}}-\varepsilon}. \tag{51}$$

Hence from (49), (50), and (51), applying Theorem 30 to (48) gives that $\sup_{x\in\mathbb{X}}|\hat{p}_h(x)-p_h(x)|$ is upper bounded with probability at least $1-\delta$ as

$$\begin{aligned}
&\sup_{x\in\mathbb{X}}|\hat{p}_h(x)-p_h(x)| \\
&\leq C\left(\frac{2d\,\|K\|_\infty \log\left(\frac{2RM_K\|K\|_\infty}{\sqrt{C_{k=2,P,K,\varepsilon}}h^{1+(d_{\mathrm{vol}}-\varepsilon)/2}}\right)}{nh^d}+\sqrt{\frac{2dC_{k=2,P,K,\varepsilon}\log\left(\frac{2RM_K\|K\|_\infty}{\sqrt{C_{k=2,P,K,\varepsilon}}h^{1+(d_{\mathrm{vol}}-\varepsilon)/2}}\right)}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}\right. \\
&\qquad\left.+\sqrt{\frac{C_{k=2,P,K,\varepsilon}\log(\frac{1}{\delta})}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\|K\|_\infty\log(\frac{1}{\delta})}{nh^d}\right) \\
&\leq C_{R,M_K,\|K\|_\infty,d,\nu,d_{\mathrm{vol}},C_{k=2,P,K,\varepsilon},\varepsilon}\left(\frac{\left(\log\left(\frac{1}{h}\right)\right)_+}{nh^d}+\sqrt{\frac{\left(\log\left(\frac{1}{h}\right)\right)_+}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\log\left(\frac{2}{\delta}\right)}{nh^d}\right),
\end{aligned}$$

where $C_{R,M_K,\|K\|_\infty,d,\nu,d_{\mathrm{vol}},C_{k=2,P,K,\varepsilon},\varepsilon}$ depends only on R, M_K, $\|K\|_\infty$, d, ν, d_{vol}, $C_{k=2,P,K,\varepsilon}$, ε. □

F Proof for Section 5

Proposition 16 is shown by finding $x_0\in\mathbb{X}$ where the volume dimension is obtained, and analyzing the behavior of $|\hat{p}_{h_n}(x_0)-p_{h_n}(x_0)|$ by applying Central Limit Theorem.

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

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

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

Proof of Proposition 16. Note that $\lim_{t\to 0}\inf_{\|x\|\leq t}K(x)>0$ implies that there exists $t_0,K_0\in(0,\infty)$ such that

$$K(x)\geq K_0 I(\|x\|\leq t_0). \tag{52}$$

Also, from $\sup_{x\in\mathbb{X}}\liminf_{r\to 0}\frac{P(\mathbb{B}_{\mathbb{R}^d}(x,r))}{r^{d_{\mathrm{vol}}}}>0$, we can choose $x_0\in\mathbb{X}$ such that $\liminf_{r\to 0}\frac{P(\mathbb{B}_{\mathbb{R}^d}(x_0,r))}{r^{d_{\mathrm{vol}}}}>0$. From $\{h_n\}_{n\in\mathbb{N}}$ bounded, there exists $r_0>0$ and $p_0>0$ such that $r_0\geq h_n t_0$ for all $n\in\mathbb{N}$ and for all $r\leq r_0$,

$$P(\mathbb{B}_{\mathbb{R}^d}(x_0,r))\geq p_0 r^{d_{\mathrm{vol}}}. \tag{53}$$

For $x \in \mathbb{X}$ and $h > 0$, let $f_{x,h} : \mathbb{R}^d \to \mathbb{R}$ be $f_{x,h} = \frac{1}{h^d}\left(K_{x,h} - \mathbb{E}_P[K_{x,h}]\right)$, so that at $x_0 \in \mathbb{X}$, $\hat{p}_{h_n}(x_0) - p_{h_n}(x_0)$ is expanded as

$$\hat{p}_{h_n}(x_0) - p_{h_n}(x_0) = \frac{1}{n}\sum_{i=1}^{n} f_{x_0,h_n}(X_i).$$

Below we get a lower bound for $\mathbb{E}_P[f_{x_0,h_n}^2]$. First, fix $\varepsilon < \frac{d_{\text{vol}}}{2}$. Then from Lemma 11,

$$\mathbb{E}_P\left[\left|K_{x_0,h}\right|\right] \leq C_{k=1,P,K,\varepsilon} h^{d_{\text{vol}}-\varepsilon}. \tag{54}$$

Now, we lower bound $\mathbb{E}_P[K_{x_0,h}^2]$. By applying (52), $\mathbb{E}_P[K_{x_0,h}^2]$ is lower bounded as

$$\begin{aligned}\mathbb{E}_P\left[K_{x_0,h}^2\right] &\geq \mathbb{E}_P\left[K_0 I\left(\left\|\frac{x_0 - X_i}{h}\right\| \geq t_0\right)\right] \\ &= K_0^2 P(\mathbb{B}_{\mathbb{R}^d}(x_0, ht_0)).\end{aligned}$$

Then applying (53) gives a further lower bound as

$$\mathbb{E}_P\left[K_{x_0,h}^2\right] \geq K_0^2 p_0 t_0^{d_{\text{vol}}} h^{d_{\text{vol}}}. \tag{55}$$

Then combining (54) and (55) gives a lower bound of $\mathbb{E}_P[f_{x_0,h}^2]$ as

$$\begin{aligned}\mathbb{E}_P\left[f_{x_0,h}^2\right] &= \frac{1}{h^{2d}}\left(\mathbb{E}_P\left[K_{x_0,h}^2\right] - \left(\mathbb{E}_P\left[K_{x_0,h}\right]\right)^2\right) \\ &\geq h^{d_{\text{vol}}-2d}(K_0^2 p_0 t_0^{d_{\text{vol}}} - C_{k=1,P,K,\varepsilon}^2 h^{d_{\text{vol}}-2\varepsilon}).\end{aligned}$$

Hence from $d_{\text{vol}} - 2\varepsilon > 0$, there exists $h_{P,K}$ and $C'_{P,K}$ depending only on P and K such that $h_n \leq h_{P,K}$ implies

$$\mathbb{E}_P\left[f_{x_0,h_n}^2\right] \geq C'_{P,K} h_n^{d_{\text{vol}}-2d}. \tag{56}$$

Now, let $s_n := \sqrt{\sum_{i=1}^n \mathbb{E}_P[f_{x_0,h_n}^2(X_i)]}$. Then (56) gives

$$s_n \geq \sqrt{C'_{P,K} n h_n^{d_{\text{vol}}-2d}}.$$

Then for any $\varepsilon > 0$, when n is large enough so that $nh_n^{d_{\text{vol}}} > \frac{\|K\|_\infty^2}{\varepsilon^2 C'_{P,K}}$, then

$$\left\|f_{x_0,h_n}\right\|_\infty \leq h^{-d}\left\|K\right\|_\infty < \varepsilon\sqrt{C'_{P,K} n h_n^{d_{\text{vol}}-2d}} \leq s_n.$$

Hence Lindeberg condition holds as for n large enough so that $nh_n^{d_{\text{vol}}} > \frac{\|K\|_\infty^2}{\varepsilon^2 C'_{P,K}}$, then

$$\frac{1}{s_n^2}\sum_{i=1}^{n} \mathbb{E}\left[f_{x_0,h_n}^2(X_i) I\left(\left|f_{x_0,h_n}(X_i)\right| \geq \varepsilon s_n\right)\right] = 0.$$

Hence, Lindeberg-Feller Central Limit Theorem gives

$$\sqrt{\frac{n}{\mathbb{E}_P[f_{x_0,h_n}^2]}}\left(\hat{p}_{h_n}(x_0) - p_{h_n}(x_0)\right) \xrightarrow{d} N(0,1).$$

Hence, for fixed $\delta \in (0,1)$, let $q_{\delta/2} \in \mathbb{R}$ be such that $P(|Z| \leq q_{\delta/2}) = \frac{\delta}{2}$ for $Z \sim N(0,1)$, then

$$\lim_{n\to\infty} P\left(\left|\sqrt{\frac{n}{\mathbb{E}_P[f_{x_0,h_n}^2]}}\left(\hat{p}_{h_n}(x_0) - p_{h_n}(x_0)\right)\right| \geq q_{\delta/2}\right) = 1 - \frac{\delta}{2}.$$

And hence there exists $N < \infty$ that for all $n \geq N$,

$$P\left(|\hat{p}_{h_n}(x_0) - p_{h_n}(x_0)| \geq q_{\delta/2}\sqrt{\frac{\mathbb{E}_P[f_{x_0,h_n}^2]}{n}}\right) \geq 1 - \delta.$$

Then applying (56) implies that with probability at least $1-\delta$,

$$|\hat{p}_{h_n}(x_0) - p_{h_n}(x_0)| \geq \sqrt{\frac{q_{\delta/2}^2 C'_{P,K}}{n h_n^{2d-d_{\mathrm{vol}}}}} = C_{P,K,\delta}\sqrt{\frac{1}{n h_n^{2d-d_{\mathrm{vol}}}}},$$

where $C_{P,K,\delta} = q_{\delta/2}\sqrt{C'_{P,K}}$ depends only on P, K, and δ. Then from

$$\sup_{x\in\mathbb{X}} |\hat{p}_{h_n}(x) - p_{h_n}(x)| \geq |\hat{p}_{h_n}(x_0) - p_{h_n}(x_0)|,$$

we get the same lower bound for $\sup_{x\in\mathbb{X}} |\hat{p}_{h_n}(x) - p_{h_n}(x)|$ with probability at least $1-\delta$ as

$$\sup_{x\in\mathbb{X}} |\hat{p}_{h_n}(x) - p_{h_n}(x)| \geq \sqrt{\frac{q_{\delta/2}^2 C'_{P,K}}{n h_n^{2d-d_{\mathrm{vol}}}}} = C_{P,K,\delta}\sqrt{\frac{1}{n h_n^{2d-d_{\mathrm{vol}}}}}.$$

□

G Proof for Section 6

For showing Lemma 19, we proceed similarly to proof of Lemma 11, where we plug in $D^s K$ in the place of K.

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*

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

for any $\varepsilon \in (0, d_{\mathrm{vol}})$, where $C_{s,P,K,\varepsilon}$ 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).

Proof of Lemma 19. We first consider the case when $d_{\mathrm{vol}} = 0$. Then $\mathbb{E}_P\left[\left(D^s K\left(\frac{x-X}{h}\right)\right)^2\right]$ is simply bounded as

$$\mathbb{E}_P\left[\left(D^s K\left(\frac{x-X}{h}\right)\right)^2\right] \leq \|D^s K\|_\infty^2 h^0.$$

Now, we consider the case when $d_{\mathrm{vol}} > 0$. Fix $\varepsilon \in (0, d_{\mathrm{vol}})$. Under Assumption 1, ε can be chosen to be 0.

Let $C_{s,K,d_{\mathrm{vol}},\varepsilon} := \int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t}(D^s K(x))^2 dt$, then it is finite from (17) and $\|D^s K\|_\infty < \infty$ in Assumption 7 as

$$\begin{aligned}\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t}(D^s K(x))^2 dt &\leq \int_0^1 t^{d_{\mathrm{vol}}-\varepsilon-1} \|D^s K\|_\infty dt + \int_1^\infty t^{d_{\mathrm{vol}}-1} \sup_{\|x\|\leq t}(D^s K(x))^2 dt \\ &\leq \frac{\|D^s K\|_\infty}{d_{\mathrm{vol}}-\varepsilon} + \int_0^\infty t^{d_{\mathrm{vol}}-1} \sup_{\|x\|\leq t}(D^s K(x))^2 dt < \infty.\end{aligned}$$

Fix $\eta > 0$, and let $\tilde{K}_\eta : [0,\infty) \to \mathbb{R}$ be a continuous and strictly decreasing function satisfying $\tilde{K}_\eta(t) > \sup_{\|x\|\geq t}(D^s K(x))^2$ for all $t \geq 0$ and $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}(\tilde{K}_\eta(t) - \sup_{\|x\|\geq t}(D^s K(x))^2)dt = \eta$. Such existence is possible since $t \mapsto \sup_{\|x\|\geq t}(D^s K(x))^2$ is nonincreasing function, so have at most countable discontinuous points, and $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t}(D^s K(x))^2 dt < \infty$. Then it is immediate to check that

$$(D^s K(x))^2 < \tilde{K}_\eta(\|x\|) \text{ for all } x \in \mathbb{R}. \tag{57}$$

Then $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(t)dt$ can be expanded as

$$\begin{aligned}\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}_\eta(t)dt &= \int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1} \sup_{\|x\|\leq t}(D^s K(x))^2 dt + \int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}(\tilde{K}_\eta(t) - \sup_{\|x\|\geq t}(D^s K(x))^2)dt \\ &= C_{s,K,d_{\mathrm{vol}},\varepsilon} + \eta < \infty. \end{aligned} \tag{58}$$

Now since $\tilde{K}_\eta$ is continuous and strictly decreasing, change of variables $t = \tilde{K}_\eta(u)$ is applicable, and then $\mathbb{E}_P\left[\left(D^s K\left(\frac{x-X}{h}\right)\right)^2\right]$ can be expanded as

$$\begin{aligned}\mathbb{E}_P\left[\left(D^s K\left(\frac{x-X}{h}\right)\right)^2\right] &= \int_0^\infty P\left(\left(D^s K\left(\frac{x-X}{h}\right)\right)^2 > t\right) dt \\ &= \int_\infty^0 P\left(\left(D^s K\left(\frac{x-X}{h}\right)\right)^2 > \tilde{K}_\eta(u)\right) d\tilde{K}_\eta(u).\end{aligned}$$

Now, from (57) and $\tilde{K}_\eta$ being a strictly decreasing, we can upper bound $\mathbb{E}_P\left[\left(D^sK\left(\frac{x-X}{h}\right)\right)^2\right]$ as

$$\begin{aligned}\mathbb{E}_P\left[\left(D^sK\left(\frac{x-X}{h}\right)\right)^2\right] &\leq \int_\infty^0 P\left(\tilde{K}_\eta\left(\frac{\|x-X\|}{h}\right) > \tilde{K}_\eta(u)\right) d\tilde{K}_\eta(u)\\ &= \int_\infty^0 P\left(\frac{\|x-X\|}{h} < u\right) d\tilde{K}_\eta(u)\\ &= \int_\infty^0 P\left(\mathbb{B}_{\mathbb{R}^d}(x,hu)\right) d\tilde{K}_\eta(u).\end{aligned}$$

Now, from Lemma 4 (and (6) for Assumption 1 case), there exists $C_{d_{\mathrm{vol}}-\varepsilon,P}<\infty$ with $P\left(\mathbb{B}_{\mathbb{R}^d}(x,r)\right)\leq C_{d_{\mathrm{vol}}-\varepsilon,P}r^{d_{\mathrm{vol}}-\varepsilon}$ for all $x\in\mathbb{X}$ and $r>0$. Then $\mathbb{E}_P\left[\left(D^sK\left(\frac{x-X}{h}\right)\right)^2\right]$ is further upper bounded as

$$\begin{aligned}\mathbb{E}_P\left[\left(D^sK\left(\frac{x-X}{h}\right)\right)^2\right] &\leq \int_\infty^0 C_{d_{\mathrm{vol}}-\varepsilon,P}(hu)^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u)\\ &= C_{d_{\mathrm{vol}}-\varepsilon,P}h^{d_{\mathrm{vol}}-\varepsilon}\int_\infty^0 u^{d_{\mathrm{vol}}-\varepsilon} d\tilde{K}(u). \qquad (59)\end{aligned}$$

Now, $\int_\infty^0 u^{d_{\mathrm{vol}}-\varepsilon}d\tilde{K}(u)$ can be computed using integration by part. Note first that $\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(t)dt<\infty$ implies

$$\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)=0.$$

To see this, note that $t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)$ is expanded as

$$t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)=\int_0^t u^{d_{\mathrm{vol}}-\varepsilon}d\tilde{K}(u)+\int_0^t (d_{\mathrm{vol}}-\varepsilon)u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u)du,$$

then $\int_0^\infty (d_{\mathrm{vol}}-\varepsilon)u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u)du<\infty$ and $\int_0^t u^{d_{\mathrm{vol}}-\varepsilon}d\tilde{K}(u)$ being monotone function of t imply that $\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)$ exists. Now, suppose $\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)=a>0$, then we can choose $t_0>0$ such that $t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)>\frac{a}{2}$ for all $t\geq t_0$, and then

$$\infty>\int_0^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(t)dt\geq\int_{t_0}^\infty t^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(t)dt\geq\frac{a}{2}\int_{t_0}^\infty t^{-1}dt=\infty,$$

which is a contradiction. Hence $\lim_{t\to\infty} t^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(t)=0$. Now, applying integration by part to $\int_\infty^0 u^{d_{\mathrm{vol}}-\varepsilon}d\tilde{K}(u)$ with $d_{\mathrm{vol}}-\varepsilon>0$ gives

$$\begin{aligned}\int_\infty^0 u^{d_{\mathrm{vol}}-\varepsilon}d\tilde{K}(u) &= \left[u^{d_{\mathrm{vol}}-\varepsilon}\tilde{K}(u)\right]_\infty^0-\int_\infty^0 (d_{\mathrm{vol}}-\varepsilon)u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u)du\\ &= \int_0^\infty (d_{\mathrm{vol}}-\varepsilon)u^{d_{\mathrm{vol}}-\varepsilon-1}\tilde{K}(u)du. \qquad (60)\end{aligned}$$

Then applying (58) and (60) to (59) gives an upper bound for $\mathbb{E}_P\left[\left(D^sK\left(\frac{x-X}{h}\right)\right)^2\right]$ as

$$\mathbb{E}_P\left[\left(D^sK\left(\frac{x-X}{h}\right)\right)^2\right]\leq C_{d_{\mathrm{vol}}-\varepsilon,P}(d_{\mathrm{vol}}-\varepsilon)h^{d_{\mathrm{vol}}-\varepsilon}(C_{s,K,d_{\mathrm{vol}},\varepsilon}+\eta). \qquad (61)$$

And then note that RHS of (61) holds for any $\eta > 0$, and hence $\mathbb{E}_P\left[\left(D^s K\left(\frac{x-X}{h}\right)\right)^2\right]$ is further upper bounded as

$$\begin{aligned}\mathbb{E}_P\left[\left(D^s K\left(\frac{x-X}{h}\right)\right)^2\right] &\leq \inf_{\eta>0}\left\{C_{d_{\rm vol}-\varepsilon,P}(d_{\rm vol}-\varepsilon)h^{d_{\rm vol}-\varepsilon}(C_{s,K,d_{\rm vol},\varepsilon}+\eta)\right\} \\ &= C_{d_{\rm vol}-\varepsilon,P}(d_{\rm vol}-\varepsilon)C_{s,K,d_{\rm vol},\varepsilon}h^{d_{\rm vol}-\varepsilon} \\ &= C_{s,P,K,\varepsilon}h^{d_{\rm vol}-\varepsilon},\end{aligned}$$

where $C_{k,P,K,\varepsilon} = C_{d_{\rm vol}-\varepsilon,P}(d_{\rm vol}-\varepsilon)C_{s,K,d_{\rm vol},\varepsilon}$.

□

For proving Theorem 20, we proceed similarly to the proof of Theorem 12. Analogous to bounding $\mathbb{E}_P[K_{x,h}^2]$ by Lemma 11, we bound $\mathbb{E}_P[(D^s K_{x,h})^2]$ by Lemma 19.

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

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

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

Proof of Theorem 20. For $x \in \mathbb{X}$ and $h \geq l_n$, let $D^s K_{x,h} : \mathbb{R}^d \to \mathbb{R}$ be $D^s K_{x,h}(\cdot) = D^s K\left(\frac{x-\cdot}{h}\right)$, and let $\tilde{\mathcal{F}}_{K,[l_n,\infty)}^s := \left\{\frac{1}{h^{d+|s|}} D^s K_{x,h} : x \in \mathbb{X}, h \geq l_n\right\}$ be a class of normalized kernel functions centered on $\mathbb{X}$ and bandwidth in $[l_n, \infty)$. Note that $D^s\hat{p}_h(x) - D^s p_h(x)$ can be expanded as

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

Hence $\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)|$ can be expanded as

$$\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)| = \sup_{f\in\tilde{\mathcal{F}}_{K,[l_n,\infty)}^s}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_P[f(X)]\right|. \tag{62}$$

Now, it is immediate to check that

$$\|f\|_\infty \leq l_n^{-d-|s|}\|D^s K\|_\infty. \tag{63}$$

For bounding the VC dimension of $\tilde{\mathscr{F}}^s_{K,[l_n,\infty)}$, consider $\mathscr{F}^s_{K,[l_n,\infty)} := \left\{D^s K_{x,h} : x \in \mathbb{X}, h \geq l_n\right\}$ be a class of unnormalized kernel functions centered on $\mathbb{X}$ and bandwidth in $[l_n,\infty)$. Fix $\eta < l_n^{-d-|s|}\left\|D^s K\right\|_\infty$ and a probability measure Q on $\mathbb{R}^d$. Suppose $\left[l_n, \left(\frac{\eta}{2\|D^s K\|_\infty}\right)^{-1/(d+|s|)}\right]$ is covered by balls $\left\{\left(h_i - \frac{l_n^{d+|s|+1}\eta}{2(d+|s|)\|D^s K\|_\infty}, h_i + \frac{l_n^{d+|s|+1}\eta}{2(d+|s|)\|D^s K\|_\infty}\right) : 1 \leq i \leq N_1\right\}$ and $(\mathscr{F}^s_{K,[l_n,\infty)}, L_2(Q))$ is covered by balls $\left\{\mathbb{B}_{L_2(Q)}\left(f_j, \frac{l_n^{d+|s|}\eta}{2}\right) : 1 \leq j \leq N_2\right\}$, and let $f_{i,j} := h_i^{-d-|s|} f_j$ for $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$. Also, choose $h_0 > \left(\frac{\eta}{2\|D^s K\|_\infty}\right)^{-1/(d+|s|)}$, $x_0 \in \mathbb{X}$, and let $f_0 = \frac{1}{h_0^{d+|s|}} D^s K_{x_0,h_0}$. We will show that

$$\left\{\mathbb{B}_{L_2(Q)}\left(f_{i,j},\eta\right) : 1 \leq i \leq N_1, 1 \leq j \leq N_2\right\} \cup \left\{\mathbb{B}_{L_2(Q)}\left(f_0,\eta\right)\right\} \text{ covers } \tilde{\mathscr{F}}^s_{K,[l_n,\infty)}. \tag{64}$$

For the first case when $h \leq \left(\frac{\eta}{2\|D^s K\|_\infty}\right)^{-1/(d+|s|)}$, find h_i and f_j with $h \in \left(h_i - \frac{l_n^{d+|s|+1}\eta}{2(d+|s|)\|D^s K\|_\infty}, h_i + \frac{l_n^{d+|s|+1}\eta}{2(d+|s|)\|D^s K\|_\infty}\right)$ and $K_{x,h} \in \mathbb{B}_{L_2(Q)}\left(f_j, \frac{l_n^{d+|s|}\eta}{2}\right)$. Then the distance between $\frac{1}{h^{d+|s|}} D^s K_{x,h}$ and $\frac{1}{h_i^{d+|s|}} f_j$ is upper bounded as

$$\begin{aligned}
&\left\|\frac{1}{h^{d+|s|}} D^s K_{x,h} - \frac{1}{h_i^{d+|s|}} f_j\right\|_{L_2(Q)} \\
&\leq \left\|\frac{1}{h^{d+|s|}} D^s K_{x,h} - \frac{1}{h_i^{d+|s|}} D^s K_{x,h}\right\|_{L_2(Q)} + \left\|\frac{1}{h_i^{d+|s|}} D^s K_{x,h} - \frac{1}{h_i^{d+|s|}} f_j\right\|_{L_2(Q)}.
\end{aligned} \tag{65}$$

Now, the first term of (65) is upper bounded as

$$\begin{aligned}
\left\|\frac{1}{h^{d+|s|}} D^s K_{x,h} - \frac{1}{h_i^{d+|s|}} D^s K_{x,h}\right\|_{L_2(Q)} &= \left|\frac{1}{h^{d+|s|}} - \frac{1}{h_i^{d+|s|}}\right| \left\|D^s K_{x,h}\right\|_{L_2(Q)} \\
&= |h_i - h| \sum_{k=0}^{d+|s|-1} h_i^{k-d-|s|} h^{-1-k} \left\|D^s K_{x,h}\right\|_{L_2(Q)} \\
&\leq |h_i - h|\,(d+|s|) l_n^{-d-|s|-1} \left\|D^s K\right\|_\infty < \frac{\eta}{2}.
\end{aligned} \tag{66}$$

Also, the second term of (65) is upper bounded as

$$\begin{aligned}
\left\|\frac{1}{h_i^{d+|s|}} D^s K_{x,h} - \frac{1}{h_i^{d+|s|}} f\right\|_{L_2(Q)} &= \frac{1}{h_i^{d+|s|}} \left\|D^s K_{x,h} - f\right\|_{L_2(Q)} \\
&\leq l_n^{-d-|s|} \left\|D^s K_{x,h} - f\right\|_{L_2(Q)} < \frac{\eta}{2}.
\end{aligned} \tag{67}$$

Hence applying (66) and (67) to (65) gives

$$\left\|\frac{1}{h^{d+|s|}} D^s K_{x,h} - \frac{1}{h_i^{d+|s|}} f_j\right\|_{L_2(Q)} < \eta.$$

For the second case when $h > \left(\frac{\eta}{2\|D^s K\|_\infty}\right)^{-1/(d+|s|)}$, $\left\|\frac{1}{h^{d+|s|}} D^s K_{x,h}\right\|_{L_2(Q)} \leq \left\|\frac{1}{h^{d+|s|}} D^s K_{x,h}\right\|_\infty < \frac{\eta}{2}$ holds, and hence

$$\left\|\frac{1}{h^{d+|s|}} D^s K_{x,h} - f_0\right\|_{L_2(Q)} \leq \left\|\frac{1}{h^{d+|s|}} D^s K_{x,h}\right\|_{L_2(Q)} + \|f_0\|_{L_2(Q)} < \eta.$$

Therefore, (64) is shown. Hence combined with Assumption 7 gives that for every probability measure Q on $\mathbb{R}^d$ and for every $\eta \in (0, h^{-d}\|D^s K\|_\infty)$, the covering number $\mathcal{N}(\tilde{\mathcal{F}}_{K,[l_n,\infty)}, L_2(Q), \eta)$ is upper bounded as

$$\begin{aligned}
&\sup_Q \mathcal{N}(\tilde{\mathcal{F}}_{K,[l_n,\infty)}, L_2(Q), \eta) \\
&\leq \mathcal{N}\left(\left[l_n, \left(\frac{\eta}{2\|D^s K\|_\infty}\right)^{-1/(d+|s|)}\right], |\cdot|, \frac{l_n^{d+|s|+1}\eta}{2(d+|s|)\|D^s K\|_\infty}\right) \sup_Q \mathcal{N}\left(\mathcal{F}_{K,[l_n,\infty)}, L_2(Q), \frac{l_n^{d+|s|}\eta}{2}\right) + 1 \\
&\leq \frac{2(d+|s|)\|D^s K\|_\infty}{l_n^{d+|s|+1}\eta}\left(\frac{2\|D^s K\|_\infty}{\eta}\right)^{1/(d+|s|)}\left(\frac{2A\|D^s K\|_\infty}{l_n^{d+|s|}\eta}\right)^{\nu} + 1 \\
&\leq \left(\frac{2A(d+|s|)\|D^s K\|_\infty}{l_n^{d+|s|}\eta}\right)^{\nu+2}.
\end{aligned} \tag{68}$$

Also, Lemma 19 implies that under Assumption 6, for any $\varepsilon \in (0, d_{\text{vol}})$ (and ε can be 0 if $d_{\text{vol}} = 0$ or under Assumption 1),

$$\mathbb{E}_P\left[\left(\frac{1}{h^{d+|s|}} D^s K_{x,h}\right)^2\right] \leq C_{s,P,K,\varepsilon} l_n^{-2d-2|s|+d_{\text{vol}}-\varepsilon}. \tag{69}$$

Hence from (63), (68), and (69), applying Theorem 30 to (62) gives that $\sup_{h\geq l_n, x\in\mathbb{X}}\left|D^s\hat{p}_h(x) - D^s p_h(x)\right|$ is upper bounded with probability at least $1-\delta$ as

$$\begin{aligned}
&\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)| \\
&\leq C\left(\frac{2(\nu+2)\|D^s K\|_\infty \log\left(\frac{2A(d+|s|)\|D^s K\|_\infty}{\sqrt{C_{s,P,K,\varepsilon}} l_n^{(d_{\text{vol}}-\varepsilon)/2}}\right)}{n l_n^{d+|s|}} + \sqrt{\frac{2(\nu+2) C_{s,P,K,\varepsilon} \log\left(\frac{2A(d+|s|)\|D^s K\|_\infty}{\sqrt{C_{s,P,K,\varepsilon}} l_n^{(d_{\text{vol}}-\varepsilon)/2}}\right)}{n l_n^{2d+2|s|-d_{\text{vol}}+\varepsilon}}}\right. \\
&\qquad \left. + \sqrt{\frac{C_{s,P,K,\varepsilon}\log(\frac{1}{\delta})}{n l_n^{2d+2|s|-d_{\text{vol}}+\varepsilon}}} + \frac{\|D_s K\|_\infty \log(\frac{1}{\delta})}{n l_n^{d+|s|}}\right) \\
&\leq C_{A,\|D^s K\|_\infty, d, \nu, d_{\text{vol}}, C_{s,P,K,\varepsilon}}\left(\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{n l_n^{d+|s|}} + \sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{n l_n^{2d+2|s|-d_{\text{vol}}+\varepsilon}}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n l_n^{2d+2|s|-d_{\text{vol}}+\varepsilon}}} + \frac{\log\left(\frac{2}{\delta}\right)}{n l_n^{d+|s|}}\right),
\end{aligned}$$

where $C_{A,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon},\varepsilon}$ depends only on A, $\|D^sK\|_\infty$, d, ν, d_{vol}, $C_{s,P,K,\varepsilon}$, ε.

$\square$

For showing Corollary 21, we proceed similarly to the proof of Corollary 13, where we plug in D^sK in the place of K.

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

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

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

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

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

Proof of Corollary 21. From (19) in Theorem 20, $\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)|$ is upper bounded with probability at least $1-\delta$ as

$$\begin{aligned}
&\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)| \\
&\leq C_{A,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon}} \left(\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{d+|s|}} + \sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}} + \frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{d+|s|}} \right) \\
&= C_{A,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon}} \\
&\quad \times \left(\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}} \left(\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{d_{\mathrm{vol}}-\varepsilon}}} + 1 \right) + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}} \left(\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{d_{\mathrm{vol}}-\varepsilon}}} + 1 \right) \right).
\end{aligned}$$

Then from $\limsup_n \frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+ + \log\left(\frac{2}{\delta}\right)}{nl_n^{d_{\mathrm{vol}}-\varepsilon}} < \infty$, there exists some constant C' with $\left(\log\left(\frac{1}{l_n}\right)\right)_+ + \log\left(\frac{2}{\delta}\right) \leq C' n l_n^{d_{\mathrm{vol}}+\varepsilon}$. And hence $\sup_{h\geq l_n, x\in\mathbb{X}} |D^s\hat{p}_h(x) - D^s p_h(x)|$ is upper bounded with probability $1-\delta$ as

$$
\begin{aligned}
&\sup_{h\geq l_n, x\in\mathbb{X}} \left|D^s\hat{p}_h(x)-D^s p_h(x)\right| \\
&\leq C_{A,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon}} \left(\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}}\left(\sqrt{C'}+1\right)+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nl_n^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}}\left(\sqrt{C'}+1\right)\right) \\
&\leq C'_{A,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon}} \sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+ + \log\left(\frac{1}{\delta}\right)}{nl_n^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}},
\end{aligned}
$$

where $C'_{A,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon}}$ depending only on A, $\|D^sK\|_\infty$, d, ν, d_{vol}, $C_{s,P,K,\varepsilon}$, ε. □

For proving Lemma 22, we proceed similarly to the proof of Lemma 14, where we plug in D^sK in the place of K.

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

$$\left\|D^sK(x)-D^sK(y)\right\|_2 \leq M_K\left\|x-y\right\|_2.$$

Then for all $\eta\in(0,\|D^sK\|_\infty)$, 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

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

Proof of Lemma 22. For fixed $\eta>0$, let $x_1,\ldots,x_M$ be the maximal η-covering of $\mathbb{B}_{\mathbb{R}^d}(0,R)$, with $M=\mathcal{M}(\mathbb{B}_{\mathbb{R}^d}(0,R),\|\cdot\|_2,\eta)$ being the packing number of $\mathbb{B}_{\mathbb{R}^d}(0,R)$. Then $\mathbb{B}_{\mathbb{R}^d}(x_i,\eta)$ and $\mathbb{B}_{\mathbb{R}^d}(x_j,\eta)$ do not intersect for any i,j and $\bigcup_{i=1}^M \mathbb{B}_{\mathbb{R}^d}(x_i,\eta)\subset\mathbb{B}_{\mathbb{R}^d}(x_i,R+\eta)$, and hence

$$\sum_{i=1}^M \lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(x_i,\eta)\right) \leq \lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(x_i,R+\eta)\right). \tag{70}$$

Then $\lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(x,r)\right)=r^d\lambda_d\left(\mathbb{B}_{\mathbb{R}^d}(0,1)\right)$ gives the upper bound on $\mathcal{M}(\mathbb{B}_{\mathbb{R}^d}(0,R),\|\cdot\|_2,\eta)$ as

$$\mathcal{M}\left(\mathbb{B}_{\mathbb{R}^d}(0,R),\|\cdot\|_2,\eta\right) \leq \left(1+\frac{R}{\eta}\right)^d.$$

Then $\mathbb{X}\subset\mathbb{B}_{\mathbb{R}^d}(0,R)$ and the relationship between covering number and packing number gives the upper bound on the covering number $\mathcal{N}(\mathbb{X},\|\cdot\|_2,\eta)$ as

$$\mathcal{N}\left(\mathbb{X},\|\cdot\|_2,\eta\right) \leq \mathcal{N}\left(\mathbb{B}_{\mathbb{R}^d}(0,R),\|\cdot\|_2,\eta\right) \leq \mathcal{M}\left(\mathbb{B}_{\mathbb{R}^d}(0,R),\|\cdot\|_2,\frac{\eta}{2}\right) \leq \left(1+\frac{2R}{\eta}\right)^d. \tag{71}$$

Now, note that for all $x, y \in \mathbb{X}$ and for all $z \in \mathbb{R}^d$, $\left|D^s K_{x,h}(z) - D^s K_{y,h}(z)\right|$ is upper bounded as

$$\begin{aligned}\left|D^s K_{x,h}(z) - D^s K_{y,h}(z)\right| &= \left|D^s K\left(\frac{x-z}{h}\right) - D^s K\left(\frac{y-z}{h}\right)\right| \\ &\leq \frac{M_K}{h}\left\|(x-z)-(y-z)\right\|_2 = \frac{M_K}{h}\left\|x-y\right\|_2.\end{aligned}$$

Hence for any measure Q on $\mathbb{R}^d$, $\left\|D^s K_{x,h} - D^s K_{y,h}\right\|_{L_2(Q)}$ is upper bounded as

$$\left\|D^s K_{x,h} - D^s K_{y,h}\right\|_{L_2(Q)} = \sqrt{\int (D^s K_{x,h}(z) - D^s K_{y,h}(z))^2 dQ(z)} \leq \frac{M_K}{h}\left\|x-y\right\|_2.$$

Hence applying this to (71) implies that for all $\eta > 0$, the supremum of the covering number $\mathcal{N}(\mathcal{F}_{K,h}, L_2(Q), \eta)$ over all measure Q is upper bounded as

$$\sup_Q \mathcal{N}(\mathcal{F}^s_{K,h}, L_2(Q), \eta) \leq \mathcal{N}\left(\mathbb{X}, \left\|\cdot\right\|_2, \frac{h\eta}{M_K}\right) \leq \left(1 + \frac{2RM_K}{h\eta}\right)^d.$$

Hence for all $\eta \in (0, \left\|D^s K\right\|_\infty)$,

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

□

For Corollary 23, we proceed similarly to the proof of Corollary 15, where we plug in $D^s K$ in the place of K.

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*

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

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

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

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

Proof of Corollary 23. For $x \in \mathbb{X}$, let $D^s K_{x,h} : \mathbb{R}^d \to \mathbb{R}$ be $D^s K_{x,h}(\cdot) = D^s K\left(\frac{x-\cdot}{h}\right)$, and let $\tilde{\mathcal{F}}^s_{K,h} := \left\{\frac{1}{h^{d+|s|}} D^s K_{x,h} : x \in \mathbb{X}\right\}$ be a class of normalized kernel functions centered on $\mathbb{X}$ and bandwidth h.

Note that $D^s \hat{p}_h(x) - D^s p_h(x)$ can be expanded as

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

Hence $\sup_{x\in\mathbb{X}} |D^s \hat{p}_h(x) - D^s p_h(x)|$ can be expanded as

$$\sup_{x\in\mathbb{X}} |D^s \hat{p}_h(x) - D^s p_h(x)| = \sup_{f\in\tilde{\mathscr{F}}_{K,h}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_P[f(X)] \right|. \tag{72}$$

Now, it is immediate to check that

$$\|f\|_\infty \leq h^{-d-|s|} \|D^s K\|_\infty. \tag{73}$$

Also, since $\tilde{\mathscr{F}}^s_{K,h} = h^{-d-|s|}\mathscr{F}^s_{K,h}$, VC dimension is uniformly bounded as Lemma 22 gives that for every probability measure Q on $\mathbb{R}^d$ and for every $\eta \in (0, h^{-d-|s|}\|D^s K\|_\infty)$, the covering number $\mathscr{N}(\tilde{\mathscr{F}}^s_{K,h}, L_2(Q), \eta)$ is upper bounded as

$$\begin{aligned} \sup_Q \mathscr{N}(\tilde{\mathscr{F}}^s_{K,h}, L_2(Q), \eta) &= \sup_Q \mathscr{N}(\mathscr{F}_{K,h}, L_2(Q), h^{d+|s|}\eta) \\ &\leq \left(\frac{2RM_K h^{-1} + \|D^s K\|_\infty}{h^{d+|s|}\eta}\right)^d \\ &\leq \left(\frac{2RM_K \|D^s K\|_\infty}{h^{d+|s|+1}\eta}\right)^d. \end{aligned} \tag{74}$$

Also, Lemma 19 implies that under Assumption 3, for any $\varepsilon \in (0, d_{\text{vol}})$ (and ε can be 0 if $d_{\text{vol}} = 0$ or under Assumption 1),

$$\mathbb{E}_P\left[\left(\frac{1}{h^{d+|s|}} D^s K_{x,h}\right)^2\right] \leq C_{s,P,K,\varepsilon} h^{-2d-2|s|+d_{\text{vol}}-\varepsilon}. \tag{75}$$

Hence from (73), (74), and (75), applying Theorem 30 to (72) gives that $\sup_{x\in\mathbb{X}} |D^s \hat{p}_h(x) - D^s p_h(x)|$

is upper bounded with probability at least $1-\delta$ as

$$
\begin{aligned}
&\sup_{x\in\mathbb{X}}|\hat{p}_h(x)-p_h(x)| \\
&\leq C\left(\frac{2d\left\|D^sK\right\|_\infty \log\left(\frac{2RM_K\|D^sK\|_\infty}{\sqrt{C_{s,P,K,\varepsilon}}h^{1+(d_{\mathrm{vol}}-\varepsilon)/2}}\right)}{nh^{d+|s|}}+\sqrt{\frac{2dC_{s,P,K,\varepsilon}\log\left(\frac{2RM_K\|D^sK\|_\infty}{\sqrt{C_{s,P,K,\varepsilon}}h^{1+(d_{\mathrm{vol}}-\varepsilon)/2}}\right)}{nh^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}}\right. \\
&\qquad\left.+\sqrt{\frac{C_{s,P,K,\varepsilon}\log(\frac{1}{\delta})}{nh^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\|D^sK\|_\infty\log(\frac{1}{\delta})}{nh^d}\right) \\
&\leq C_{R,M_K,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon},\varepsilon} \\
&\quad\times\left(\frac{\left(\log\left(\frac{1}{h}\right)\right)_+}{nh^d}+\sqrt{\frac{\left(\log\left(\frac{1}{h}\right)\right)_+}{nh^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}}+\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nh^{2d+2|s|-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\log\left(\frac{2}{\delta}\right)}{nh^d}\right),
\end{aligned}
$$

where $C_{R,M_K,\|D^sK\|_\infty,d,\nu,d_{\mathrm{vol}},C_{s,P,K,\varepsilon},\varepsilon}$ depends only on R, M_K, $\|D^sK\|_\infty$, d, ν, d_{vol}, $C_{s,P,K,\varepsilon}$, ε.

□