République algérienne démocratique et populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université Les Frères Mentouri Constantine

Faculté des sciences exactes

Département de mathématiques

N$^{o}$ d'ordre : 71/DS/2021.

N$^{o}$ de série : 07/MATHS/2021.

# THÈSE

PRÉSENTÉE POUR L'OBTENTION DU DIPLÔME DE DOCTORAT EN SCIENCES EN MATHÉMATIQUES

# « Estimation non-paramétrique dans un modèle de censure »

Par

EL-HADJALI Thouria

OPTION

Probabilités et Statistique

Devant le jury :

| | | | | |
|---|---|---|---|---|
| Présidente | M$^{me}$ | S. BELALOUI | Prof. | Université frères Mentouri, Constantine 1, |
| Encadreur de thèse | M$^{me}$ | F. MESSACI | Prof. | Université Salah Boubnider, Constantine 3, |
| Co-encadreur de thèse | M. | S. BOUZEBDA | Prof. | Université de technologie de Compiègne, |
| Examinatrice | M$^{me}$ | S. KHARFOUCHI | Prof. | Université Salah Boubnider, Constantine 3. |
| Examinateur | M. | Z. MOHDEB | Prof. | École polytechnique de Constantine, |
| Examinatrice | M$^{me}$ | I.LESCHEB | M.C.A. | Université frères Mentouri, Constantine 1. |

Soutenue le : 01/07/2021.

# Dedication

*This thesis is dedicated to my parents.*

# Acknowledgments

I am extraordinarily grateful to my co-supervisor, Prof. Salim BOUZEBDA, for his insightful guidance, generous support and kind encouragement, without which I couldn't have achieved what I have now. He is a talented researcher and a patient teachers in statistics. Working with him is a real pleasure.

I would like to express my sincere gratitude to my advisor Prof. Fatiha MESSACI who has guided and taught me a lot throughout the journey of this Ph.D. thesis, which allowed me to learn and acquire the spirit of scientific research.

I would like to thank Mrs S. Belaloui for doing me the honor of accepting to chair the jury, and for the time given to the attentive reading of my thesis.

Besides, I would like to thank Mr. Z. Mohdeb, Mrs S. Kharfouchi and Mrs I. Lescheb for doing me the honor of being part of the jury and for offering their valuable time to review and examine my thesis.

My sincere thanks also go to all the members of the the LMAC laboratory for their kind reception and welcome during my PNE scholarship of 18 months.

Thanks you to my parents and my husband for their love and encouragement, without which I never could have succeeded. Their constant support throughout my academic achievements has been incredible.

Finally, I would also like to express my gratitude and love to my sisters Ines, Nesrine, Lamia and my brother Skandar for their love and support. In addition, I would like to thank Fateh Djabli and Lamia Aouicha for their continued help. A special think to my dear and adorable sons Yasser and Zakaria.

# Contents

# General Introduction

The objective of this thesis is the nonparametric estimation of the regression function by the kernel method based on the theory of empirical processes and in presence of censored data.

The theory of statistical estimation is one of the fundamental elements of mathematical statistics. This theory is subdivided into parametric and nonparametric estimation. A nonparametric procedure is usually defined as a procedure which is valid independently of the distribution of the sampled observations and it consists in estimating from the observations some unknown function pertaining to a class of functions which is not in bijection with a finite-dimensional space.

One of the main problems in nonparametric estimation is the estimation of functional characteristics associated with the law of observations, such as, for example, the density function or the regression function (in a multivariate framework).

For the density, there exists a number of methods for nonparametric estimation, based on e.g., kernel smoothing, histograms, orthogonal series, splines, frequency polygons, wavelets or the penalized likelihood.

In this work, we are interested in nonparametric kernel estimation for multivariate models. The kernel estimator of density $f_{\mathbf{X}}(\cdot)$ was introduced by the Akaike-Parzen-Rosenblatt (Akaike (1954), Rosenb1att (1956) and Parzen (1962)) and it can be formulated as follows. Let the kernel $K(\cdot)$ be any function satisfying some regularity conditions and $(h_n)_{n\geq 1}$ be a sequence of positive constants converging to zero and

$$nh_n^d \to \infty \text{ as } n \to \infty.$$

The kernel-type estimator of the density function $f_{\mathbf{X}}(\cdot)$ of $\mathbf{X}$ is given, for $\mathbf{x} \in \mathbb{R}^d$, by

$$f_{\mathbf{X};n}(\mathbf{x};h_n) := \frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right). \tag{0.0.1}$$

Parzen (1962) has shown, under some assumptions on $K(\cdot)$, that $f_{\mathbf{X};n}(\cdot;h_n)$ is an asymptotically unbiased and consistent estimator for $f_{\mathbf{X}}(\cdot)$ whenever $h_n \to 0$, $nh_n^d \to \infty$ and $\mathbf{x}$ is a continuity point of $f_{\mathbf{X}}(\cdot)$. Under some additional assumptions on $f_{\mathbf{X}}(\cdot)$ and $h_n$, he obtained an asymptotic normality result, too. This estimator has been widely studied thereafter and has recently been the subject of an extensive research since it can be easily interpreted and is very often used in practical applications.

Kernel estimation method has been extended to numerous methods for nonparametric estimation of regression function , distribution functions, failure rates, etc. The topic of interest in our thesis is the kernel nonparametric regression.
In statistics, regression analysis models the predictive relationship between responses $\mathbf{Y}$ and predictors $\mathbf{X}$, that is

$$\mathbf{Y} \approx m(\mathbf{X}) + \epsilon,$$

where $(\mathbf{X},\mathbf{Y}) \in \mathbb{R}^d \times \mathbb{R}^q$, $\epsilon$ is unknown parameter represents the error made during model selection and $m(\cdot)$ is a multivariate regression function expressed as a conditional expectation function, that is

$$m(\mathbf{X}) = \mathbb{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}).$$

The primary goal of the regression analysis is to provide an estimate $\widehat{m}(\cdot)$ of $m(\cdot)$ from i.i.d samples $(\mathbf{X}_i,\mathbf{Y}_i)$, and there are three different ways to estimate $m$: parametric approach, semiparametric approach and nonparametric approach.
As with the density, it is worth noticing that the parametric regression models provide useful tools for analyzing practical data when the models are correctly specified, but may suffer from large modelling biases if the structures of the models are misspecified, which is the case in many practical problems. As an alternative, nonparametric smoothing methods ease the concerns on modelling biases. Kernel nonparametric function estimation methods are popular presenting only one of many approaches to the construction of good function estimators, including nearest-neighbor, spline and wavelet methods. These methods have been applied to a wide variety of data.
In this thesis, we shall restrict attention to the construction of consistent kernel-type estimators for multivariate models.
Let $\{\mathbf{X}_i,\mathbf{Y}_i\}_{i\geq 1}$ be an $\mathbb{R}^d \times \mathbb{R}^q$-valued independent vectors defined on a probability space $(\Omega,\mathcal{A},\mathbb{P})$, with $d,q \geq 1$ and Let $\Psi : \mathbb{R}^q \to \mathbb{R}$ be a measurable function,

such that the random variable $\Psi(\mathbf{Y})$ verifies some conditions that will be specified later. We can define the regression function of $\Psi(\mathbf{Y})$ knowing $\mathbf{X}$ (whenever it exists) by

$$m_{\Psi}(\mathbf{x}) := \mathbb{E}(\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}). \tag{0.0.2}$$

A well-known estimator for the regression function $m_{\Psi}(\cdot)$, which is often used in non-parametric statistics, is the so-called kernel regression function estimator introduced by Nadaraya-Watson (Nadaraya (1964) and Watson (1964)). Nadaraya (1964) established similar results to those of Parzen (1962) for $\widehat{m}_{n;h_n}(\mathbf{x})$ as an estimator for $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$.

In this work, we are interested in general kernel-type estimator of regression $m_{\Psi}(\cdot)$ defined, for a bandwidth $h > 0$ $\mathbf{x} \in \mathbb{R}^d$, by

$$r_{\Psi;n}(\mathbf{x}; h) = \frac{1}{nh^d} \sum_{i=1}^{n} \Psi(\mathbf{Y}_i) K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right), \tag{0.0.3}$$

$$m_{\Psi;n}(\mathbf{x}; h) = \begin{cases} \dfrac{r_{\Psi;n}(\mathbf{x}; h)}{f_{\mathbf{X};n}(\mathbf{x}; h)} = \dfrac{\displaystyle\sum_{i=1}^{n} \Psi(\mathbf{Y}_i) K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right)}, & \text{for } f_{\mathbf{X};n}(\mathbf{x}; h) \neq 0, \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \Psi(\mathbf{Y}_i), & \text{for } f_{\mathbf{X};n}(\mathbf{x}; h) = 0. \end{cases} \tag{0.0.4}$$

By setting $\Psi(y) = y$ (or $\Psi(y) = y^k$ ) into (0.0.4), $y \in \mathbb{R}$, we get the classical Nadaraya-Watson kernel regression function estimator $\breve{m}_{n;h_n}(\mathbf{x})$ of $m(\mathbf{x}) := \mathbb{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ given by

$$\widehat{m}_{n;h_n}(\mathbf{x}) := \frac{\displaystyle\sum_{i=1}^{n} Y_i K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)}, \tag{0.0.5}$$

or

$$\breve{m}_{n;h_n}(\mathbf{x}) := \frac{\displaystyle\sum_{i=1}^{n} Y_i^k K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)}. \tag{0.0.6}$$

The asymptotic behavior of nonparametric density estimators, regression function or other functionals, including consistency and limit laws has been studied intensively over the last twenty years. For good sources of references to research literature in this area along with statistical applications consult Tapia and

Thompson (1978), Wertz (1978), Collomb (1981), Devroye and Györfi (1985), Devroye (1987), Silverman (1986), Müller (1988), Nadaraya (1989), Härdle (1990), Wand and Jones (1995), Eggermont *et al.* (2001), Devroye and Lugosi (2001), Györfi *et al.* (2002), Scott (2015), Chacón and Duong (2018) and the references therein.

In practice, a major difficulty in estimating density and regression by the kernel method lies in the choice of the bandwidth $h > 0$. It is important to choose $h$ so that there is a good compromise between the order of bias and the order of variance. The smaller $h$ is, the larger random term (the variance whose order of magnitude is evaluated) and the smaller deterministic term (corresponding to the estimator bias) is. For a large $h$, the opposite effect occurs. One of the first solutions proposed to solve this problem of choice of $h$ was to choose a bandwidth $h$ that minimizes the mean square error [MSE](see Wand and Jones (1995)). The (asymptotically) optimal choice of $h$ depends on the unknown density. To resolve this problem, adaptive bandwidths were introduced in the 1990's, which depends on the available observations and/or their location. There are various adaptive methods of choosing the bandwidth $h$ for $f_{\mathbf{X};n}(\cdot)$. We cite for example the *plug-in* and *cross-validation* methods which do not depend on the location of $x \in \mathbb{R}$, and the *nearest-neighbor* method which depends on the location of $x \in I$. In 2005, following the precursory work of (Deheuvels (2000)), that Einmahl *et al.* (2005) have proposed a practical justification for these solutions which is the "uniform in bandwidth consistency" of kernel density estimators. Using the theory of empirical processes indexed by functions, they have shown that for $c > 0$,

$$\limsup_{n\to\infty} \sup_{c\log n/n\leq h\leq 1} \frac{\sqrt{nh^d}\sup_{\mathbf{x}\in I}|f_{\mathbf{X};n}(\mathbf{x};h)-\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x};h))|}{\sqrt{\log(1/h)\vee\log\log n}} < \infty, \tag{0.0.7}$$

for regular kernels and bounded Lebesgue densities. The interest of this type of result is to ensure the convergence of the estimators under general conditions, even when $h$ is random.

They made use of the indexing of the empirical process by functions combined with the class properties of Vapnik- Červonenkis (see, e.g., Van Der Vaart and Wellner (1996)) and the exponential inequalities of Talagrand type (see Talagrand (1994)).

In the last decades, empirical process theory has provided very useful and powerful tools to analyze the large sample properties of a nonparametric estimators of the regression function and the density function, refer to Pollard (1984), Pollard (1990), Van Der Vaart and Wellner (1996), Kosorok (2008), Dudley (2014). Nolan and Pollard (1987) were the first to introduce the notion of uniform in bandwidth consistency for kernel density estimators, and they applied empirical process methods in their study. In the series of papers, Deheuvels (2000), Einmahl and Mason (2000), Deheuvels and Mason (2004), Giné *et al.* (2004), Einmahl *et al.* (2005), Dony *et al.* (2009), Maillot and Viallon (2009), Giné *et al.* (2009), Mason and Swanepoel (2011), Mason (2012), Giné *et al.* (2013), Mason *et al.* (2015), Bouzebda and Elhattab (2011), Bouzebda (2012), Bouzebda *et al.* (2018) the authors established uniform consistency results for such kernel estimators, where $h_n$ varies within suitably chosen intervals indexed by $n$. In the functional setting, we can refer, among many others, to Kara-Zaitri *et al.* (2017), Ling *et al.* (2019), Novo *et al.* (2019), Bouzebda and Nemouchi (2020). In our work, we will consider one of the most commonly used classes of estimators that is formed by the so-called kernel-type estimators. There are basically no restrictions on the choice of the kernel function in our setup, apart from satisfying some mild conditions those we will give after.

## Major contribution of the thesis

It is worth noticing that the high-dimensional data sets have several unfortunate properties that make them hard to analyze. One of the limiting aspects of density (regression)-based approaches is their performance in high dimensions. It takes an exponential in dimension number of samples to estimate the density (regression).

The phenomenon that the computational and statistical efficiency of statistical techniques deterioate rapidly with the dimension is often referred to as the "curse of dimensionality". Dimensionality reduction methods aim to project the original data set $\Omega \subset \mathbb{R}^d$ without information loss, on to a lower $M$-dimensional manifold of $\mathbb{R}^d$. Since the value of $M$ is unknown, there are several techniques that provide them in advance. Following Fukunaga (1990), the minimum number of parameters required to account for the observed properties of data is the intrinsic (or effective) dimension $d_M$ of the data set. The notion of intrinsic dimension $d_M$ has been studied in the statistical machine learning literature to establish fast estimation rates in high-dimensional kernel density and regression settings. There are numerous known techniques for doing so e.g., Kégl (2003), Levina and Bickel (2004), Hein and Audibert (2005), Farahmand *et al.* (2007).

However, understanding density estimation in situations where the intrinsic dimension can be much lower than the ambient dimension is becoming ever more important: modern systems are able to capture data at an increasing resolution while the number of degrees of freedom stays relatively constant.

Jiang (2017) drifted finite-sample- high-probability density estimation bounds for multivariate kernel density estimators under mild density assumptions (he only required $f_{\mathbf{X}}$ to be bounded) that hold uniformely on $h$ and under appropriate assumptions on $K$, and he extended this results to the manifold setting and these for local instrinsic dimension.

In 2018 Kim *et al.* (2018) have extended the existing uniform in $\mathbf{x} \in \mathbb{R}^d$ and the bandwidth $h$ bounds of kernel density estimators given by Einmahl *et al.* (2005) and Jiang (2017) to more general cases such as the ones of distribution with unbounded densities or supported on a mixture of manifolds with different dimensions and these by weakening the conditions on the kernel and making it adaptive to the intrinsic dimension of the underlying distribution.

Our principal aim in this thesis, is to establish uniform in bandwidth consistency result for some general kernel-type estimators under weaker conditions on the kernel than previously used in the literature and without assumptions on the distribution. We extend the work of Kim *et al.* (2018) to the more general estimators including the kernel density estimator as a particular case (studied in Kim *et al.* (2018)), this generalization is far from being trivial and harder to control some complex classes of functions, which form a basically unsolved open problem in the literature. We aim at filling this gap in the literature by combining results Kim *et al.* (2018) with techniques developed in Einmahl *et al.* (2005). However, as will be seen in chapter 2, the problem requires much more than “simply” combining ideas from the existing results. In fact, delicate mathematical derivations will be required to cope with the empirical processes that we consider in this extended setting. In addition, we will consider the nonparametric Inverse Probability of Censoring Weighted (I.P.C.W.) estimators of the multivariate regression function under random censorship and obtain uniform in bandwidth consistency results which are of independent interest.
The main results of this thesis have been published in an article (Bouzebda and El-hadjali (2020)), written in collaboration with my co-supervisor BOUZEBDA Salim, and published in the journal JOURNAL OF NONPARAMETRIC STATISTICS.

## Organization of the dissertation

This thesis is organized as follows

## Chapter 1. Mathematical framework

In chapter one we will briefly give the general framework considered in this thesis. We define the empirical processes and we are particularly interested in empirical processes indexed by some special classes of functions (called Vapnik-Červonenkis classes) and we give some of their instrumental properties needed for the main results. To be self contained, we will also give some notions about the kernel estimator of density and regression and we will need to know a bit more about volume dimension.

## Chapter 2. Uniform in bandwidth consistency results for general kernel-type estimators

In this chapter we establish two results of the consistency of kernel-type estimators by introducing a generalized multivariate empirical process indexed by a VC class of functions, which will be useful to study various types of kernel-type estimators in chapter 3. In particular, we will treat the uniform consistency in two cases: first in Section 2.3, we consider a bounded class of functions (Theorem 2.3.0.1) and secondly in Section 2.4 the unbounded class of functions (Theorem 2.4.0.1), whenever some moment conditions are satisfied for the envelope function together with entropy conditions.

## Chapter 3. Uniform in bandwidth consistency for nonparametric kernel-type estimators

This chapter is an application of the main results of chapter 2. In particular, we will study the uniform in bandwidth consistency of the kernel type estimators for density, regression, the conditional distribution, multivariate mode and Shannon's entropy. Analogous results are derived for the $s$th derivatives of density and regression functions (see section 3.2), those are of independent interest in the setting of the nonparametric estimation. We will also study in section 3.6, the consistency of the additive regression estimation and we discuss briefly in Section 3.7 the bandwidth selection criterion .

## Chpater 4. Uniform in bandwidth consistency for nonparamteric I.P.C.W. estimators of the regression function in censored case

The purpose of this chapter is to show the uniform in bandwidth consistency for non parametric regression function $m_\psi(\mathbf{x}) = \mathbb{E}(\psi(Y) \mid \mathbf{X} = \mathbf{x})$, when $Y$ is right-censored . To cope with this problem, we will apply the results of Chapter 3, in particular, we establish, using these methods, the asymptotic behaviour of a regression estimator of the type "Inverse Probability of Censoring Weighted estimators" [I.P.C.W.] introduced by Kohler *et al.* (2002). Applications of the main results include the kernel type estimators of the conditional density and the conditional distribution.

## Conclusions and perspectives

In this chapter, we conclude this thesis and we present some open questions and perspectives which appeared during the preparation of this thesis.

# Chapter 1

# Mathematical framework

Empirical process theory plays a central role in statistics, since it concerns the set of general borderline results relating to random samples. As a result, it has innumerable applications to specific problems. Among the more important properties of these mathematical models, which have been the subject of extensive research since the origin of modern statistics, include, among others, results related to the theorems of Glivenko-Cantelli (Glivenko (1933),Cantelli (1933)), the Donsker's theorems (Donsker (1951)), to the laws of the iterated logarithm (Chung (1949)), see also Deheuvels (1991)) and the accompanying bibliography), or functional limit laws, global or local, (Finkelstein *et al.* (1971), Deheuvels (1992), Deheuvels and Mason (1992), Deheuvels (2000)). In the early 1980s, Stute (see Stute (1982), Stute (1986a), Stute (1986b), and Stute (1986c)) was one of the first statisticians to make systematic use of methods derived from empirical process theory in the study of the asymptotic properties of nonparametric functional estimators and more particularly kernel estimators (see also Csörgő and Révész (1981) and Deheuvels (1974)).

In this chapter, we will describe the frameworks in which this thesis works take place, and we will present mathematical tools used in the following chapters.

## 1.1 Empirical processes

In this part, we will mainly be involved with "empirical measures" and "empirical processes".

Empirical process theory began in the 1930's and 1940's with the study of the empirical distribution function and the corresponding empirical process, this theory is very useful because many statistics can be expressed as functionals of the empirical distribution function denoted $\mathbb{F}_n(\cdot)$.

Now, suppose that $X_1, \dots, X_n$ are i.i.d random variables defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution function $F(\cdot)$.

A *stochastic process* is a collection of random variables $\{X(t) : t \in T\}$ on the same probability space, indexed by an arbitrary index set $T$. If $X_1, \dots, X_n$ are i.i.d real-valued random variables, then *empirical distribution function* is

$$\mathbb{F}_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}, \tag{1.1.1}$$

where the index $t$ is allowed to vary over $T = \mathbb{R}$, the real line. The corresponding empirical processes $\mathbb{G}_n(t)$ is defined as

$$\mathbb{G}_n(t) = \sqrt{n}(\mathbb{F}_n(t) - F(t)),$$

By the law of large numbers, we know that, for every $t \in \mathbb{R}$

$$\mathbb{F}_n(t) \xrightarrow{a.s} F(t),$$

where $\xrightarrow{a.s}$ denotes almost sure convergence. By the central limit theorem, for each $t \in \mathbb{R}$,

$$\mathbb{G}_n(t) \xrightarrow{d} N(0, F(t)(1 - F(t))),$$

where $\xrightarrow{d}$ denotes convergence in distribution. To generalize the above two results to processes that hold for all $t$ simultaneously, two important results in empirical process theory concerning $\mathbb{F}_n(\cdot)$ and $\mathbb{G}_n(\cdot)$ are given below.

**Theorem 1.1.0.1 (Glivenko (1933), Cantelli(1933))**

$$\sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)| \xrightarrow{a.s} 0.$$

**Theorem 1.1.0.2 (Donsker, 1952)**

$$\mathbb{G}_n(t) \xrightarrow{d} G, \quad in \quad \ell^\infty(\mathbb{R}),$$

*where, for any index set $T$, $\ell^\infty(T)$ is the collection of all bounded functions $f : T \mapsto \mathbb{R}$.*

In the 1950's and 1960's, the need for generalization of Theorem 1.1.0.1 and Theorem 1.1.0.2 became apparent that when the observations take values in a more general arbitrary sample space $\mathcal{X}$ (such as $\mathbb{R}^d$ or a Riemannian manifold, etc.), in this case the empirical distribution function is not as natural, it becomes much more natural to define the *empirical measure* $\mathbb{P}_n$ indexed by some class of real-valued functions $\mathcal{F}$ defined on $\mathcal{X}$.
We consider a random sample $X_1, \dots, X_n$ of independent draws from a probability measure $P$ on $\mathcal{X}$. The *empirical measure* is defined by linear combination of the Dirac measures $\delta_x$ as

$$\mathbb{P}_n := n^{-1} \sum_{i=1}^{n} \delta_{X_i}.$$

For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, the empirical measure induces a map from $\mathcal{F}$ to $\mathbb{R}$ given by

$$\mathbb{P}_n f := n^{-1} \sum_{i=1}^{n} f(X_i).$$

If $\mathcal{F}$ is a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$, then $\{\mathbb{P}_n f : f \in \mathcal{F}\}$ is the empirical process indexed by a class of functions $\mathcal{F}$. This definition is more general as it describes all measurable functions of the sample. The definition (1.1.1) is obtained if $\mathcal{X} = \mathbb{R}$, and we re-express $\mathbb{F}_n(\cdot)$ as the empirical process $\{\mathbb{P}_n(f) : f \in \mathcal{F}\}$, where $\mathcal{F} = \{\mathbb{1}\{x \leq t\} : t \in \mathbb{R}\}$.
The *general empirical processes* $\mathbb{G}_n$ is defined by

$$\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - \mathbb{P}),$$

and the collection of random variables $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ as $f$ varies over $\mathcal{F}$ is called the *empirical process* indexed by the class of functions $\mathcal{F}$, where

$$\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbb{P}f).$$

*The objective of empirical process theory is to study the properties of the approximation of $\mathbb{P}f$ by $\mathbb{P}_n f$, uniformly in $\mathcal{F}$.*

With the notation $\|Q\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |Qf|$, a class $\mathcal{F}$ of measurable functions is said to be a *P-Glivenko-Cantelli* class if

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s} 0, \tag{1.1.2}$$

$\mathcal{F}$ is called P-Donsker if under suitable conditions

$$\mathbb{G}_n \xrightarrow{d} \mathbb{G}, \quad \text{in } \ell^{\infty}(\mathcal{F}), \tag{1.1.3}$$

where the limit process $\mathbb{G}$ is a tight Borel measurable element in $\ell^{\infty}(\mathcal{F})$.

## Size of a function class

In empirical process theory, it is important to measure the size of a given class $\mathcal{F}$ of measurable functions defined on $\mathcal{X}$. Whether a given class $\mathcal{F}$ is Glivenko-Cantelli or Donsker depends on the "size" (or"complexity") of the class. A finite class of integrable functions is always Glivenko- Cantelli and a finite class of square integrable functions is always Donsker. A relatively simple way to measure the size of a class $\mathcal{F}$ is to use entropy numbers. The $\epsilon$-entropy of $\mathcal{F}$ is essentially the logarithm of the number of "balls" or "brackets" of size $\epsilon$ needed to cover $\mathcal{F}$. From this informal definition, it is already clear that the entropy numbers increase as $\epsilon$ decreases to zero.

In the following paragraph, we give a way to measure the size of a class $\mathcal{F}$ in terms of entropy.

### Covering numbers

For $1 \leq r$, let $L_r(\mathbb{P})$ denote the collection of functions $g : \mathcal{X} \mapsto \mathbb{R}$ such that $\|g\|_{r,p} := [\int_{\mathcal{X}} |g(x)|^p d\mathbb{P}(x)]^{1/r} < \infty$.

**Definition 1.1.0.3** *(Kosorok (2008))*
*For a probability measure $\mathbb{Q}$, the covering number $\mathcal{N}(\mathcal{F}, L_r(\mathbb{Q}), \varepsilon)$ is the minimum number of $L_r(\mathbb{Q})$-balls of radius $\epsilon$ needed to cover $\mathcal{F}$, where an $L_r(\mathbb{Q})$-balls of radius $\varepsilon$ is the set $\{h \in L_r(\mathbb{Q}) : \|h - g\|_{\mathbb{Q},r} < \varepsilon\}$.*

The centers of the balls need not belong to $\mathcal{F}$, but they should have finite norms. The *uniform covering numbers* are given by

$$\sup_{\mathbb{Q}} \mathcal{N}(\mathcal{F}, L_r(\mathbb{Q})), \epsilon \|F\|_{\mathbb{Q},r}), \tag{1.1.4}$$

where $F : \mathcal{X} \mapsto \mathbb{R}$ is an *envelope* for $\mathcal{F}$, meaning that $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$, and where the supremum is taken over all finitely discrete probability measure $\mathbb{Q}$ with $\|F\|_{\mathbb{Q},r} > 0$. The minimal envelope function is $x \mapsto \sup_f |f(x)|$. It will usually be assumed that this function is finite for every $x$. Notice that the uniform covering number does not depend on the probability measure $\mathbb{P}$ for the observed data.

The *entropy* is the logarithm of of the covering number and the uniform entropy numbers are defined as

$$\sup_{\mathbb{Q}} \log \mathcal{N}(\mathcal{F}, L_r(\mathbb{Q}), \varepsilon \|F\|_{\mathbb{Q},r}),$$

where the supremum is over all probability measures $\mathbb{Q}$ on $(\mathcal{X}, \mathcal{A})$, with $0 < \mathbb{Q}F^r < \infty$.

### Vapnik- Červonenkis (VC) classes of functions

One of the starting points for controlling uniform covering numbers is the so-called Vapnik- Červonenkis classes, or VC *classes*. There are several non- equivalent definitions of VC classes of functions. We will only mention here the so-called VC subgraph classes of functions (for the definitions of the VC classes of functions called *major* and *hull*, we refer the reader to Van Der Vaart and Wellner (1996) and Kosorok (2008)). In our framework, the most interesting definition is that of the VC subgraph classes of functions, since they present the property of polynomial covering number. The other types of VC-classes of functions usually have only polynomial entropy numbers, which is essentially enough to show that, for example, a class is Glivenko- Cantelli or Donsker. For an extensive exposition on VC-classes, we refer the reader to the books of Van Der Vaart and Wellner (1996), Kosorok (2008).
First, we introduce VC classes of sets, and VC classes of functions.

**Definition 1.1.0.4** *Say that the collection $\mathcal{C}$ of subsets of the sample space $\mathcal{X}$ picks out a certain subset $A$ of the finite set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ if it can be written as $A = \{x_1, \ldots, x_n\} \cap C$, where $C \in \mathcal{C}$.*

**Definition 1.1.0.5** *The collection $\mathcal{C}$ is said to shatter $\{x_1, \ldots, x_n\}$ if $\mathcal{C}$ picks out each of its $2^n$ subsets.*

**Definition 1.1.0.6** *The VC index $V(\mathcal{C})$ of $\mathcal{C}$ is the smallest $n$ for which no set of size $n$ is shattered by $\mathcal{C}$.*

A collection $\mathcal{C}$ of measurable sets is called a *VC class* if its index $V(\mathcal{C})$ is finite.

**Definition 1.1.0.7** *A collection $\mathcal{F}$ is a VC class of functions if the collection of all subgraphs $\{(x,t) : f(x) < t\}$, if $f$ ranges over $\mathcal{F}$, forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$.*

It has been shown that any class $\mathcal{F}$ of measurable functions on a measure space $(\mathcal{X}, \mathcal{A})$ is Vapnik- Červonenkis (VC) class of functions with respect to an envelope function $F$ if there exists a measurable function $F$ everywhere finite such that

$|f| < F$, for all $f \in \mathcal{F}$ and if there exists finite numbers $A$ and $\nu$ such that $0 < \varepsilon < 1$, and for all $\mathbb{Q}$ probability measure on $(\mathcal{X}, \mathcal{A})$ such that $\int \mathbf{F}^2 d\mathbb{Q} < \infty$, we have

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_{L_2(\mathbb{Q})}, \varepsilon\|F\|_{L_2(\mathbb{Q})}) \leq \left(\frac{A}{\varepsilon}\right)^{\nu}.$$

Consequently, VC-classes are examples of *polynomial classes* in the sense that their covering numbers are bounded by a polynomial in $1/\varepsilon$.
Now, we present two important examples of VC-classes of functions:

**Example 1.1.0.8** *A finite-dimensional vector of measurable functions from $\mathcal{X} \times \mathbb{R}$ is VC subgraph with $V(\mathcal{F}) \leq \dim(\mathcal{F}) + 2$.*

**Example 1.1.0.9** *The class of indicator functions of the type $\mathbb{1}\{(-\infty, t]\}$, for $t \in \mathbb{R}$, or $\mathbb{1}\{(s, t]\}$, for $(s, t] \in \mathbb{R}^2$ is a VC class of functions.*

The next lemmas (LEMMA 9.7 and 9.9 of Kosorok (2008)), consist of useful tools for building VC-classes from other VC-classes.

**Lemma 1.1.0.10** *Let $\mathcal{C}$ and $\mathcal{D}$ be VC-classes of sets in a set $\mathcal{X}$, with respective VC-indices $V_{\mathcal{C}}$ and $V_{\mathcal{D}}$; and let $\mathcal{E}$ be a VC-class of sets in $\mathcal{W}$, with VC-index $V_{\mathcal{E}}$. Also let $\phi : \mathcal{X} \mapsto \mathcal{Y}$ be a fixed function. Then*

**(a)** *$\mathcal{C} \sqcap \mathcal{D} := \{C \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC with index $\leq V_{\mathcal{C}} + V_{\mathcal{D}} - 1$;*

**(b)** *$\mathcal{C} \sqcup \mathcal{D} := \{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC with index $\leq V_{\mathcal{C}} + V_{\mathcal{D}} - 1$;*

**(c)** *$\mathcal{D} \times \mathcal{E}$ is VC index in $\mathcal{X} \times \mathcal{W}$ with VC index $\leq V_{\mathcal{D}} + V_{\mathcal{E}} - 1$;*

**(d)** *$\phi(\mathcal{C})$ is VC with index $V_{\mathcal{C}}$ if $\phi$ is one to one.*

**Lemma 1.1.0.11** *Let $\mathcal{F}$ and $\mathcal{G}$ be VC-subgraph classes of functions on a set $\mathcal{X}$, with respective VC indices $V_{\mathcal{F}}$ and $V_{\mathcal{G}}$. Let $g : \mathcal{X} \mapsto \mathbb{R}, \phi : \mathbb{R} \mapsto \mathbb{R}$, and $\psi : \mathcal{Z} \mapsto \mathcal{X}$ be fixed functions. Then*

**(a)** *$\mathcal{F} \wedge \mathcal{G} := \{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$ is VC -subgraph with index $\leq V_{\mathcal{F}} + V_{\mathcal{G}} - 1$;*

**(b)** *$\mathcal{F} \vee \mathcal{G}$ is VC with index $\leq V_{\mathcal{F}} + V_{\mathcal{G}} - 1$;*

**(c)** *$\{\mathcal{F} > 0\} := \{\{f > 0\} : f \in \mathcal{F}\}$ is a VC-class of sets with index $V_{\mathcal{F}}$;*

**(d)** *$-\mathcal{F}$ is VC-subgraph with index $V_{\mathcal{F}}$;*

**(e)** *$\mathcal{F} + g := \{f + g : f \in \mathcal{F}\}$ is VC with index $V_{\mathcal{F}}$;*

**(f)** *$\mathcal{F} \cdot g := \{fg : f \in \mathcal{F}\}$ is VC with index $\leq 2V_{\mathcal{F}} - 1$;*

**(g)** *$\mathcal{F} \circ \psi := \{f(\psi) : f \in \mathcal{F}\}$ is VC with index $\leq V_{\mathcal{F}}$.*

The following lemma proved in Einmahl and Mason (2000) is very helpful in our proofs. It provides stability by Cartesian products of classes of functions with a polynomial covering number.

**Lemma 1.1.0.12** *Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of real valued measurable functions on $\mathcal{H}$ satisfying*

$$|f(x)| \leq F(x), \quad f \in \mathcal{F}, \quad x \in \mathcal{H}, \tag{1.1.5}$$

*where $F$ is a finite valued measurable envelope function on $\mathcal{H}$;*

$$\|g\|_\infty \leq M, \quad g \in \mathcal{G},$$

*where $M > 0$ is a finite constant. Assume that for all p-measures $\mathbb{Q}$ with $0 < \mathbb{Q}(F^2) < \infty$,*

$$\mathcal{N}(\mathcal{F}, \varepsilon(\mathbb{Q}(F^2))^{1/2}, d_{\mathbb{Q}}) \leq C_1 \varepsilon^{-\nu_1}, \quad 0 < \varepsilon < 1,$$

*and for all probability measure $\mathbb{Q}$,*

$$\mathcal{N}(\mathcal{G}, \varepsilon M, d_{\mathbb{Q}}) \leq C_2 \varepsilon^{-\nu_2}, \quad 0 < \varepsilon < 1,$$

*where $\nu_1, \nu_2, C_1, C_2 \geq 1$ are suitable constants. Then we have for probability measures $\mathbb{Q}$, with $\mathbb{Q}(F^2) < \infty$,*

$$\mathcal{N}(\mathcal{F}\mathcal{G}, \varepsilon M(\mathbb{Q}(F^2))^{1/2}, d_{\mathbb{Q}}) \leq C_3 \varepsilon^{-\nu_1 - \nu_2}, \quad 0 < \varepsilon < 1,$$

*for some finite constant $0 < C_3 < \infty$.*

## 1.2 Non-parametric kernel estimators

The problem of non-parametric estimation consists, in most cases, in estimating, from observations, an unknown function, an element of a certain functional class. Recall that a non-parametric procedure is defined independently of the distribution or law of the sample of observations. More specifically, a non-parametric estimation method is defined when it does not boil down to the estimation of a finite number of real parameters associated with the distribution of the sample.

### 1.2.1 Kernel density estimator

There exists a number of methods for non-parametric density estimation, based on e.g., kernel smoothing, histograms, orthogonal series, splines, frequency polygons, wavelets or the penalized likelihood. An extensive survey of these topics can be found for example in Devroye and Lugosi (2001). In 1956, Rosenblatt proposed a kernel density estimators KDEs obtained by a convolving the empirical distribution function and an appropriate function, called kernel $K(\cdot)$, [Akaike (1954), Rosenb1att (1956), Parzen (1962)]. This method of estimation is popular because of their conceptual simplicity and nice theoretical properties. Formally, let $f(\cdot)$ be a probability density with respect to Lebesgue measure on $\mathbb{R}^d$, and let $(\mathbf{X}, \mathbf{X}_i, i \geq 1)$ be an independent and identically distributed (i.i.d) $\mathbb{R}^d-$valued random variables with unknown Borel probability distribution $\mathbb{P}$. For a given kernel $K(\cdot)$, where $K(\cdot)$ is an appropriate function on $\mathbb{R}^d$ (often a density), the kernel density estimator (KDE) of $f_{\mathbf{X}}(\cdot)$ known as the Akaike-Parzen-Rosenblatt (see Akaike (1954), Rosenb1att (1956), Parzen (1962)) with kernel $K$ and bandwidth $(h_n)_{n\geq 1}$, is defined as

$$\mathbf{x} \in \mathbb{R}^d \mapsto f_{\mathbf{X};n}(\mathbf{x}; h_n) := \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right). \tag{1.2.1}$$

where $(h_n)_{n\geq 1}$ is a sequence of positive constants converging to zero and $nh_n^d \to \infty$ as $n \to \infty$.

These kernel density estimators have remained popular over time due to their consistency properties. It is well known that when the bandwidth $h_n > 0$ converges to 0 and $nh_n^d$ converges to infinity, these estimators are convergent in $L_1$, whether in probability, average or almost surely.

To establish the strong consistency, one usually writes the difference $f_{\mathbf{X};n} - f_{\mathbf{X}}$ as the sum of a probabilistic term $f_{\mathbf{X};n} - \mathbb{E}f_{\mathbf{X};n}$ and the bias (deterministic term) $\mathbb{E}f_{\mathbf{X};n} - f_{\mathbf{X}}$. The order of the bias depends on the smoothness properties of the density $f_{\mathbf{X}}(\cdot)$, whereas the random term can be studied, under existence of Lebesgue density and fixed bandwidth (see, Prakasa Rao (1983), Giné and Guillou (2002), Sriperumbudur and Steinwart (2012), Steinwart *et al.* (2017)) or via empirical process techniques (see Stute (1982), Stute (1984), Pollard (1984)), among other authors. Deheuvels (2000) for one-dimensional case, and Giné and Guillou (2002) have shown that if $K(\cdot)$ is a regular kernel, the density function $f_{\mathbf{X}}(\cdot)$ is bounded and $h_n$ satisfies some regularity conditions,

$$\|f_{\mathbf{X};n} - \mathbb{E}f_{\mathbf{X};n}\|_\infty = O(\sqrt{|\log h_n|/nh_n}).$$

In addition, this rate cannot be improved. Interestingly, there is no need for continuity of $f_{\mathbf{X}}(\cdot)$ for this result. (Of course, continuity of $f_{\mathbf{X}}(\cdot)$ is crucial for controlling the bias).

### Adaptive kernel density estimators

Varying the bandwidth along the support of the sample data gives flexibility to reduce the variance of the estimates in areas with few observations, and reducing the bias of the estimates in areas with many observations. Kernel density estimation methods relying on such varying bandwidths are generally referred to as *adaptive kernel* density estimation methods. Uniform in bandwidth consistency for KDE have been received relatively less attention, although such consistency of KDEs with adaptive bandwidth may depend on the location of $\mathbf{x}$. Deheuvels (2000), Einmahl *et al.* (2005) proved the almost sure uniform convergence of the kernel density estimator, with bounded Lebesgue densities if the bandwidth varies within an interval $[a_n, b_n]$, they have shown that

$$\limsup_{n\to\infty} \sup_{c\log n/n\leq h\leq 1} \frac{\sqrt{nh^d}\|f_{\mathbf{X};n} - \mathbb{E}f_{\mathbf{X};n}\|_\infty}{\sqrt{\log(1/h) \vee \log\log n}} < \infty,$$

Jiang (2017) provided a finite-sample bound of $\|f_{\mathbf{X};n} - \mathbb{E}f_{\mathbf{X};n}\|_\infty$ that holds uniformly on $h$ and under appropriate assumptions on $K(\cdot)$, and extended it to case of densities over well-behaved manifolds. In 2019, Kim *et al.* (2019) extend existing uniform bounds on KDEs by weakening the conditions on $K(\cdot)$ and making it adaptive to the so-called *intrinsic dimension* of the underlying distribution noted by $d_{vol}$, and they obtained a result similar to that of Einmahl *et al.* (2005), given by

$$\sup_{h\geq l_n} \sup_{\mathbf{x}\in\mathbb{X}} |f_{\mathbf{X};n}(\mathbf{x}; h) - \mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h))| \leq C\sqrt{\frac{\log(1/l_n)_+ + \log(2/\delta)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{1.2.2}$$

## 1.2.2 Kernel regression estimator

Several paradigms of non-parametric regression estimation $m(\mathbf{x}) := \mathbb{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ are available, including, the wight estimation, least squares estimation, and penalized least square estimation or smoothing spline. Weight estimation is declined in several models including the partition estimator, the nearest neighbor $k$ estimator and the Nadaraya-Watson kernel estimator. The latter, which was introduced independently by Nadaraya (1964) and Watson (1964), is one of the most popular non-parametric regression models because it is uniformly best in terms of integrated mean square error,in the sens that it turns out to be optimal in the minimax sense.

Let $\{\mathbf{X}_i, Y_i\}_{i\geq 1}$ be an $\mathbb{R}^d \times \mathbb{R}$-valued independent vectors defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with $d \geq 1$, the expression of the classical Nadaraya-Watson kernel regression function estimator of $m(\mathbf{x})$ is given by

$$\widehat{m}_{n;h_n}(\mathbf{x}) := \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}, \tag{1.2.3}$$

which is in the form of a weighted local average of the $Y_i$ values. The convergence of this estimator to $m(\mathbf{x})$, has been studied intensively over the last few years. We refer, in particular, to Collomb (1981), Bosq (1987),Nadaraya (1964), Härdle (1990), Györfi *et al.* (2002), for additional details and references. The consistency of kernel regression estimators when the bandwidth $h_n$ varies within suitably chosen intervals indexed by $n$, has been established using empirical process methods, in the series of papers, Einmahl and Mason (2000), Deheuvels and Mason (2004), Einmahl *et al.* (2005), Dony *et al.* (2009), Maillot and Viallon (2009), Mason and Swanepoel (2011), Mason (2012), Giné *et al.* (2013), Bouzebda and Elhattab (2011), Bouzebda (2012), Bouzebda *et al.* (2018) and Bouzebda and Nemouchi (2020).

## 1.3 Volume dimension

If $\mathbf{X}$ takes values in a high-dimensional space (i.e., if $d$ is large), it is particularly difficult to estimate the regression function. The reason for this is that in the case of large $d$ it is, in general, not possible to densely pack the space of $\mathbf{X}$ with finitely many sample points, even if the sample size $n$ is very large.

Density and regression estimation on manifolds has received much less attention than the "fulldimensional" counterpart.

In this section, we have to know the intrinsic dimension $d_M$, such a notion has been studied in the statistical machine learning literature, so as to establish fast estimation rates in high-dimensional kernel regression settings.

We first introduce a concept proposed by Kim *et al.* (2018, 2019), the so called *volume dimension*, to characterize the intrinsic dimension of the underlying distribution. To be more specific, the volume dimension $d_{\text{vol}}$ is the rate of decay of the probability of vanishing Euclidean balls.

Let $\|\cdot\|$ be the Euclidean 2-norm. For $\mathbf{x} \in \mathbb{R}^d$ and $r > 0$, we use the notation

$\mathbb{B}_{\mathbb{R}^d}(\mathbf{x}, r)$ for the open Euclidean ball centered at $\mathbf{x}$ and radius $r$, i.e.,

$$\mathbb{B}_{\mathbb{R}^d}(\mathbf{x}, r) = \left\{ y \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| < r \right\}.$$

When a probability distribution $\mathbb{P}$ has a bounded density $f_{\mathbf{X}}(\cdot)$ with respect to a well-behaved manifold $M$ of dimension $d_M$, it is known that, for any point $\mathbf{x} \in M$, the measure on the ball $\mathbb{B}_{\mathbb{R}^d}(\mathbf{x}, r)$ centered at $\mathbf{x}$ and radius $r$ decays as

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

when $r$ is small enough. From this, Kim *et al.* (2018) define the volume dimension of a probability distribution $\mathbb{P}$ to be the maximum possible exponent rate that can dominate the probability volume decay on balls, i.e., fix a subset $\mathbb{X} \subset \mathbb{R}^d$, then

$$d_{\text{vol}}(\mathbb{P}) := \sup \left\{ \nu \geq 0 : \limsup_{r \to 0} \sup_{\mathbf{x} \in \mathbb{X}} \frac{\mathbb{P}(\mathbb{B}_{\mathbb{R}^d}(\mathbf{x}, r))}{r^{\nu}} < \infty \right\}. \tag{1.3.1}$$

Now, we establish some notation that are used throughout the manuscript. For more detailed definitions, see (Kim *et al.* (2018), Appendix A). The Hausdorff measure is a generalization of the Lebesgue measure to lower dimensional subsets of $\mathbb{R}^d$. For $\nu \in \{1, \ldots, d\}$, let $\lambda_\nu$ be a normalized $\nu$-dimensional Hausdorff measure on $\mathbb{R}^d$ satisfying that its measure on any $\nu$-dimensional unit cube is $1$. We use the notation

$$\omega_\nu := \lambda_\nu(\mathbb{B}_{\mathbb{R}^\nu}(0, 1)) = \frac{\pi^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2} + 1\right)},$$

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. The volume dimension is a natural generalization of the dimension of a manifold. If a probability distribution has a positive measure on a manifold with positive reach satisfying

$$\mathbb{P}(M \cap \mathbb{X}) > 0,$$

then $0 \leq d_{\text{vol}} \leq d_M$ and in particular, the volume dimension of any probability distribution is between $0$ and the ambient dimension $d$.

**Remark 1.3.0.1** *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 $d$. But, if the probability distribution $\mathbb{P}$ has an unbounded density, 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, (for more details see Proposition 3 in* Kim et al. (*2018*)). *we give an example from* Kim et al. (*2018*) *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 1.3.0.2 (Kim et al. (2018))*** *Let $\mathbb{P}$ 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*

$$p(\mathbf{x}) = \frac{(d-\beta)\Gamma\left(\frac{d}{2}\right)}{2\pi^{\frac{d}{2}}} \|\mathbf{x}\|_2^{-\beta} \mathbb{1}(\|\mathbf{x}\|_2 \leq 1).$$

*Then, for each fixed $r > 0$,*

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

*Hence from definition in 1.3.1, the volume dimension is*

$$d_{\text{vol}}(\mathbb{P}) = d - \beta.$$

# Chapter 2

# Uniform in bandwidth consistency results for general kernel-type estimators

## Introduction

In this chapter, we establish a general theorems (Theorem 2.3.0.1 and Theorem 2.4.0.1) of the consistency of kernel type estimators. To obtain these theorems, we will make use of properties, suitably selected, of the multivariate empirical process, indexed by classes of functions, which will serve as a working basis. This process is defined in (2.0.1) below. Set, for $\mathbf{x} \in \mathbf{I}$ and any $h > l_n$

$$\begin{aligned} W_{n;h}(\mathbf{x};\Psi) &= \sum_{j=1}^{n}(c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_j) + d_\Psi(\mathbf{x}))K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right) \\ &\quad -n\mathbb{E}\left\{(c_\Psi(\mathbf{x})\Psi(\mathbf{Y}) + d_\Psi(\mathbf{x}))K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right\} \qquad (2.0.1) \end{aligned}$$

where $h > 0$ is a bandwidth parameter, $\Psi \in \mathcal{F}_q$, $l_n$ is a positive sequence approaching $0$ and $c_\Psi(\cdot)$ and $d_\Psi(\cdot)$ are continuous functions on a compact set $\mathbf{I} \subset \mathbb{X}$.

In view of (2.0.1), the introduction of the process $W_{n;h}(\mathbf{x};\Psi)$ provides a suitable and general set-up to study various types of kernel estimators (e.g. the density and regression functions, the conditional distribution, multivariate mode and Shannon's entropy) discussed in Chapter 3, for related details the reader is referred to Einmahl and Mason (2000), Deheuvels and Mason (2004), Einmahl *et al.* (2005), Bouzebda and Elhattab (2009, 2011), Bouzebda (2012), Bouzebda and Nemouchi (2020).

In order to present the main results, we start by presenting the hypotheses in Section 2.1. We give some notation used in this manuscript and we provide also the hypotheses on the class of functions and the kernel functions. We follow Kim *et al.* (2018) for weakening the conditions on the kernel functions (integrability conditions on the kernel, see Assumption 3). In Section 2.3, we establish the consistency of the estimator $W_{n;h}(\mathbf{x};\Psi)$ when the class of functions $\mathcal{F}_q$ is bounded. In Section 2.4, we treat the unbounded case, where the envelope functions satisfy some moment conditions.

## 2.1 Notation and assumptions

Let $(\mathbf{X}_1,\mathbf{Y}_1),(\mathbf{X}_2,\mathbf{Y}_2),\ldots,$ be a sequence of independent and identically distributed $\mathbb{R}^d\times\mathbb{R}^q$-valued random variables with $d,q\geq 1$. Let $\mathbb{P}_{\mathbf{X}}=\mathbb{P}$ be an unknown marginal Borel probability distribution in $\mathbb{R}^d$.

For a specified measurable function $\Psi$, we consider the regression function of $\Psi(\mathbf{Y})$ evaluated in $\mathbf{X}=\mathbf{x}$, for $\mathbf{x}\in\mathbb{X}=\mathrm{supp}(\mathbb{P})$ and $\Psi\in\mathcal{F}_q$,

$$m_\Psi(\mathbf{x})=\mathbb{E}(\Psi(\mathbf{Y})\mid\mathbf{X}=\mathbf{x}),$$

where $\mathcal{F}_q$ is a class of measurable functions defined on $\mathbb{R}^q$ with a measurable envelope function $F$ that is,

$$F(\mathbf{y})\geq\sup_{\Psi\in\mathcal{F}_q}\mid\Psi(\mathbf{y})\mid,\quad \mathbf{y}\in\mathbb{R}^q, \tag{F.i}$$

which fulfills the following assumptions. We will work under assumption (F.ii) and (F.iii) below.

First, to avoid measurability problems, we assume that

$$\mathcal{F}_q \text{ is pointwise measurable.} \tag{F.ii}$$

That is, there exists a countable subclass $\mathcal{F}_0$ of $\mathcal{F}_q$ such that we can find, for any function $g\in\mathscr{F}_q$, a sequence of functions $g_m\in\mathcal{F}_0$ for which

$$g_m(\mathbf{z})\rightarrow g(\mathbf{z}),\ \ \mathbf{z}\in\mathbb{R}^q.$$

This condition is discussed in (Van Der Vaart and Wellner, 1996, Example 2.3.4. p 110) and (Kosorok, 2008, §8.2. p. 110).

We assume that $\mathcal{F}_q$ is of VC-type, with characteristics $A_1$ and $\nu_1$ ("VC" for Vapnik and Červonenkis), meaning that for some $A_1\geq 3$ and $\nu_1\geq 1$,

$$\mathcal{N}(\mathcal{F}_q,L_2(\mathbb{Q}),\varepsilon)\leq\left(\frac{A_1\|F\|_{L_2(\mathbb{Q})}}{\varepsilon}\right)^{\nu_1},\ \ \text{for}\ \ 0<\varepsilon\leq 2\|F\|_{L_2(\mathbb{Q})}, \tag{F.iii}$$

where $\mathbb{Q}$ is any probability measure on $(\mathbb{R}^q, \mathcal{B})$, where $\mathcal{B}$ represents the $\sigma$-field of Borel sets of $\mathbb{R}^q$, such that $0 < \|F\|_{L_2(\mathbb{Q})} < \infty$, and where for $\varepsilon > 0, \mathcal{N}(\mathcal{F}_q, L_2(\mathbb{Q}), \varepsilon)$ is defined as the smallest number of $L_2(\mathbb{Q})$-open balls of radius $\varepsilon$ required to cover $\mathcal{F}_q$.

If (F.iii) holds for $\mathcal{F}_q$, then we say that the VC-type class $\mathcal{F}_q$ admits the characteristics $A_1$ and $\nu_1$. For instance, see (Pollard, 1984, Examples 26 and 38), (Nolan and Pollard, 1987, Lemma 22), (Dudley, 2014, §4.7.), (Van Der Vaart and Wellner, 1996, Theorem 2.6.7), (Kosorok, 2008, §9.1) provide a number of sufficient conditions under which (F.i) holds, we may refer also to (Deheuvels, 2011, §3.2) for further discussions. For instance, it is satisfied, for general $d \geq 1$, whenever $g(\mathbf{x}) = \phi(\tau(\mathbf{x}))$, with $\tau(\mathbf{x})$ being a polynomial in $d$ variables and $\phi(\cdot)$ being a real-valued function of bounded variation, we refer the reader to (Einmahl *et al.*, 2005, p. 1381). Notice that condition (F.ii) implies that the supremum in (F.i) is measurable.

Under (F.ii), (F.iii) the regression function $m_\Psi(\mathbf{x})$ is defined as

$$m_\Psi(\mathbf{x}) = \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \int_{\mathbb{R}^q} \Psi(\mathbf{y}) f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \frac{r_\Psi(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})},$$

where

$$r_\Psi(\mathbf{x}) = \int_{\mathbb{R}^q} \Psi(\mathbf{y}) f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

The conditional variance $\sigma^2_{\mathbf{\Psi}}(x)$ of $\Psi(\mathbf{Y})$ given that $\mathbf{X} = \mathbf{x} \in \mathbb{X}$, is defined to be

$$\sigma^2_\Psi(x) = \operatorname{Var}(\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}) = \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \int_{\mathbb{R}^q} \{\Psi(\mathbf{y}) - m_\Psi(\mathbf{x})\}^2 f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

We fix a subset $\mathbb{X} \subset \mathbb{R}^d$ on wich we are considering the uniform convergence of the kernel regression estimator. We first characterize the intrinsic dimension of the distribution $\mathbb{P}$, proposed by Kim *et al.* (2018), by its rate of the probability volume growth on balls. According to the remark given in the first chapter, that is, if a probability distribution has a positive measure on a manifold with positive reach, 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 $d$.

**Lemma 2.1.0.1 (Kim *et al.* (2018))** *Let $\mathbb{P}$ be a probability distribution on $\mathbb{R}^d$, and $d_{\mathrm{vol}}$ be its volume dimension. Then for any $\nu \in [0, d_{\mathrm{vol}})$, there exists a constant $C_{\nu,\mathbb{P}}$ depending only on $\mathbb{P}$ and $\nu$ such that for all $\mathbf{x} \in \mathbb{X}$ and $r > 0$,*

$$\frac{\mathbb{P}(\mathbb{B}_{\mathbb{R}^d}(\mathbf{x}, r))}{r^\nu} \leq C_{\nu,\mathbb{P}}. \tag{2.1.1}$$

For the exact optimal rate, we impose conditions on how the probability volume decay in (2.1.1).

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

$$\limsup_{r\to 0} \sup_{\mathbf{x}\in\mathbb{X}} \frac{\mathbb{P}(\mathbb{B}_{\mathbb{R}^d}(\mathbf{x},r))}{r^{\nu}} < \infty. \tag{2.1.2}$$

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

$$\sup_{\mathbf{x}\in\mathbb{X}} \liminf_{r\to 0} \frac{\mathbb{P}(\mathbb{B}_{\mathbb{R}^d}(\mathbf{x},r))}{r^{\nu}} > 0. \tag{2.1.3}$$

These assumptions are in fact weak and hold for common probability distributions. In particular, Assumptions 1 and 2 hold when the probability distribution has a bounded density with respect to the $d$-dimensional Lebesgue measure. Let $K(\cdot)$ be a kernel function defined on $\mathbb{R}^d$, that is a measurable function such that

**(K.1)**

$$\int_{\mathbb{R}^d} K(\mathbf{x})d\mathbf{x} = 1.$$

From the Nadaraya-Watson Kernel estimator, we consider the kernel estimators of $f_{\mathbf{X}}(\mathbf{x})$, $r_{\Psi}(\mathbf{x})$, $m_{\Psi}(\mathbf{x})$, given respectively, for a bandwidth $h > 0$, by

$$f_{\mathbf{X};n}(\mathbf{x};h) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right), \tag{2.1.4}$$

$$r_{\Psi;n}(\mathbf{x};h) = \frac{1}{nh^d}\sum_{i=1}^{n} \Psi(\mathbf{Y}_i)K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right), \tag{2.1.5}$$

$$m_{\Psi;n}(\mathbf{x};h) = \begin{cases} \dfrac{r_{\Psi;n}(\mathbf{x};h)}{f_{\mathbf{X};n}(\mathbf{x};h)} = \dfrac{\sum_{i=1}^{n}\Psi(\mathbf{Y}_i)K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}, & \text{for } f_{\mathbf{X};n}(\mathbf{x};h)\neq 0,\\ \dfrac{1}{n}\sum_{i=1}^{n}\Psi(\mathbf{Y}_i), & \text{for } f_{\mathbf{X};n}(\mathbf{x};h)=0. \end{cases} \tag{2.1.6}$$

In this section, we follow Kim *et al.* (2018) for weakening the conditions on the kernel and making it adaptive to the intrinsic dimension of the underlying distribution and without assumptions on the distribution. It is worth noticing that for general distributions (such as the ones supporte don a lower-dimensional manifold), the usual change of variables argument is no longer directly applicable. However, we can provide a bound based on the volume dimension under an integrability condition on the kernel, given bellow.

**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*

$$\int_0^\infty t^{d_{\mathrm{vol}}-1} \sup_{\|\mathbf{x}\| \geq t} |K|^k(\mathbf{x}) dt < \infty. \tag{2.1.7}$$

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

$$\mathcal{K} := \left\{ (\mathbf{x}, h) \mapsto K\left(\frac{\mathbf{x} - \cdot}{h}\right) : \mathbf{x} \in \mathbb{X}, h \geq l_n \right\}$$

*is a uniformly bounded VC-class with dimension $\nu_2$, i.e., there exists positive number $A_2 \geq 1$ and $\nu_2 \geq 1$ such that, for every probability measure $\mathbb{Q}$ on $\mathbb{R}^d$ and for every $\varepsilon \in (0, \|K\|_\infty)$, the covering numbers $\mathcal{N}(\mathcal{K}, L_2(\mathbb{Q}), \varepsilon)$ satisfies*

$$\mathcal{N}(\mathcal{K}, L_2(\mathbb{Q}), \varepsilon) \leq \left(\frac{A_2 \|K\|_\infty}{\varepsilon}\right)^{\nu_2}.$$

*By combining Assumption 3 and Lemma 2.1.0.1, we can bound $\mathbb{E}_{\mathbb{P}}\left[K^2\right]$ in terms of the volume dimension $d_{\mathrm{vol}}$.*

***Lemma 2.1.0.2*** *Let $(\mathbb{R}^d, \mathbb{P})$ be a probability space and let $X \sim \mathbb{P}$. For any kernel $K(\cdot)$ satisfying Assumption 3 with $k > 0$, the expectation of the k-moment of the kernel is upper bounded as*

$$\mathbb{E}_{\mathbb{P}}\left[\left|K\left(\frac{\mathbf{x} - \mathbf{X}}{h}\right)\right|^k\right] \leq C_{k,\mathbb{P},K,\varepsilon} h^{d_{\mathrm{vol}}-\varepsilon} \tag{2.1.8}$$

*for any $\varepsilon \in (0, d_{\mathrm{vol}})$, where $C_{k,\mathbb{P},K,\varepsilon}$ is a constant depending only on $k, \mathbb{P}, K$ and $\varepsilon$. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, $\varepsilon$ can be 0 in (2.1.8).*

**Remark 2.1.0.3 (Kim *et al.* (2018))** *It is important to emphasize that Assumption 3 is weak, as it is satisfied by commonly used kernels. For instance, if the kernel function $K(\mathbf{x})$ decays at a polynomial rate strictly faster than $d_{\mathrm{vol}}/k$ (which is at most $d/k$) as $\mathbf{x} \to \infty$, that is, if*

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

*for any $\epsilon > 0$, the integrability condition (2.1.7) is satisfied. Also, if the kernel function $K(\mathbf{x})$ is spherically symmetric, that is, if there exists $\widetilde{K} : [0, \infty) \to \mathbb{R}$ with $K(\mathbf{x}) = \widetilde{K}(\|\mathbf{x}\|_2)$, then the integrability condition (2.1.7) is satisfied provided $\|K\|_k < \infty$. Kernels with bounded support also satisfy the condition (2.1.7). Thus, most of the commonly used kernels including Uniform, Epanechnikov, and Gaussian kernels satisfy the above integrability condition.*

## 2.2 The generalized empirical process

Compared to the univariate case, where the observations take their values in $\mathbb{R}$, the study of empirical processes becomes more delicate in $\mathbb{R}^d$, $d \geq 1$. The general theoretical framework allowing the study of i.i.d. random variables with values in a more general measurable space than $\mathbb{R}$ (such as, for example, $\mathbb{R}^d$) is based on the notion of generalized empirical process.
In this chapter, we establish a general theorems which, under weaker conditions on the kernel, will yield as special cases (given in chapter 3) the consistency of kernel-type estimators for density, regression, the conditional distribution, multivariate mode and Shannon's entrpoy and we will work in high- dimensional data sets particularly when the data generating distribution is supported on manifolds.
Recalling the definition of general empirical process

$$\begin{aligned} W_{n;h}(\mathbf{x};\Psi) &= \sum_{j=1}^{n}(c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_j)+d_\Psi(\mathbf{x}))K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right) \\ &\quad -n\mathbb{E}\left\{(c_\Psi(\mathbf{x})\Psi(\mathbf{Y})+d_\Psi(\mathbf{x}))K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right\}. \end{aligned}$$

For the future use, introduce the classes of continuous functions on a compact $\mathbf{J}=\mathbf{I}^\eta$, for some $0<\eta<1$, indexed by $\mathcal{F}_q$

$$\mathcal{F}_C = \{c_\Psi(\mathbf{x}) : \Psi \in \mathcal{F}_q\}, \qquad \mathcal{F}_D = \{d_\Psi(\mathbf{x}) : \Psi \in \mathcal{F}_q\}.$$

We shall always assume that the classes $\mathcal{F}_C$ and $\mathcal{F}_D$ are relatively compact with respect to the sup-norm topology on $\mathbf{J}=\mathbf{I}^\eta$, which by the Arzela-Ascoli theorem is equivalent to these classes being uniformly bounded and uniformly equicontinuous on $\mathbf{J}$. Let

$$C_{\mathcal{F}_q} := \sup\{\|c_\Psi\|_{\mathbf{J}} : \Psi \in \mathcal{F}_q\} \quad \text{and} \quad D_{\mathcal{F}_q} := \sup\{\|d_\Psi\|_{\mathbf{J}} : \Psi \in \mathcal{F}_q\}. \tag{2.2.1}$$

We can now state, in theorems 2.3.0.1 and 2.4.0.1 below, a technical result, which will be particularly useful in proving the other results of chapter 3.

## 2.3 Theorem in bounded case

To establish consistency of kernel-type estimators when the class of functions is bounded, we combine Talagrand's inequality (see appendix) and a VC type bound, following the approach of Sriperumbudur and Steinwart (2012). We apply Talagrand's inequality due to Bousquet (2002) and simplified in Steinwart and Christmann (2008).

**Theorem 2.3.0.1** *Let $\mathbb{P}$ be a probability distribution and let $K(\cdot)$ be a kernel function satisfying Assumption 3 and 4. Assume that $\mathcal{F}_q$ satisfy the above conditions and the classes of continuous functions $\mathcal{F}_C$ and $\mathcal{F}_D$ are as above, that is, relatively compact with respect to the sup-norm topology. Also assume that the envelope function $F$ of the class $\mathcal{F}_q$ satisfies*

$$\exists M > 0, \quad F(\mathbf{Y})\mathbb{1}\{\mathbf{x} \in \mathbf{J}\} \leq M, \quad a.s. \tag{H.i}$$

*Then, for any $\delta > 0$, we have with probability at least $1-\delta$,*

$$\begin{aligned}&\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\frac{1}{nh^d}\left|W_{n,h}(\mathbf{x},\Psi)\right| \\ \leq\ & C_1\left(\frac{(\log(1/l_n))_+}{nl_n^d}+\sqrt{\frac{(\log(1/l_n))_+}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\sqrt{\frac{\log(2/\delta)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\log(2/\delta)}{nl_n^d}\right)\end{aligned} \tag{2.3.1}$$

*for any $\varepsilon \in (0, d_{\mathrm{vol}})$, where $C_1$ is a constant depending only on $A$, $\|\vartheta\|_\infty$, $d$, $\nu_1$, $\nu_2$, $d_{\mathrm{vol}}$, $C_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon}$, $\varepsilon$.*
*Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, $\varepsilon$ can be 0 in (2.3.1).*

When $\delta$ is fixed and $l_n < 1$, the dominating terms in (2.3.1) are $\frac{\log(1/l_n)}{nl_n^d}$ and $\sqrt{\frac{(\log(1/l_n))}{nl_n^{2d-d_{\mathrm{vol}}}}}$. If $l_n$ does not vanish too rapidly, then the second term dominates the upper bound in (2.3.1) as in the following result that has been proven in Kim *et al.* (2018) for a density, where it is stated as Corrolary 13.

**Corollary 2.3.0.2** *Let $\mathbb{P}$ be a probability distribution and let $K(\cdot)$ be a kernel function satisfying Assumption 3 and 4. Fix $\varepsilon \in (0, d_{\mathrm{vol}})$. Further, if $d_{\mathrm{vol}} = 0$ or under Assumption 1, $\varepsilon$ can be 0. Suppose*

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

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

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\frac{1}{nh^d}\left|W_{n,h}(\mathbf{x},\Psi)\right| \leq C_2\sqrt{\frac{\log(1/l_n)_+ + \log(2/\delta)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}, \tag{2.3.2}$$

*where $C_2$ is a constant depending only on $A$, $\|\vartheta\|_\infty$, $d$, $\nu_1$, $\nu_2$, $d_{\mathrm{vol}}$, $C_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon}$, $\varepsilon$.*

## Proof of Theorem 2.3.0.1

We first note that under the Assumptions 3 and 4, and making use of Theorem 12 in Kim *et al.* (2018), we have with probability at least $1-\delta$

$$\begin{aligned}
&\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbb{X}}\frac{d_\Psi(\mathbf{x})}{nh^d}\left|\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)-n\mathbb{E}K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right|\\
&\leq D_{\mathcal{F}_q}\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbb{X}}\frac{1}{nh^d}\left|\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)-n\mathbb{E}K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right|\\
&\leq A\left(\frac{(\log(1/l_n))_+}{nl_n^d}+\sqrt{\frac{(\log(1/l_n))_+}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}\right.\\
&\qquad\left.+\sqrt{\frac{\log(2/\delta)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\log(2/\delta)}{nl_n^d}\right), \qquad (2.3.3)
\end{aligned}$$

where $\varepsilon\in(0,d_{\mathrm{vol}})$ and $A$ is a constant depending only on $C_{\mathcal{F}_q}$,$A_2$, $\|K\|_\infty$, $d$, $\nu_2$, $d_{\mathrm{vol}}$, $C_{k=2,\mathbb{P},K,\varepsilon}$, $\varepsilon$.

Consider the following class of functions,

$$\mathcal{G}:=\{\vartheta_{\Psi,h,\mathbf{x}}(\cdot,\cdot):\mathbf{x}\in\mathbb{X},h\geq l_n,\Psi\in\mathcal{F}_q\},$$

where, for $\mathbf{x}\in\mathbf{I}\subset\mathbb{X}$, $\Psi\in\mathcal{F}_q$ and $h\geq l_n$, $\vartheta_{\Psi,h,\mathbf{x}}(\mathbf{z},\mathbf{y}):\mathbb{R}^d\times\mathbb{R}^q\to\mathbb{R}$

$$\vartheta_{\Psi,h,\mathbf{x}}(\mathbf{z},\mathbf{y})=c_\Psi(\mathbf{x})\Psi(\mathbf{y})K\left(\frac{\mathbf{x}-\mathbf{z}}{h}\right). \qquad (2.3.4)$$

And we set

$$\widetilde{\eta}_{\Psi,h,\mathbf{x}}:=\frac{1}{nh^d}\sum_{i=1}^n c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_i)K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)=\frac{1}{nh^d}\sum_{i=1}^n\vartheta_{\Psi,h,\mathbf{x}}(\mathbf{X}_i,\mathbf{Y}_i).$$

Let us introduce

$$\widetilde{\mathcal{G}}:=\left\{\frac{1}{h^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right):\mathbf{x}\in\mathbf{I},h\geq l_n,\Psi\in\mathcal{F}_q\right\},$$

a class of normalized functions. Notice that we have

$$\begin{aligned}
\widetilde{\eta}_{\Psi,h,\mathbf{x}}-\mathbb{E}\widetilde{\eta}_{\Psi,h,\mathbf{x}} &= \frac{1}{nh^d}\sum_{i=1}^n c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_i)K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)-\mathbb{E}\left(\frac{1}{h^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right)\\
&= \frac{1}{n}\sum_{i=1}^n\frac{1}{h^d}\vartheta_{\Psi,h,\mathbf{x}}(\mathbf{X}_i,\mathbf{Y}_i)-\mathbb{E}\left(\frac{1}{h^d}\vartheta_{\Psi,h,\mathbf{x}}(\mathbf{X},\mathbf{Y})\right). \qquad (2.3.5)
\end{aligned}$$

It clearly suffices to show the following proposition.

**Proposition 2.3.0.3** *Under the assumptions of Theorem 2.3.0.1 and for all $A^{'} > 0$, we have with probability at least $1-\delta$,*

$$\begin{aligned}&\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\left|\widetilde{\eta}_{\Psi,h,\mathbf{x}}-\mathbb{E}\widetilde{\eta}_{\Psi,h,\mathbf{x}}\right|\\ \leq\ & A'\left(\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^d}+\sqrt{\frac{\left(\log\left(\frac{1}{l_n}\right)\right)_+}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}+\frac{\log\left(\frac{2}{\delta}\right)}{nl_n^d}\right)\end{aligned} \tag{2.3.6}$$

*for any $\varepsilon\in(0,d_{\mathrm{vol}})$, where the positive constant $A'$ depends only on $A_1, A_2, \|\vartheta\|_\infty, d$, $\nu_1,\nu_2,d_{\mathrm{vol}},C_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon},\varepsilon$.*

**Proof of Proposition 2.3.0.3**

Notice that for $g\in\mathcal{G}$, $\sup_{\mathbf{x}\in\mathbf{I}}\left|\widetilde{\eta}_{\Psi,h,\mathbf{x}}-\mathbb{E}\widetilde{\eta}_{\Psi,h,\mathbf{x}}\right|$ can be written as follows

$$\sup_{\mathbf{x}\in\mathbf{I}}\left|\widetilde{\eta}_{\Psi,h,\mathbf{x}}-\mathbb{E}\widetilde{\eta}_{\Psi,h,\mathbf{x}}\right|=\sup_{g\in\widetilde{\mathcal{G}}}\left|\frac{1}{n}\sum_{i=1}^n g(\mathbf{X}_i,\mathbf{Y}_i)-\mathbb{E}\left(g(\mathbf{X},\mathbf{Y})\right)\right|. \tag{2.3.7}$$

Now, it is immediate to check that

$$\|g\|_\infty\leq l_n^{-d}\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty\leq l_n^{-d}C_{\mathcal{F}_q}M\,\|K\|_\infty. \tag{2.3.8}$$

In order to bound the VC dimension of $\widetilde{\mathcal{G}}$, we consider

$$\widetilde{\mathcal{G}}:=\left\{\vartheta_{\Psi,h,\mathbf{x}}:\mathbf{x}\in\mathbf{I},h\geq l_n,\Psi\in\mathcal{F}_q\right\},$$

be a class of unnormalized functions. Fix $\eta<l_n^{-d}\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty$ and a probability measure $\mathbb{Q}$ on $\mathbb{R}^d$. Suppose

$$\left[l_n,\left(\frac{\eta}{2\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty}\right)^{-1/d}\right]$$

is covered by balls

$$\left\{\left(h_i-\frac{\eta l_n^{d+1}}{3d\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty},h_i+\frac{\eta l_n^{d+1}}{3d\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty}\right),1\leq i\leq N_1\right\},$$

and $(\mathcal{G},\mathbb{L}_2(\mathbb{Q}))$ is covered by

$$\left\{\mathbb{B}_{\mathbb{L}_2(\mathbb{Q})}\left(K_j,\frac{l_n^d\eta}{3C_{\mathcal{F}_q}M}\right)\cup\mathbb{B}_{\mathbb{L}_2(\mathbb{Q})}\left(\Psi_k,\widetilde{\varepsilon}\right),1\leq j\leq N_2,1\leq k\leq N_3\right\},$$

where

$$\widetilde{\varepsilon} \leq \frac{l_n^d \eta}{3C_{\mathcal{F}_q}\left\|K\right\|_\infty}.$$

For $1 \leq i \leq N_1, 1 \leq j \leq N_2$ and $1 \leq k \leq N_3$, we let

$$g_{i,j,k} = \frac{1}{h_i^d} g_{j,k} = \frac{1}{h_i^d} c_\Psi(\mathbf{x})\Psi_k(\mathbf{Y})K_j\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right).$$

Also, choose $h_0 > \left(\frac{\eta}{2\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty}\right)^{-1/d}$, $\mathbf{x}_0 \in \mathbf{I}$, $\Psi_0 \in \mathcal{F}_q$ and let

$$g_0 = \frac{1}{h_0}\vartheta_{\Psi_0,h_0,\mathbf{x}_0}.$$

We will show that

$$\left\{\mathbb{B}_{\mathbb{L}_2(\mathbb{Q})}\left(g_{i,j,k},\eta\right) : 1 \leq i \leq N_1, 1 \leq j \leq N_2 \quad \text{and} \quad 1 \leq k \leq N_3\right\} \cup \left\{\mathbb{B}_{\mathbb{L}_2(\mathbb{Q})}\left(g_0,\eta\right)\right\} \tag{2.3.9}$$

covers $\widetilde{\mathcal{G}}$.

For the first case when $h \leq \left(\frac{\eta}{2\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty}\right)^{-1/d}$, find $h_i$, $K_j$ and $\Psi_k$ with

$$\begin{aligned} h &\in \left(h_i - \frac{\eta l_n^{d+1}}{3d\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty}, h_i + \frac{\eta l_n^{d+1}}{3d\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_\infty}\right), \\ \Psi &\in \mathbb{B}_{\mathbb{L}_2(\mathbb{Q})}\left(\Psi_k, \widetilde{\varepsilon}\right), \\ K &\in \mathbb{B}_{\mathbb{L}_2(\mathbb{Q})}\left(K_j, \frac{l_n^d \eta}{3C_{\mathcal{F}_q}M}\right). \end{aligned}$$

Then the distance between $\frac{1}{h^d}\vartheta_{\Psi,h,\mathbf{x}}$ and $\frac{1}{h_i^d}g_{i,k}$ is upper bounded as follows

$$\begin{aligned} &\left\|\frac{1}{h^d}\vartheta_{\Psi,h,\mathbf{x}}(\mathbf{X},\mathbf{Y}) - \frac{1}{h_i^d}g_{i,k}(\mathbf{X},\mathbf{Y})\right\|_{\mathbb{L}_2(\mathbb{Q})} \\ &= \left\|\frac{1}{h^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right) - \frac{1}{h_i^d}g_{i,k}(\mathbf{X},\mathbf{Y})\right\|_{\mathbb{L}_2(\mathbb{Q})} \\ &\leq \left\|\frac{1}{h^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right) - \frac{1}{h_i^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right\|_{\mathbb{L}_2(\mathbb{Q})} \\ &\quad + \left\|\frac{1}{h_i^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right) - \frac{1}{h_i^d}g_{i,k}(\mathbf{X},\mathbf{Y})\right\|_{\mathbb{L}_2(\mathbb{Q})} \\ &\leq \left\|\frac{1}{h^d}\vartheta_{\Psi,h,\mathbf{x}} - \frac{1}{h_i^d}\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\mathbb{L}_2(\mathbb{Q})} + \left\|\frac{1}{h_i^d}\vartheta_{\Psi,h,\mathbf{x}} - \frac{1}{h_i^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K_j\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right\|_{\mathbb{L}_2(\mathbb{Q})} \\ &\quad + \left\|\frac{1}{h_i^d}c_\Psi(\mathbf{x})\Psi(\mathbf{Y})K_j\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right) - \frac{1}{h_i^d}c_\Psi(\mathbf{x})\Psi_k(\mathbf{Y})K_j\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right\|_{\mathbb{L}_2(\mathbb{Q})}. \end{aligned} \tag{2.3.10}$$

Now the first term of (3.6.3) is upper bounded as

$$\left\|\frac{1}{h^d}\vartheta_{\Psi,h,\mathbf{x}} - \frac{1}{h_i^d}\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\mathbb{L}_2(\mathbb{Q})} = |\frac{1}{h^d} - \frac{1}{h_i^d}|\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\mathbb{L}_2(\mathbb{Q})}$$

$$
\begin{aligned}
&= |h_i - h| \sum_{k=0}^{d-1} h_i^{k-d} h^{-1-k} \left\| \vartheta_{\Psi,h,\mathbf{x}} \right\|_{\mathbb{L}_2(\mathbb{Q})} \\
&\leq |h_i - h| d l_n^{-d-1} \left\| \vartheta_{\Psi,h,\mathbf{x}} \right\|_\infty < \frac{\eta}{3}. \qquad (2.3.11)
\end{aligned}
$$

Also, the second term of (3.6.3) is upper bounded as

$$
\begin{aligned}
&\left\| \frac{1}{h_i^d} \vartheta_{\Psi,h,\mathbf{x}} - \frac{1}{h_i^d} c_\Psi(\mathbf{x}) \Psi(\mathbf{Y}) K_j \left( \frac{\mathbf{x} - \mathbf{X}}{h} \right) \right\|_{\mathbb{L}_2(\mathbb{Q})} \\
&\leq \frac{1}{h_i^d} C_{\mathcal{F}_q} M \left\| K \left( \frac{\mathbf{x} - \mathbf{X}}{h} \right) - K_j \left( \frac{\mathbf{x} - \mathbf{X}}{h} \right) \right\|_{\mathbb{L}_2(\mathbb{Q})} < \frac{\eta}{3}. \qquad (2.3.12)
\end{aligned}
$$

And the last term of (3.6.3) is upper bounded as

$$
\begin{aligned}
&\left\| \frac{1}{h_i^d} c_\Psi(\mathbf{x}) \Psi(\mathbf{Y}) K_j \left( \frac{\mathbf{x} - \mathbf{X}}{h} \right) - \frac{1}{h_i^d} c_\Psi(\mathbf{x}) \Psi_k(\mathbf{Y}) K_j \left( \frac{\mathbf{x} - \mathbf{X}}{h} \right) \right\|_{\mathbb{L}_2(\mathbb{Q})} \\
&\qquad \leq l_n^{-d} C_{\mathcal{F}_q} \|K\|_\infty \left\| \Psi(\mathbf{Y}) - \Psi_k(\mathbf{Y}) \right\|_{\mathbb{L}_2(\mathbb{Q})} \\
&\qquad < l_n^{-d} C_{\mathcal{F}_q} \|K\|_\infty \widetilde{\varepsilon} \leq \frac{\eta}{3}. \qquad (2.3.13)
\end{aligned}
$$

Combining (3.6.4), (3.6.6) and (3.6.7) to (3.6.3), we arrive at the following bound

$$
\left\| \frac{1}{h^d} \vartheta_{\Psi,h,\mathbf{x}}(\mathbf{X}, \mathbf{Y}) - \frac{1}{h_i^d} g_{i,k}(\mathbf{X}, \mathbf{Y}) \right\|_{\mathbb{L}_2(\mathbb{Q})} < \eta.
$$

For the second case when $h > \left( \frac{\eta}{2 \left\| \vartheta_{\Psi,h,\mathbf{x}} \right\|_\infty} \right)^{-1/d}$, we have

$$
\left\| \frac{1}{h^d} \vartheta_{\Psi,h,\mathbf{x}} \right\|_{\mathbb{L}_2(\mathbb{Q})} \leq \left\| \frac{1}{h^d} \vartheta_{\Psi,h,\mathbf{x}} \right\|_\infty < \frac{\eta}{2},
$$

holds, and hence

$$
\left\| \frac{1}{h^d} \vartheta_{\Psi,h,\mathbf{x}} - g_0 \right\|_{\mathbb{L}_2(\mathbb{Q})} \leq \left\| \frac{1}{h^d} \vartheta_{\Psi,h,\mathbf{x}} \right\|_{\mathbb{L}_2(\mathbb{Q})} + \|g_0\|_{\mathbb{L}_2(\mathbb{Q})} < \eta.
$$

Therefore (2.3.9) is shown. Hence combined with (F.iii), (2.2.1) and Assumption 4 and due to Lemma 9.9, p.160 of Kosorok (2008), gives that every probability measure $\mathbb{Q}$ on $\mathbb{R}^d$ and for every $\eta \in \left(0, h^{-d} \left\| \vartheta_{\Psi,h,\mathbf{x}} \right\|_\infty\right)$, the covering number $\mathcal{N}(\widetilde{\mathcal{G}}, L_2(\mathbb{Q}), \eta)$ is upper bounded as

$$
\begin{aligned}
&\sup_{\mathbb{Q}} \mathcal{N}(\widetilde{\mathcal{G}}, L_2(\mathbb{Q}), \eta) \\
&\leq \mathcal{N} \left( \left[ l_n, \left( \frac{\eta}{2 \left\| \vartheta_{\Psi,h,\mathbf{x}} \right\|_\infty} \right)^{-1/d} \right], |\cdot|, \frac{\eta l_n^{d+1}}{3d \left\| \vartheta_{\Psi,h,\mathbf{x}} \right\|_\infty} \right) \sup_{\mathbb{Q}} \mathcal{N} \left( c_\Psi(\mathbf{x}) \mathcal{F}_q, L_2(\mathbb{Q}), C_{\mathcal{F}_q} \widetilde{\varepsilon} \right) \\
&\qquad \sup_{\mathbb{Q}} \mathcal{N} \left( \mathcal{K}, L_2(\mathbb{Q}), \frac{l_n^d \eta}{3 C_{\mathcal{F}_q} M} \right) + 1
\end{aligned}
$$

$$\begin{aligned}
&\leq \frac{3d\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\infty}}{\eta l_n^{d+1}}\left(\frac{2\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\infty}}{\eta}\right)^{1/d}\left(\frac{3A_1M\|K\|_{\infty}}{l_n^d\eta}\right)^{2\nu_1-1}\left(\frac{3A_2C_{\mathcal{F}_q}M\left\|K\right\|_{\infty}}{l_n^d\eta}\right)^{\nu_2}+1\\
&\leq \left(\frac{3Ad\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\infty}}{l_n^d\eta}\right)^{2\nu_1+\nu_2-1}\left[\left(\frac{3d^{2-2\nu_1-\nu_2}\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\infty}}{\eta l_n^{d+1}}\right)\left(\frac{2\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\infty}}{\eta}\right)^{1/d}\right.\\
&\quad\left.+\left(\frac{3Ad\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\infty}}{l_n^d\eta}\right)^{-2\nu_1-\nu_2+1}\right]\\
&\leq \left(\frac{3Ad\left\|\vartheta_{\Psi,h,\mathbf{x}}\right\|_{\infty}}{l_n^d\eta}\right)^{2\nu_1+\nu_2-1}, \qquad (2.3.14)
\end{aligned}$$

for some finite constant $0 < A < \infty$. For $p > 2$, note that assumption (H.i) implies that

$$\sup_{\mathbf{x}\in\mathbf{I}}\mathbb{E}(F^p(\mathbf{Y})\mid\mathbf{X}=\mathbf{x}) < M^p < \infty.$$

From Lemma 2.1.0.2 and using a conditioning argument, we observe that, for $p \geq k$

$$\begin{aligned}
\mathbb{E}\left|c_{\Psi}(\mathbf{x})^k\Psi^k(\mathbf{Y})K^k\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right| &\leq C_{\mathcal{F}_q}^k\mathbb{E}\left|\Psi^k(\mathbf{Y})K^k\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right|\\
&\leq C_{\mathcal{F}_q}^k\mathbb{E}\left|K^k\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\mathbb{E}\left(\Psi^k(\mathbf{Y})\mid\mathbf{X}=\mathbf{x}\right)\right|,
\end{aligned}$$

where

$$\mathbb{E}(\Psi^k(\mathbf{Y})\mid\mathbf{X}=\mathbf{x}) \leq \sup_{\mathbf{x}\in\mathbf{J}}\mathbb{E}(F^k(\mathbf{Y})\mid\mathbf{X}=\mathbf{x}).$$

Then, for $k = 2$, we have

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{h^d}c_{\Psi}(\mathbf{x})^2\Psi^2(\mathbf{Y})K^2\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right| &\leq M^pC_{\mathcal{F}_q}^2\left|\mathbb{E}\frac{1}{h^d}K^2\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right|\\
&\leq M^pC_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon}^2 l_n^{-2d+d_{\mathrm{vol}}-\varepsilon}. \qquad (2.3.15)
\end{aligned}$$

Now from (2.3.8), (2.3.14), and (2.3.15), applying Theorem A.1.0.1 to (2.3.7) gives that $\|\widetilde{\eta}_{\Psi,h,\mathbf{x}}-\mathbb{E}\widetilde{\eta}_{\Psi,h,\mathbf{x}}\|_{\infty}$ is upper bounded with probability at least $1-\delta$ as

$$\begin{aligned}
&\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\left|\widetilde{\eta}_{\Psi,h,\mathbf{x}}-\mathbb{E}\widetilde{\eta}_{\Psi,h,\mathbf{x}}\right|\\
&\leq C\left(\frac{2(2\nu_1+\nu_2-1)\left\|\vartheta\right\|_{\infty}\log\left(\dfrac{2Ad\left\|\vartheta\right\|_{\infty}}{M^{\frac{p}{2}}C_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon}l_n^{(d_{\mathrm{vol}}-\epsilon)/2}}\right)}{nl_n^d}\right.\\
&\quad+\sqrt{\frac{2(2\nu_1+\nu_2-1)C_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon}^2\log\left(\dfrac{2Ad\left\|\vartheta\right\|_{\infty}}{M^{\frac{p}{2}}C_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon}l_n^{(d_{\mathrm{vol}}-\epsilon)/2}}\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}
\end{aligned}$$

$$
\left. + \sqrt{\frac{C^2_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon} \log\left(\frac{1}{\delta}\right)}{n l_n^{2d-d_{\mathrm{vol}}+\varepsilon}}} + \frac{\|\vartheta\|_\infty \log\left(\frac{1}{\delta}\right)}{n l_n^d} \right)
$$

$$
\leq C_{A,\|\vartheta\|_\infty,d,\nu_1,\nu_2,d_{\mathrm{vol}},C_{k=2,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon},\varepsilon}
$$

$$
\times \left( \frac{\left(\log\left(\dfrac{1}{l_n}\right)\right)_+}{n l_n^d} + \sqrt{\frac{\left(\log\left(\dfrac{1}{l_n}\right)\right)_+}{n l_n^{2d-d_{\mathrm{vol}}+\varepsilon}}} + \sqrt{\frac{\log\left(\dfrac{2}{\delta}\right)}{n l_n^{2d-d_{\mathrm{vol}}+\varepsilon}}} + \frac{\log\left(\frac{2}{\delta}\right)}{n l_n^d} \right).
$$

Theorem 2.3.0.1 now follows from (2.3.3) and Proposition 2.3.0.3. □

## 2.4 Theorem in unbounded case

In this section, we will treat the unbounded case, where (H.i) is replaced by (H.ii). Our proofs are based on an extension of the methods developed in Einmahl *et al.* (2005). We will use the same idea with suitable modifications, from the proof of Theorem 02 of Einmahl *et al.* (2005), namely, combining an exponential inequality of Talagrand with a suitable moment inequality.

To prove Theorem 2.3.0.1 in this case,we assume, for some $p > 2$, that

$$
\sup_{\mathbf{x}\in\mathbf{J}} \mathbb{E}(F^p(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}) < \infty.
$$

**Theorem 2.4.0.1** *Assume that $\mathcal{F}_q$ and $\mathcal{K}$ satisfies the above conditions, and the kernel $K(\cdot)$ satisfying assumption 3. Further assume that the classes of continuous functions $\mathcal{F}_C$ and $\mathcal{F}_D$ are as above, that is, relatively compact with respect to the sup-norm topology. Assume that $\mathcal{F}_q$ satisfy the above conditions and further assume that the envelope function $F$ of the class $\mathcal{F}_q$ satisfies for some $p > 2$*

$$
\beta_{\mathbb{P}}(F) := \sup_{\mathbf{x}\in\mathbf{J}} \mathbb{E}(F^p(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}) < \infty. \tag{H.ii}
$$

*Then we have for any $c > 0$ and $0 < h_0 < (2\eta)^{d_{\mathrm{vol}}-\varepsilon}$, with probability $1$,*

$$
\limsup_{n\to\infty} \sup_{c(\log n/n)^\gamma \leq h \leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{1}{nh^d} |W_{n,h}(\mathbf{x}, \Psi)| =: \mathfrak{K}(c) \sqrt{\frac{(\log(1/h) \vee \log\log n)}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}, \tag{2.4.1}
$$

*for any $\varepsilon \in (0, d_{\mathrm{vol}})$, and $\gamma = 1 - 2/p$.*

**Remark 2.4.0.2** *Note that the condition (H.ii) may be replaced by more general hypotheses upon moments of* $\mathbf{Y}$ *as in Deheuvels (2011). That is*

**(M.1)″** *We denote by* $\{\mathcal{M}(x) : x \geq 0\}$ *a nonnegative continuous function, increasing on* $[0, \infty)$*, and such that, for some* $s > 2$*, ultimately as* $x \uparrow \infty$*,*

$$(i)\ x^{-s}\mathcal{M}(x) \downarrow; (ii)\ \ x^{-1}\mathcal{M}(x) \uparrow . \tag{2.4.2}$$

*For each* $t \geq \mathcal{M}(0)$*, we define* $\mathcal{M}^{inv}(t) \geq 0$ *by* $\mathcal{M}(\mathcal{M}^{inv}(t)) = t$*. We assume further that:*

$$\mathbb{E}\left(\mathcal{M}\left(|F(\mathbf{Y})|\right)\right) < \infty.$$

*The introduction of the function* $\Psi(\cdot)$ *in our setting is motivated by Remark 1.2 of Deheuvels and Mason (2004) or Remark 1.1 of Deheuvels (2011).*

## Proof of Theorem 2.4.0.1 :

We shall prove Theorem 2.4.0.1 under assumption (H.ii). We first note that

$$\limsup_{n\to\infty} \sup_{c(\log n/n)^{\gamma}\leq h\leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{\sqrt{nh^{2d-d_{\text{vol}}+\varepsilon}}\left|\sum_{i=1}^{n} d_{\Psi}(\mathbf{x})\left\{K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) - \mathbb{E}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)\right\}\right|}{nh^d\sqrt{(\log(1/h)\vee\log\log n)}} \tag{2.4.3}$$

$$\leq D_{\mathcal{F}_q} \limsup_{n\to\infty} \sup_{c(\log n/n)^{\gamma}\leq h\leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \frac{\sqrt{nh^{2d-d_{\text{vol}}+\varepsilon}}\left|\sum_{i=1}^{n} \left\{K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) - \mathbb{E}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)\right\}\right|}{nh^d\sqrt{(\log(1/h)\vee\log\log n)}}. \tag{2.4.4}$$

In view of Theorem 1 in Einmahl *et al.* (2005) by an obvious modification of the proof, it is easy to see that this quantity is finite with probability 1.

**Proposition 2.4.0.3** *Under the assumptions of Theorem 2.4.0.1, for all* $c > 0$*, there exists a* $\mathfrak{Q}_0(c) > 0$ *such that with probability 1,*

$$\limsup_{n\to\infty} \sup_{c(\log n/n)^{\gamma}\leq h\leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} |\tilde{\eta}_{\Psi,h,\mathbf{x}} - \mathbb{E}\tilde{\eta}_{\Psi,h,\mathbf{x}}| =: \mathfrak{Q}_0(c)\sqrt{\frac{(\log(1/h)\vee\log\log n)}{nh^{2d-d_{\text{vol}}+\varepsilon}}}. \tag{2.4.5}$$

*where* $\mathfrak{Q}_0(c) < \infty$.

**Proof of Proposition 2.4.0.3**

For $\mathbf{x} \in \mathbf{I}$, $a_k = c\left(\frac{\log n_k}{n_k}\right)^\gamma \leq h \leq h_0$ and $\Psi \in \mathcal{F}_q$, let

$$\vartheta^{(k)}_{\Psi,h,\mathbf{x}}(\mathbf{u},\mathbf{v}) = c_\Psi(\mathbf{x})\Psi_k(\mathbf{v})K\left(\frac{\mathbf{x}-\mathbf{u}}{h}\right), \tag{2.4.6}$$

and

$$\vartheta_{\Psi,h,\mathbf{x}}(\mathbf{u},\mathbf{v}) = c_\Psi(\mathbf{x})\Psi(\mathbf{v})K\left(\frac{\mathbf{x}-\mathbf{u}}{h}\right). \tag{2.4.7}$$

Let $\alpha_n$ be the empirical process based on the sample $(\mathbf{X}_1,\mathbf{Y}_1),\ldots,(\mathbf{X}_n,\mathbf{Y}_n)$, i.e., if $\mathbf{g}:\mathbb{R}^d\times\mathbb{R}^q\rightarrow\mathbb{R}$, we have

$$\alpha_n(\mathbf{g}) = \sum_{i=1}^n \left(\mathbf{g}(\mathbf{X}_i,\mathbf{Y}_i) - \mathbb{E}\mathbf{g}(\mathbf{X},\mathbf{Y})\right)/\sqrt{n}. \tag{2.4.8}$$

To prove our uniform in bandwidth results, we shall apply Talagrand (1994) exponential inequality for the empirical process combined with a moment bound due to Einmahl *et al.* (2005). See, respectively, Proposition A.2.0.1 and Talagrand's inequality in the Appendix of this manuscript.
Let $\gamma = 1-p/2$, $\Psi\in\mathcal{F}_q$ and $n_k = 2^k$, $k\geq 1$. Set for $j\geq 0$ and $c>0$

$$h_{j,k} = 2^j a_k = 2^j c\left(\frac{\log n_k}{n_k}\right)^\gamma$$

and

$$\Psi_k(\mathbf{y}) = \Psi(\mathbf{y})\mathbb{1}\{F(\mathbf{y}) < (n_k/k)^{1/p}\}. \tag{2.4.9}$$

The proof of Proposition 2.4.0.3 will be divided into two parts which are a consequence of two lemmas. We begin with a truncation argument. Set for $n_{k-1}\leq n\leq n_k$, $\mathbf{x}\in\mathbf{I}$, $a_k\leq h\leq h_0$ and $\Psi\in\mathcal{F}_q$

$$\begin{aligned}
\tilde{\eta}_{\Psi,h,n}(\mathbf{x}) &:= \frac{1}{nh^d}\sum_{i=1}^n c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_i)K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)\\
&:= \frac{1}{nh^d}\sum_{i=1}^n c_\Psi(\mathbf{x})\Psi_k(\mathbf{Y}_i)K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) + \frac{1}{nh^d}\sum_{i=1}^n c_\Psi(\mathbf{x})\overline{\Psi}_k(\mathbf{Y}_i)K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)\\
&=: \frac{1}{nh^d}\sum_{i=1}^n c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_i)\mathbb{1}\{F(\mathbf{Y}_i)<(n_k/k)^{1/p}\}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)\\
&\quad+\frac{1}{nh^d}\sum_{i=1}^n c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_i)\mathbb{1}\{F(\mathbf{Y}_i)\geq(n_k/k)^{1/p}\}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)\\
&=: \tilde{\eta}_{\Psi,h,n,k}(\mathbf{x}) + \overline{\tilde{\eta}}_{\Psi,h,n,k}(\mathbf{x}).
\end{aligned}$$

**Lemma 2.4.0.4** *There exists a constant $\mathfrak{Q}_1(c) < \infty$, such that with probability 1,*

$$\limsup_{k\to\infty} \max_{n_{k-1}\le n\le n_{k-1}} \sup_{a_k\le h\le h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} |\tilde{\eta}_{\Psi,h,n,k} - \mathbb{E}\tilde{\eta}_{\Psi,h,n,k}| =: \mathfrak{Q}_1(c)\sqrt{\frac{(\log(1/h)\vee\log\log n)}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{2.4.10}$$

**Proof of Lemma 2.4.0.4**

Notice that

$$\frac{1}{nd^d}(\tilde{\eta}_{\Psi,h,n,k} - \mathbb{E}\tilde{\eta}_{\Psi,h,n,k}) = \sqrt{n}\alpha_n(\vartheta_{\Psi,h,\mathbf{x},k}), \tag{2.4.11}$$

where $\alpha_n$ is the empirical process based on $(\mathbf{X}_1,\mathbf{Y}_1),\ldots,(\mathbf{X}_n,\mathbf{Y}_n)$ defined by (2.4.8). For $k\ge 1$, let

$$\mathcal{G}_k(h) := \{\vartheta_{\Psi,h,\mathbf{x},k} : \Psi\in\mathcal{F}_q, \mathbf{x}\in\mathbf{I}\}.$$

Remark that we have

$$\begin{aligned}
&\max_{n_{k-1}\le n\le n_{k-1}} \sup_{a_k\le h\le h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}\,|\tilde{\eta}_{\Psi,h,n,k} - \mathbb{E}\tilde{\eta}_{\Psi,h,n,k}|}{\sqrt{(\log(1/h)\vee\log\log n)}}\\
&= \max_{n_{k-1}\le n\le n_{k-1}} \sup_{a_k\le h\le h_0} \frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}\,\|\sqrt{n}\alpha_n\|_{\mathcal{G}_k(h)}}{\sqrt{(\log(1/h)\vee\log\log n)}}.
\end{aligned}$$

Note that for each $\vartheta_{\Psi,h,\mathbf{x},k}\in\mathcal{G}_k(h)$

$$\|\vartheta_{\Psi,h,\mathbf{x},k}\|_\infty \le \|K\|_\infty C_{\mathcal{F}_q}(n_k/k)^{1/p} =: B_0(n_k/k)^{1/p}. \tag{2.4.12}$$

By Assumption 3 and from Lemma 2.1.0.2, we get that

$$\begin{aligned}
\mathbb{E}\left[(\vartheta_{\Psi,h,\mathbf{x},k})^2(\mathbf{X},\mathbf{Y})\right] \le \mathbb{E}\left[\vartheta^2_{\Psi,h,\mathbf{x}}(\mathbf{X},\mathbf{Y})\right] &\le C^2_{\mathcal{F}_q}\beta_{\mathbb{P}}^{2/p}\mathbb{E}K^2\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\\
&\le C^2_{\mathcal{F}_q}\beta_{\mathbb{P}}^{2/p}C_{\mathbb{P},K,\varepsilon}h^{d_{\mathrm{vol}}-\varepsilon}\\
&=: h^{d_{\mathrm{vol}}-\varepsilon}B_1.
\end{aligned}$$

Thus

$$\sup_{\vartheta\in\mathcal{G}_k(h)} \mathbb{E}\vartheta^2_{\Psi,h,\mathbf{x}}(\mathbf{X},\mathbf{Y}) \le h^{d_{\mathrm{vol}}-\varepsilon}B_1. \tag{2.4.13}$$

Recalling that for $j,k\ge 0$,

$$h_{j,k} = 2^j a_k = 2^j c\left(\frac{\log n_k}{n_k}\right)^\gamma,$$

and set

$$\mathcal{G}_{j,k}(h) := \{\vartheta_{\Psi,h,\mathbf{x},k} : \Psi\in\mathcal{F}_q, \mathbf{x}\in\mathbf{I} \text{ and } h_{j,k}\le h\le h_{j+1,k}\}.$$

Clearly by (2.4.13), for all $h_{j,k} \leq h \leq h_{j+1,k}$

$$\mathbb{E}\left[(\vartheta_{\Psi,h,\mathbf{x},k})^2(\mathbf{X},\mathbf{Y})\right] \leq h^{d_{\text{vol}}-\varepsilon}B_1 \leq 2B_1 h_{j,k}^{d_{\text{vol}}-\varepsilon} =: \sigma_{j,k}^2.$$

We shall use Corollary A.2.0.2 in the Appendix to bound

$$\mathbb{E}\left\|\sum_{i=1}^{n_k}\varepsilon_i\vartheta(\mathbf{X}_i,\mathbf{Y}_i)\right\|_{\mathcal{G}_{j,k}}.$$

Note first that by arguing as in the proof of Lemma 5 of Einmahl and Mason (2000), each $\mathcal{G}_{j,k} \subset \mathcal{G}$, where $\mathcal{G}$ is pointwise measurable class of functions with envelope function

$$G(\mathbf{x},\mathbf{y}) = C_{\mathcal{F}_q}\|K\|_\infty F(\mathbf{y}),$$

and satisfies the uniform entropy conditions. Further, we note that each $\mathcal{G}_{j,k}$ satisfies **(i)** and **(v)** of the corollary A.2.0.2 with

$$\beta^2 = \beta_{\mathbb{P}}^{2/p}, \quad \sigma^2 = \sigma_{j,k}^2 \text{ and } U = B_0(n_k/k)^{1/p},$$

we obtain after a small calculation that for a suitable positive constants $B_2$ and $B_3$ and for $k \geq 1$ $j \geq 0$

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{n_k}\varepsilon_i\vartheta(\mathbf{X}_i,\mathbf{Y}_i)\right\|_{\mathcal{G}_{j,k}} &\leq B_3\sqrt{n_k h_{j,k}^{d_{\text{vol}}-\varepsilon}\log((B_2 h_{j,k}^{d_{\text{vol}}-\varepsilon})^{-1} \vee C_1)} \\ &\leq B_3 a_{j,k}, \qquad (2.4.14)\end{aligned}$$

where

$$a_{j,k} = \sqrt{n_k h_{j,k}^{d_{\text{vol}}-\varepsilon}\log((B_2 h_{j,k}^{d_{\text{vol}}-\varepsilon})^{-1} \vee \log\log n_k)}.$$

Applying Talagrand's inequality in the Appendix with

$$M = B_0(n_k/k)^{1/p} \quad \text{and} \quad \sigma^2 = \sigma_{j,k}^2 \leq 2B_1 h_{j,k}^{d_{\text{vol}}-\varepsilon},$$

we get for any $t > 0$ and large enough $k$

$$\begin{aligned}&\mathbb{P}\left\{\max_{n_{k-1}\leq n\leq n_k}\left\|\sqrt{n}\alpha_n\right\|_{\mathcal{G}_{j,k}} \geq A_1\left(B_3 a_{j,k}+t\right)\right\} \\ &\leq \; 2\left\{\exp\left(-A_2 t^2/(2B_1 n_k h_{j,k}^{d_{\text{vol}}-\varepsilon})\right) + \exp\left(-A_2 t k^{1/p}/(B_0 n_k)^{1/p}\right)\right\}.\end{aligned}$$

Set for any $\rho > 1$, $j \geq 0$ and $k \geq 1$,

$$p_{j,k}(\rho) = \mathbb{P}\left\{\max_{n_{k-1}\leq n\leq n_k}\left\|\sqrt{n}\alpha_n\right\|_{\mathcal{G}_{j,k}} \geq A_1\left(D_3+\rho\right)a_{j,k}\right\}.$$

As we have

$$a_{j,k}/\sqrt{n_k h_{j,k}^{d_{\mathrm{vol}}-\varepsilon}} \geq \sqrt{\log\log n_k} \;\text{ and }\; h_{j,k} \geq c(\log n_k/n_k)^{1-2/p},$$

we readily obtain for large $k$ and $j \geq 0$ that

$$p_{j,k}(\rho) \leq 2\exp\left(-(\rho^2 A_2/B_1)\log\log n_k\right) + 2\exp\left((\sqrt{c}\rho A_2/B_0)\sqrt{\log n_k \log\log n_k}\right),$$

which for $\gamma = \frac{A_2}{B_1} \wedge \sqrt{c}A_2/B_0$ implies

$$p_{j,k}(\rho) \leq 4\exp(-\rho\gamma\log\log n_k).$$

Let for large enough $k$

$$l_k = \max\{j : h_{j,k} \leq 2h_0\}.$$

Then, we have $l_k \leq k$ for all large $k$ and large enough $\rho$ consequently,

$$P_k(\rho) := \sum_{j=0}^{l_k-1} p_{j,k}(\rho) \leq 4l_k(\log n_k)^{-\rho\gamma} \leq \frac{1}{k^2}.$$

Notice that by definition of $l_k$ for large $k$,

$$2h_{l_k,k} = h_{l_k+1,k} \geq 2h_0,$$

which implies that we have for $n_{k-1} \leq n \leq n_k$,

$$\left[\frac{c\log n}{n}, h_0\right] \subset \left[\frac{c\log n_k}{n_k}, h_{l_k,k}\right].$$

Thus for all large enough $k$,

$$\begin{aligned} A_k(\rho) := &\left\{\max_{n_{k-1}\leq n\leq n_{k-1}} \sup_{a_k\leq h\leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}\,|\tilde{\eta}_{\Psi,h,n,k} - \mathbb{E}\tilde{\eta}_{\Psi,h,n,k}|}{\sqrt{(\log(1/h^{d_{\mathrm{vol}}-\varepsilon}) \vee \log\log n}} > 2A_1(B_3+\rho)\right\} \\ &\subset \bigcup_{j=0}^{l_k-1}\left\{\max_{n_{k-1}\leq n\leq n_k} \left\|\sqrt{n}\alpha_n\right\|_{\mathcal{G}_{j,k}} \geq A_1\left(B_3+\rho\right)a_{j,k}\right\}. \end{aligned}$$

It follows for large enough $\rho$ that

$$\mathbb{P}\left(A_k(\rho)\right) \leq P_k(\rho) \leq \frac{1}{k^2},$$

wich by Borel-Cantelli lemma implies Lemma 2.4.0.4. □

**Lemma 2.4.0.5** *With probability 1,*

$$\lim_{k\to\infty} \max_{n_{k-1}\leq n\leq n_{k-1}} \sup_{a_k\leq h\leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}\left|\bar{\tilde{\eta}}_{\Psi,h,n,k} - \mathbb{E}\bar{\tilde{\eta}}_{\Psi,h,n,k}\right|}{\sqrt{(\log(1/h) \vee \log\log n)}} = 0. \quad (2.4.15)$$

### Proof of Lemma 2.4.0.5

First, note that for any $h \leq h_0$, $\Psi \in \mathcal{F}_q$ and $n_{k-1} \leq n \leq n_k$

$$\sup_{\mathbf{x}\in\mathbf{I}} \left| nh^d \mathbb{E}\bar{\tilde{\eta}}_{\Psi,h,n,k} \right| \leq \|K\|_\infty C_{\mathcal{F}_q} n_k \mathbb{E}\left[ F(\mathbf{Y})\mathbb{1}\{\mathbf{X} \in J, F(\mathbf{Y} \geq (n_k/k)^{1/p}\} \right].$$

We further have by (H.ii),

$$\mathbb{E}F^p(\mathbf{Y})\mathbb{1}\{\mathbf{X} \in J\} \leq \infty,$$

and we see that uniformly in $n_{k-1} \leq n \leq n_k$, $h \leq h_0$ and $\Psi \in \mathcal{F}_q$,

$$\sup_{\mathbf{x}\in\mathbf{I}} \left| nh^d \mathbb{E}\bar{\tilde{\eta}}_{\Psi,h,n,k} \right| = o(n_k^{1/p} k^{1-1/p}) = o\left( \sqrt{n_k a_k^{d_{\mathrm{vol}}-\varepsilon} \log(1/a_k^{d_{\mathrm{vol}}-\varepsilon})} \right)$$

as $k \to \infty$, where

$$a_k = c(\log n_k/n_k)^{1-2/p}.$$

By the monotonicity of the function $h \to h^{d_{\mathrm{vol}}-\varepsilon}\log(1/h^{d_{\mathrm{vol}}-\varepsilon})$, $h \leq 1/\exp$, we readily obtain that

$$\lim_{k\to\infty} \max_{n_{k-1}\leq n \leq n_{k-1}} \sup_{a_k \leq h \leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}} \left|\mathbb{E}\bar{\tilde{\eta}}_{\Psi,h,n,k}\right|}{\sqrt{(\log(1/h^{d_{\mathrm{vol}}-\varepsilon}) \vee \log\log n)}} = 0. \tag{2.4.16}$$

It remains to establish that, with probability 1,

$$\lim_{k\to\infty} \max_{n_{k-1}\leq n \leq n_{k-1}} \sup_{a_k \leq h \leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}} \left|\bar{\tilde{\eta}}_{\Psi,h,n,k}\right|}{\sqrt{(\log(1/h^{d_{\mathrm{vol}}-\varepsilon}) \vee \log\log n}} = 0. \tag{2.4.17}$$

Observe that we have

$$\begin{aligned} &\max_{n_{k-1}\leq n \leq n_{k-1}} \sup_{a_k \leq h \leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} nh^d \left|\bar{\tilde{\eta}}_{\Psi,h,n,k}\right| \\ \leq\ & \|K\|_\infty C_{\mathcal{F}_q} \sum_{i=1}^{n_k} F(\mathbf{Y}_i)\mathbb{1}\{\mathbf{X}_i \in J, F(\mathbf{Y}_i) > (n_k/k)^{1/p}\}. \end{aligned}$$

Inspecting the proof from Lemma 1 of Einmahl and Mason (2000), we see that the argument there also applies if we set $h_{n_k} = c(\log n_k/n_k)^{1-2/p}$ in equation (2.10) of Einmahl and Mason (2000) to give as $k \to \infty$, with probability 1,

$$\sum_{i=1}^{n_k} F(\mathbf{Y}_i)\mathbb{1}\{\mathbf{X}_i \in J, F(\mathbf{Y}_i) > (n_k/k)^{1/p}\} = o\left( \sqrt{n_k a_k^{d_{\mathrm{vol}}-\varepsilon} \log(1/a_k^{d_{\mathrm{vol}}-\varepsilon})} \right),$$

and we see by the same arguments as in (2.4.16) that (2.4.17) holds, thereby finishing the proof of Lemma 2.4.0.5. □

Proposition 2.4.0.3 now follows from Lemmas 2.4.0.4 and 2.4.0.5.

□

**Remark 2.4.0.6** *For notational convenience, we have chosen the same bandwidth sequence for each margins. This assumption can be dropped easily. If one wants to make use of the vector bandwidths (see, in particular, Chapter 12 of Devroye and Lugosi (2001)). With obvious changes of notation, our results and their proofs remain true when* $h_n$ *is replaced by a vector bandwidth* $\mathbf{h}_n = (h_n^{(1)}, \ldots, h_n^{(d)})$*, where* $\min h_n^{(i)} > 0$*. In this situation we set* $h_n = \prod_{i=1}^{d} h_n^{(i)}$*, and for any vector* $\mathbf{v} = (v_1, \ldots, v_d)$ *we replace* $\mathbf{v}/h_n$ *by* $(v_1/h_n^{(1)}, \ldots, v_1/h_n^{(d)})$*. For ease of presentation we chose to use real-valued bandwidths throughout.*

# Chapter 3

# Uniform in bandwidth consistency for nonparametric kernel-type estimators

## Introduction

In this chapter, we establish a set of uniform convergence results for some kernel estimators. To state the main contributions, we will make use of properties of the empirical process $W_{n;h}(\mathbf{x};\Psi)$, indexed by classes of functions, which will serve as our working basis. Particularly, we will show in sections 3.1 and 3.2 how Theorems 2.3.0.1 and 2.4.0.1 can be used to establish the uniform-in-bandwidth consistency for density and regression function estimators and their derivatives and we will see that the proofs will follow straightforwardly from these Theorems. In Section 3.3, we show how Theorem 2.3.0.1 can be used to establish the uniform in bandwidth consistency for the kernel estimator of the conditional distribution function.

The approach we use also allows us to deal with the uniform-in-bandwidth consistency of other estimators such as Kernel mode estimator and Shannon's entropy, introduced respectively in Sections 3.4 and 3.5.

An important result on the estimate of the additive regression function is given in section 3.6 using additive models introduced by Stone (1985), and we close this chapter with a discussion on the problem of bandwidth selection criterion.

## 3.1 Estimation of this density and regression functions

For this purpose recall that the definition of the estimator of the kernel density of $f_{\mathbf{X}}(\cdot)$ based upon the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is

$$f_{\mathbf{X};n}(\mathbf{x}; h) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right).$$

**Corollary 3.1.0.1** *Assume that $K(\cdot)$ satisfies Assumptions 3 and 4, we have for $C_3 > 0$, with probability at least $1 - \delta$,*

$$\sup_{h \geq l_n} \sup_{\mathbf{x} \in \mathbb{X}} |f_{\mathbf{X};n}(\mathbf{x}; h) - \mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h))| \leq C_3 \sqrt{\frac{\log\left(1/l_n\right)_+ + \log\left(2/\delta\right)}{n l_n^{2d - d_{\mathrm{vol}} + \varepsilon}}}. \tag{3.1.1}$$

*Moreover, we have for any $C_4 > 0$, $\mathfrak{K}_1(\mathbb{X}, C_4) < \infty$ and $0 < h_0 < (2\eta)^{d_{\mathrm{vol}} - \varepsilon}$, with probability $1$,*

$$\limsup_{n \to \infty} \sup_{c(\log n/n)^{\gamma} \leq h \leq h_0} \sup_{\mathbf{x} \in \mathbb{X}} |f_{\mathbf{X};n}(\mathbf{x}; h) - \mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h))| =: \mathfrak{K}_1(\mathbb{X}, C_4) \sqrt{\frac{(\log(1/h) \vee \log\log n)}{n h^{2d - d_{\mathrm{vol}} + \varepsilon}}}. \tag{3.1.2}$$

**Proof.**

Applying corollary 2.3.0.2 with $c_{\Psi}(\mathbf{x}) = 0$ and $d_{\Psi}(\mathbf{x}) = 1$, we get (3.1.1) with probability at least $1 - \delta$. From therorem 2.4.0.1 and by (H.ii) [setting $c_{\Psi}(\mathbf{x}) = 0$, $d_{\Psi}(\mathbf{x}) = 1$], (3.1.2) follows with probability 1. ■

**Remark 3.1.0.2** *For proving strong consistency of $|f_{\mathbf{X};n}(\mathbf{x}; h) - f_{\mathbf{X}}(\mathbf{x}; h)|$, we write the difference $f_{\mathbf{X};n}(\mathbf{x}; h) - f_{\mathbf{X}}(\mathbf{x})$ as the sum of probabilistic term $|f_{\mathbf{X};n}(\mathbf{x}; h) - \mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h))|$ and a deterministic term $|\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h)) - f_{\mathbf{X}}(\mathbf{x})|$, the so-called bias. The first (random) term has been studied via empirical process techniques, whereas, the order of the bias depends on smoothness properties of $f_{\mathbf{X}}(\cdot)$ only. That is, if $f_{\mathbf{X}}(\cdot)$ is uniformly continuous density, then Bochner's lemma (see* Einmahl et al. (2005)*) gives that, for $\check{h}_n \to 0$,*

$$\sup_{l_n \leq h \leq \check{h}_n} \sup_{\mathbf{x} \in \mathbb{X}} |\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h)) - f_{\mathbf{X}}(\mathbf{x}; h)| = o(1).$$

Next consider the kernel-type estimator of $r_{\Psi;n}(\mathbf{x};h)$ and general kernel-type estimator of the regression function (see Nadaraya (1964) and Watson (1964)), given respectively by (2.1.5) and (2.1.6).

To prove the strong consistency of $m_{\Psi;n}(\mathbf{x};h)$, we shall consider another, but more appropriate and more computationally convenient, centering factor than the expectation $\mathbb{E}m_{\Psi;n}(\mathbf{x};h)$, which is delicate to handle. This is given by

$$\widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h) = \frac{\mathbb{E}\left(\Psi(\mathbf{Y})K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right)}{\mathbb{E}\left(K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right)} = \frac{\mathbb{E}(r_{\Psi;n}(\mathbf{x};h))}{\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x};h))}.$$

**Remark 3.1.0.3** *Note that $\widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h)$ does not coincide in general with $\mathbb{E}m_{\Psi;n}(\mathbf{x};h)$. However, under mild regularity assumptions, the difference between these two quantities becomes asymptotically negligible as $h_n \to 0$ and $nh_n^d \to \infty$, (examples of this are given in Deheuvels and Mason (2004)).*

**Corollary 3.1.0.4** *Assume that the kernel function $K(\cdot)$ satisfies Assumptions 3 and 4. If there exists $M > 0$ such that*

$$F(\mathbf{Y})\mathbb{1}\{\mathbf{x} \in \mathbf{J}\} \leq M, \quad a.s.,$$

*we have for $C_5 > 0$, and under Assumption 1 with probability at least $1-\delta$,*

$$\sup_{h \geq l_n} \sup_{\mathbf{x} \in \mathbf{I}} \sup_{\Psi \in \mathcal{F}_q} |r_{\Psi;n}(\mathbf{x};h) - \mathbb{E}(r_{\Psi;n}(\mathbf{x};h))| \leq C_5 \sqrt{\frac{\log\left(1/l_n\right)_+ + \log\left(2/\delta\right)}{n l_n^{2d-d_{\mathrm{vol}}+\varepsilon}}} \tag{3.1.3}$$

*Moreover, if we assume that for some $p > 2$*

$$\beta_{\mathbb{P}}(\Psi) := \sup_{\mathbf{x} \in \mathbf{J}} \mathbb{E}(F^p(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}) < \infty$$

*we have for any $C_6 > 0$ and $0 < h_0 < (2\eta)^{d_{\mathrm{vol}}-\varepsilon}$, with probability $1$,*

$$\limsup_{n\to\infty} \sup_{C_6(\log n/n)^{\gamma} \leq h \leq h_0} \sup_{\mathbf{x} \in \mathbf{I}} \sup_{\Psi \in \mathcal{F}_q} |r_{\Psi,n}(\mathbf{x},h) - \mathbb{E}(r_{\Psi;n}(\mathbf{x};h))| =:$$

$$\mathfrak{K}_2(\mathbf{I}, C_6)\sqrt{\frac{(\log(1/h^{d_{\mathrm{vol}}-\varepsilon}) \vee \log\log n)}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}, \tag{3.1.4}$$

*where $\mathfrak{K}_2(\mathbf{I}, C_6) < \infty$.*

**Proof.** Applying theorems (2.3.0.1) and corollary 2.3.0.2 with $c_\Psi(\mathbf{x}) = 1$ and $d_\Psi(\mathbf{x}) = 0$, we get (3.1.3) with probability at least $1 - \delta$. From Theorem 2.4.0.1 and by (H.ii) [setting $c_\Psi(\mathbf{x}) = 1$ and $d_\Psi(\mathbf{x}) = 1$], (3.1.4) follows with probability 1. ■

**Corollary 3.1.0.5** *Assume that the conditions of Theorem 2.3.0.1 and 2.4.0.1 hold, in particular that $\mathcal{F}_D$ are relatively compact with respect to the sup-norm topology on $\mathbf{J}$, and the enveloppe function $F$ satisfies (H.i) or (H.ii) according as the class $\mathcal{F}_q$ is bounded or unbounded. Then for any kernel $K(\cdot)$ satisfying Assumptions 3 and 4 and under Assumption 1, with probability at least $1 - \delta$,*

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\left|m_{\Psi,n}(\mathbf{x},h) - \widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h)\right| \leq C_7\sqrt{\frac{\log\left(1/l_n\right)_+ + \log\left(2/\delta\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{3.1.5}$$

*Moreover, if $\mathcal{F}_q$ is not necessarily bounded, but satisfies (H.ii), we have almost surely*

$$\lim_{n\to\infty}\sup_{C_8(\log n/n)^\gamma\leq h\leq h_0}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\left|m_{\Psi,n}(\mathbf{x};h) - \widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h)\right| := \mathfrak{K}_3(\mathbf{I},C_8)\sqrt{\frac{(\log(1/h)\vee\log\log n)}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{3.1.6}$$

*where $\mathfrak{K}_3(\mathbf{I},C_8) < \infty$.*

**Proof.** It is easy to show that, for all $h > l_n$ the following relation holds

$$\begin{aligned}\left|m_{\Psi;n}(\mathbf{x};h) - \widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h)\right| &\leq \frac{|r_{\Psi,n}(\mathbf{x};h) - \mathbb{E}(r_{\Psi;n}(\mathbf{x};h))|}{|f_{\mathbf{X},\mathbf{n}}(\mathbf{x};h)|} \\ &+ \frac{|\mathbb{E}(r_{\Psi;n}(\mathbf{x};h))|}{|f_{\mathbf{X};n}(\mathbf{x};h)\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x};h))|}\,|f_{\mathbf{X},n}(\mathbf{x};h) - \mathbb{E}(f_{\mathbf{X};n}(\mathbf{x};h))|\end{aligned} \tag{3.1.7}$$

From (3.1.1), (3.1.3) of Corollaries 3.1.0.1 and 3.1.0.4, it follows with probability at least $1 - \delta$,

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\left|m_{\Psi,n}(\mathbf{x},h) - \widehat{\mathbb{E}}(m_{\Psi,n}(\mathbf{x};h))\right| \leq C_7\sqrt{\frac{\log\left(1/l_n\right)_+ + \log\left(2/\delta\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}$$

and from (3.1.2), (3.1.4), it follows with probability 1,

$$\lim_{n\to\infty}\sup_{C_8(\log n/n)^\gamma\leq h\leq h_0}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\left|m_{\Psi,n}(\mathbf{x};h) - \widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h)\right| := \mathfrak{K}_3(\mathbf{I},C_8)\sqrt{\frac{(\log(1/h)\vee\log\log n)}{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{3.1.8}$$

**Remark 3.1.0.6** *Under the assumption that $f_{\mathbf{X},n}(\cdot)$ is bounded away on $\mathbf{I}$, uniformly in $c(\log n/n)^{\gamma} \leq h \leq h_0$ and combining this with (3.1.0.4) or (3.1.0.4) it follows that $\frac{\sup_{\mathbf{x}\in\mathbf{I}}\left|\mathbb{E}(r_{\Psi;n}(\mathbf{x};h))\right|}{\sup_{\mathbf{x}\in\mathbf{I}}|\mathbb{E}f_{\Psi}(\mathbf{x};h)|}$ remains bounded.*

■

**Remark 3.1.0.7** *We note that the main problem in using estimator such as in (2.1.6) is to choose properly the smoothing parameter $h$. The uniform in bandwidth consistency results given in Corollary 3.1.0.5 shows that any choice of $h$ between $h_n'$ and $h_n''$ ensures the consistency of $m_{\Psi;n}(\mathbf{x};h)$. Namely, the fluctuation of the bandwidth in a small interval does not affect the consistency of the nonparametric estimator $m_{\Psi;n}(\mathbf{x};h)$ of $m_{\Psi}(\mathbf{x})$.*

## 3.2 Estimation of the density and regression derivatives

Estimation of function (density or regression) derivatives is a versatile tool in statistical data analysis. A statistical test for modes of the data density, which is based on the second order density derivative Genovese *et al.* (2013). The optimal bandwidth of kernel density estimation depends on the second-order density derivative Noh *et al.* (2018). More applications in fundamental statistical problems such as regression, Fisher information estimation, parameter estimation, and hypothesis testing are discussed in Singh (1977). The derivative of regression, that is used in modal regression, which is an alternate approach to the usual regression methods for exploring the relationship between a response variable $\mathbf{Y}$ and a predictor variable $\mathbf{X}$, we may refer to Ziegler (2003, 2002). Here and elsewhere the non negative integer vector $s \in (\{0\} \cup \mathbb{N})^d$ denotes a fixed order of differentiation in the following sense. Let $\xi$ be an arbitrary measurable function with $\xi : \mathbb{R}^d \rightarrow \mathbb{R}$. We consider estimation of functionals of $\xi$ defined at $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_d) \in \mathbf{J}$. For each $d$-uple of nonnegative integers $s_1 \geq 0, \ldots, s_d \geq 0$, $s = (s_1, \ldots, s_d)$, we define the differential operator $D^s$ of order

$$|s| = s = s_0 + s_1 + \cdots + s_d,$$

where $|s_0| = 0$ and

$$D^s\xi(\mathbf{x}) := \frac{\partial^{|s|}}{\partial\mathbf{x}_1^{s_1}\ldots\partial\mathbf{x}_d^{s_d}}.$$

For $D^s$ operator to be well defined and interchange with integration we need some smoothness condition that will be assumed later. Our aim is to investigate the kernel-type estimators of the $s$-th derivatives $D^s f_{\mathbf{X};n}(\mathbf{x};h)$ and $D^s r_{\mathbf{X};n}(\mathbf{x};h)$ of $f_{\mathbf{X};n}(\mathbf{x};h)$ and $r_{\mathbf{X};n}(\mathbf{x};h)$, respectively, given by

$$D^s f_{\mathbf{X},n}(\mathbf{x},h) = (nh^{d+|s|})^{-1}\sum_{i=1}^{n} D^s K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right), \tag{3.2.1}$$

$$D^s r_{\Psi;n}(\mathbf{x};h) = (nh^{d+|s|})^{-1}\sum_{i=1}^{n} \Psi(\mathbf{Y}_i) D^s K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right). \tag{3.2.2}$$

Introduce the following process. Given any two continuous and bounded functions $c_\Psi$ and $d_\Psi$ on $\mathbf{J}$, set for $\mathrm{x} \in \mathbf{J}$

$$\begin{aligned} W_{n;h}^{(s)}(\mathbf{x};\Psi) = & \sum_{j=1}^{n}(c_\Psi(\mathbf{x})\Psi(\mathbf{Y}_j) + d_\Psi(\mathbf{x}))D^s K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right) \\ & -n\mathbb{E}\left\{(c_\Psi(\mathbf{x})\Psi(\mathbf{Y}) + d_\Psi(\mathbf{x}))D^s K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right\}. \end{aligned} \tag{3.2.3}$$

We treat the Nadaraya-Watson kernel estimator see (Nadaraya (1964); Watson (1964)) and its partial derivatives when the predictor variables are $\mathbb{R}^d$ valued. For this purpose, we need to introduce a general kernel function $K : \mathbb{R}^d \to \mathbb{R}$, fulfilling the conditions

**(K.2)** The partial derivative $D^s K : \mathbb{R}^d \to \mathbb{R}$ exists and

$$\|D^s K\|_\infty, \|D^s K\|_2 < \infty;$$

**(K.3)** The derivative of the kernel is such that:

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

**(K.4)** We assume that

$$\mathcal{K}^{(s)} := \left\{(\mathbf{x},h) \mapsto D^s K\left(\frac{\mathbf{x}-\cdot}{h}\right) : \mathbf{x}\in\mathbb{X}, h \geq l_n\right\}$$

is a uniformly bounded VC-class with dimension $\nu_2$, such that, the covering numbers $\mathcal{N}(\mathcal{K}^{(s)}, L_2(\mathbb{Q}), \varepsilon)$ satisfies

$$\mathcal{N}(\mathcal{K}^{(s)}, L_2(\mathbb{Q}), \varepsilon) \leq \left(\frac{A_2\|D^s K\|_\infty}{\varepsilon}\right)^{\nu_2}.$$

Under Assumption **(K.3)**, we can bound $\mathbb{E}_{\mathbb{P}}\left[D^s K^2\right]$ in terms of volume dimension $d_{\mathrm{vol}}$ as follows (see Kim *et al.* (2018))

**Lemma 3.2.0.1** *Let $(\mathbb{R}^d, \mathbb{P})$ be a probability space and let $\mathbf{X} \sim \mathbb{P}$. For any kernel $K(\cdot)$ satisfying Assumption **(K.3)**, the expectation of the derivative of kernel is upper bounded as*

$$\mathbb{E}_{\mathbb{P}}\left[\left(D^s K\left(\frac{\mathbf{x}-\mathbf{X}}{h}\right)\right)^2\right] \leq C_{s,\mathbb{P},K,\varepsilon} h^{d_{\text{vol}}-\varepsilon}, \tag{3.2.4}$$

*for any $\varepsilon \in (0, d_{\text{vol}})$, where $C_{s,\mathbb{P},K,\varepsilon}$ is a constant depending only on $s, \mathbb{P}, K$ and $\varepsilon$. Further, under Assumption 1, $\varepsilon$ can be 0 in (3.2.4).*

**Proof.** We proceed similarly to proof of Lemma 11 in (Kim *et al.* (2018)), where we plug in $D^s K(\cdot)$ in the place of $K(\cdot)$. ■

Note that we will work in the multivariate framework, where $d \geq 1$, $q \geq 1$ are arbitrary integers, and we keep the assumptions of chapter 2.Throughout, $\ell_n, n = 1, 2, \dots,$ denote a nonrandom (bandwidth) sequence of positive constants satisfying the following assumptions: as $n \to \infty$,

**(H.1)** $\ell_n \searrow 0$ and $n\ell_n^d \nearrow \infty$,

**(H.2)** $n\ell_n^{2k+d}/\log(\ell_n^{-d}) \to \infty$,

**(H.3)** $\log(\ell_n^{-d})/\log\log n \to \infty$.

**(H.4)**

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

**Theorem 3.2.0.2** *Under H.i, **(K.2- 4)**, and assume that $h$ satisfies **(H.1)** and for any $\delta > 0$, we have with probability at least $1-\delta$,*

$$\begin{aligned}
&\sup_{h \geq l_n} \sup_{\mathbf{x} \in \mathbf{I}} \sup_{\Psi \in \mathcal{F}_q} \frac{1}{nh^{d+|s|}} \left| W_{n;h}^{(s)}(\mathbf{x};\Psi) \right| \\
\leq \; & D_0 \left( \frac{(\log(1/l_n))_+}{nl_n^{|s|+d}} + \sqrt{\frac{(\log(1/l_n))_+}{nl_n^{|s|+2d-d_{\text{vol}}+\varepsilon}}} + \sqrt{\frac{\log(2/\delta)}{nl_n^{|s|+2d-d_{\text{vol}}+\varepsilon}}} + \frac{\log(2/\delta)}{nl_n^{|s|+d}} \right)
\end{aligned} \tag{3.2.5}$$

*for any $\varepsilon \in (0, d_{\text{vol}})$, where $D_0$ is a constant depending only on $A$, $\|\vartheta\|_\infty$, $d$, $\nu_1$, $\nu_2$, $d_{\text{vol}}$, $C_{s,C_{\mathcal{F}_q},\mathbb{P},K,\varepsilon}$, $\varepsilon$. Further, under Assumption 1, $\varepsilon$ can be 0 in (3.2.5). Under H.ii, **(K.1-2)** and assume that $h$ satisfies **(H.1-3)**, we have for any $D_1 > 0$ and $0 < h_0 < (2\eta)^{d_{\text{vol}}-\varepsilon}$, with probability 1,*

$$\limsup_{n\to\infty} \sup_{D_1(\log n/n)^{\gamma}\leq h\leq h_0} \sup_{\mathbf{x}\in\mathbf{I}} \sup_{\Psi\in\mathcal{F}_q} \frac{1}{nh^{d+|s|}} \left|W^{(s)}_{n;h}(\mathbf{x};\Psi)\right| := \mathfrak{P}_1(D_1)\sqrt{\frac{(\log(1/h)\vee\log\log n)}{nh^{|s|+2d-d_{\mathrm{vol}}+\varepsilon}}}, \tag{3.2.6}$$

*for any* $\varepsilon\in(0,d_{\mathrm{vol}})$,$\mathfrak{P}_1(D_1)<\infty$ *and* $\gamma=1-2/p$.

**Corollary 3.2.0.3** *Assume* ***(K.1-4)*** *and* $h$ *satisfies* ***(H.1-4)****, Then, with probability at least* $1-\delta$

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbb{X}}\left|f^{(s)}_{\mathbf{X};n}(\mathbf{x};h)-\mathbb{E}(f^{(s)}_{\mathbf{X};n}(\mathbf{x};h))\right|\leq D_2\sqrt{\frac{\log(1/l_n)_+ +\log(2/\delta)}{nl_n^{|s|+2d-d_{\mathrm{vol}}+\varepsilon}}} \tag{3.2.7}$$

*where* $D_2$ *is a constant depending only on* $A$*,* $\|K\|_\infty$*,* $d$*,* $\nu_1$*,* $d_{\mathrm{vol}}$*,* $C_{s,\mathbb{P},K,\varepsilon}$*,* $\varepsilon$*. Further, under Assumption 1,* $\varepsilon$ *can be 0 in (3.2.7). Moreover, assume* ***(K.2-4)*** *and assume that and* $h$ *satisfies* ***(H.1-3)****, for any* $D_3>0$ *and* $0<h_0<(2\eta)^{d_{\mathrm{vol}}-\varepsilon}$*, with probability* $1$*,*

$$\limsup_{n\to\infty}\sup_{D_3(\log n/n)^{\gamma}\leq h\leq h_0}\sup_{\mathbf{x}\in\mathbb{X}}\left|f^{(s)}_{\mathbf{X};n}(\mathbf{x};h)-\mathbb{E}(f^{(s)}_{\mathbf{X}}(\mathbf{x};h))\right| =: \mathfrak{P}_2(\mathbf{I},D_3)\sqrt{\frac{(\log(1/h)\vee\log\log n)}{nh^{|s|+2d-d_{\mathrm{vol}}+\varepsilon}}}, \tag{3.2.8}$$

*where* $\mathfrak{P}_2(\mathbf{I},D_3)<\infty$.

**Corollary 3.2.0.4** *Assume* ***(K.1-4)****, (H.i), and* $h$ *satisfies* ***(H.1-4)****. Then, with probability at least* $1-\delta$

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\left|r^{(s)}_{\Psi;n}(\mathbf{x};h)-\mathbb{E}(r^{(s)}_{\Psi;n}(\mathbf{x};h))\right|\leq D_4\sqrt{\frac{\log(1/l_n)_+ +\log(2/\delta)}{nl_n^{|s|+2d-d_{\mathrm{vol}}+\varepsilon}}} \tag{3.2.9}$$

*where* $D_4$ *is a constant depending only on* $A$*,* $\|K\|_\infty$*,* $d$*,* $\nu_1$*,* $d_{\mathrm{vol}}$*,* $C_{s,\mathbb{P},K,\varepsilon}$*,* $\varepsilon$*. Further, under Assumption 1,* $\varepsilon$ *can be 0 in (3.2.9).*

*Moreover, assume* ***(K.2-4)****,(H.ii), and* $h$ *satisfies* ***(H.1-3)****, for any* $D_5>0$ *and* $0<h_0<(2\eta)^{d_{\mathrm{vol}}-\varepsilon}$*, we have with probability* $1$*,*

$$\limsup_{n\to\infty}\sup_{D_5(\log n/n)^{\gamma}\leq h\leq h_0}\sup_{\mathbf{x}\in\mathbf{I}}\left|r^{(s)}_{\Psi,n}(\mathbf{x},h)-\mathbb{E}(r^{(s)}_{\Psi}(\mathbf{x};h))\right| =: \mathfrak{P}_3(\mathbf{I},D_5)\sqrt{\frac{(\log(1/h)\vee\log\log n)}{nh^{|s|+2d-d_{\mathrm{vol}}+\varepsilon}}}, \tag{3.2.10}$$

*where* $\mathfrak{P}_3(\mathbf{I},D_5)<\infty$.

Now, we treat the derivatives Nadaraya-Watson estimator when the predictor variables are $\mathbb{R}^d$-valued. Unless specified, we will limit most of our exposition to the case where $s = (s_1, \ldots, s_d)$ is such that $s_j = 1$ and $s_l = 0$ for $l \neq j$ and denote by $s_j = (0, \ldots, 1, \ldots, 0)$ the corresponding $d$-uple. It will become obvious later on that our method allow to treat likewise the case of an arbitrary $s$. Thus

$$m_{\Psi;n}^{(s_j)}(\mathbf{x}; h) = D^{(s_j)}(m_{\Psi;n}(\mathbf{x}; h)) = \frac{r_{\Psi;n}^{(s_j)}(\mathbf{x}; h)}{f_{\mathbf{X};n}(\mathbf{x}; h)} - \frac{r_{\Psi;n}(\mathbf{x}; h) f_{\mathbf{X};n}^{(s_j)}(\mathbf{x}; h)}{f_{\mathbf{X};n}^2(\mathbf{x}; h)}.$$

To prove the strong consistency of $m_{\Psi;n}^{(s_j)}(\mathbf{x}; h)$, we shall consider another, but more appropriate and more computationally convenient, centering factor than the expectation $\mathbb{E} m_{\Psi;n}^{(s_j)}(\mathbf{x}; h)$, which is delicate to handle. This is given by

$$\begin{aligned}
\widehat{\mathbb{E}}[m_{\Psi;n}^{(s_j)}(\mathbf{x}; h)] &:= \frac{\mathbb{E}[r_{\Psi;n}^{(s_j)}(\mathbf{x}; h)]}{\mathbb{E}[f_{\mathbf{X};n}(\mathbf{x}; h)]} - \frac{\mathbb{E}[r_{\Psi;n}(\mathbf{x}; h) f_{\mathbf{X};n}^{(s_j)}(\mathbf{x}; h)]}{\mathbb{E}[f_{\mathbf{X};n}^2(\mathbf{x}; h)]} \\
&= \frac{\mathbb{E}(r_{\Psi;n}^{(s_j)}(\mathbf{x}; h))}{\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h))} - \frac{\mathbb{E}(r_{\Psi;n}(\mathbf{x}; h))\mathbb{E}(f_{\mathbf{X};n}^{(s_j)}(\mathbf{x}; h))}{\mathbb{E}(f_{\mathbf{X};n}^2(\mathbf{x}; h))}.
\end{aligned}$$

A similar setup applies for operators $D^s$ with $|s| = s \geq 2$. It is not too difficult to derive from the previous results uniformly consistent estimators $m_{\Psi;n}^{(s)}(\mathbf{x}; h)$ of the partial derivatives of the regression $m_{\Psi;n}(\mathbf{x}; h)$. For the future use, we consider in more detail estimators $m_{\Psi;n}^{(1)}(\mathbf{x}; h)$. We observe that

$$m_{\Psi;n}^{(1)}(\mathbf{x}; h) = \frac{r_{\Psi;n}^{(1)}(\mathbf{x}; h)}{f_{\mathbf{X};n}(\mathbf{x}; h)} - \frac{r_{\Psi;n}(\mathbf{x}; h) f_{\mathbf{X};n}^{(1)}(\mathbf{x}; h)}{f_{\mathbf{X};n}^2(\mathbf{x}; h)}. \tag{3.2.11}$$

**Corollary 3.2.0.5** *Assume the conditions of Theorem 2.3.0.1, and the envelope function $F$ satisfies (H.i). Then, we have, almost surely*

$$\sup_{h \geq l_n} \sup_{\mathbf{x} \in \mathbf{I}} \left| m_{\Psi;n}^{(1)}(\mathbf{x}; h) - \widehat{\mathbb{E}}\left[m_{\Psi;n}^{(1)}(\mathbf{x}; h)\right]\right| \leq D_6 \sqrt{\frac{\log(1/l_n)_+ + \log(2/\delta)}{n l_n^{1+2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{3.2.12}$$

**Proof.** Observe that we have the following chain of inequalities

$$\begin{aligned}
&\left|m_{\Psi,n}^{(1)} - \widehat{\mathbb{E}} m_{\Psi;n}^{(1)}\right| \\
&= \left|\frac{r_{\Psi,n}^{(1)}}{f_{\mathbf{X},n}} - \frac{r_{\Psi,n} f_{\mathbf{X},n}^{(1)}}{f_{\mathbf{X},n}^2} - \frac{\mathbb{E} r_{\Psi,n}^{(1)}}{\mathbb{E} f_{\mathbf{X},n}} + \frac{\mathbb{E} r_{\Psi,n} \mathbb{E} f_{\mathbf{X},n}^{(1)}}{\mathbb{E} f_{\mathbf{X},n}^2}\right| \\
&\leq \frac{1}{|f_{\mathbf{X},n}|}\left|r_{\Psi,n}^{(1)} - \mathbb{E} r_{\Psi,n}^{(1)}\right| + \frac{|\mathbb{E} r_{\Psi,n}^{(1)}|}{|f_{\mathbf{X},n} \mathbb{E} f_{\mathbf{X},n}|}\left|\mathbb{E} f_{\mathbf{X},n} - f_{\mathbf{X},n}\right| - \left|\frac{f_{\mathbf{X},n}^{(1)}}{f_{\mathbf{X},n}} m_{\Psi,n} - \frac{\mathbb{E} f_{\mathbf{X},n}^{(1)}}{\mathbb{E} f_{\mathbf{X},n}} \widehat{\mathbb{E}} m_{\Psi,n}\right|
\end{aligned}$$

$$\leq \frac{1}{|f_{\mathbf{X},n}|}\left|r^{(1)}_{\Psi,n}-\mathbb{E}r^{(1)}_{\Psi,n}\right|+\frac{|\mathbb{E}r^{(1)}_{\Psi,n}|}{|f_{\mathbf{X},n}\mathbb{E}f_{\mathbf{X},n}|}\left|f_{\mathbf{X},n}-\mathbb{E}f_{\mathbf{X},n}\right|-\frac{|m_{\Psi,n}|}{|f_{\Psi,n}|}\left|f^{(1)}_{\mathbf{X},n}-\mathbb{E}f^{(1)}_{\mathbf{X},n}\right|$$
$$+\frac{|\mathbb{E}f^{(1)}_{\mathbf{X},n}|}{|f_{\mathbf{X},n}|}\left|m_{\Psi,n}-\widehat{\mathbb{E}}m_{\Psi,n}\right|. \qquad (3.2.13)$$

Applying Theorem 2.3.0.1 and 2.4.0.1 and from Corollaries 3.1.0.1 and 3.1.0.5, we get that both

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbb{X}}\left|f_{\mathbf{X},n}(\mathbf{x},h)-\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x};h))\right|\leq C_3\sqrt{\frac{\log\left(1/l_n\right)_+ +\log\left(2/\delta\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}, \qquad (3.2.14)$$

and

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\Psi\in\mathcal{F}_q}\left|m_{\Psi,n}(\mathbf{x},h)-\widehat{\mathbb{E}}m_{\Psi,n}(\mathbf{x};h)\right|\leq C_7\sqrt{\frac{\log\left(1/l_n\right)_+ +\log\left(2/\delta\right)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \qquad (3.2.15)$$

Moreover, Corollary 3.2.0.3 and equation (3.2.9) of Corollary 3.2.0.4 with $s=1$, gives

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbb{X}}\left|f^{(1)}_{\mathbf{X},n}(\mathbf{x},h)-\mathbb{E}f^{(1)}_{\mathbf{X},n}(\mathbf{x},h)\right|\leq D_2\sqrt{\frac{\log\left(1/l_n\right)_+ +\log\left(2/\delta\right)}{nl_n^{1+2d-d_{\mathrm{vol}}+\varepsilon}}}, \qquad (3.2.16)$$

And

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\left|r^{(1)}_{\Psi,n}(\mathbf{x},h)-\mathbb{E}r^{(1)}_{\Psi,n}(\mathbf{x},h)\right|\leq D_4\sqrt{\frac{\log\left(1/l_n\right)_+ +\log\left(2/\delta\right)}{nl_n^{1+2d-d_{\mathrm{vol}}+\varepsilon}}}. \qquad (3.2.17)$$

Using the fact that $f_{\mathbf{X},n}(\cdot)$ is bounded away on $\mathbf{I}$. Therefore, we can infer (3.2.0.5) from (3.2.14)-(3.2.17). ■

**Remark 3.2.0.6** *The treatment of the other derivatives for $s>2$ is similar and will not be presented here for the sake of clarity. We note that, when $s\geq 2$, $m^{(s)}_{\Psi,n}(\mathbf{x},h)=D^s(m_{\Psi,n}(\mathbf{x},h))$ may be obtained likewise thought the usual Leibniz expansion of derivatives of products given by*

$$m^{(s)}_{\Psi;n}(\mathbf{x};h)=\sum_{j=0}^{s}C_s^j r^{(j)}_{\Psi;n}(\mathbf{x};h)\{f^{(-1)}_{\mathbf{X};n}(\mathbf{x};h)\}^{(s-j)},\quad f_{\mathbf{X};n}(\mathbf{x};h)\neq 0.$$

## 3.3 Kernel estimator of the conditional distribution function

By setting $\Psi_{\mathbf{t}}(\mathbf{Y}) = \mathbb{1}\{\mathbf{Y} \leq \mathbf{t}\}$, $\mathbf{t} \in \mathbb{R}^q$, into (2.1.6) we obtain the kernel estimator of the conditional distribution function given by

$$F_{n,h}(\mathbf{t} \mid \mathbf{x}) := \frac{\sum_{i=1}^{n} \mathbb{1}\{\mathbf{Y}_i \leq \mathbf{t}\} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}. \tag{3.3.1}$$

This kernel estimator is called the conditional empirical distribution function and was first studied by Stute (1986a). Considering the bounded case in Corollary 3.1.0.5.

**Corollary 3.3.0.1** *Assume the conditions of Theorem 2.3.0.1 and $F$ satisfies (H.i). Then for any kernel $K(\cdot)$ satisfying Assumptions 3 and 4 and under Assumption 1, we have almost surely*

$$\sup_{h \geq l_n} \sup_{\mathbf{x} \in \mathbf{I}} \left| F_{n,h}(t \mid \mathbf{x}) - \widehat{\mathbb{E}}(F_{n,h}(t \mid \mathbf{x})) \right| \leq C' \sqrt{\frac{\log\left(1/l_n\right)_+ + \log\left(2/\delta\right)}{n l_n^{2d - d_{\text{vol}} + \varepsilon}}}, \tag{3.3.2}$$

*where*

$$\widehat{\mathbb{E}}(F_{n,h}(t \mid \mathbf{x})) = \mathbb{E}\left[\mathbb{1}\{\mathbf{Y} \leq t\} K\left((\mathbf{x} - \mathbf{X})/h\right)\right] / h\mathbb{E}(f_{\mathbf{X};n}(\mathbf{x}; h)).$$

Corollary 3.3.0.1 being direct consequence of (3.1.0.5), the bounded case, details of its proof are omitted.

## 3.4 Multivariate mode

In the sequel, we follow Mokkadem and Pelletier (2003), and we state the problem from this paper by keeping the same notation and definitions. The kernel mode estimator is any random variable $\widehat{\boldsymbol{\Theta}}_{n,h_n}$ satisfying

$$f_{\mathbf{X};n}(\widehat{\boldsymbol{\Theta}}_{n,h_n}; h_n) = \sup_{\mathbf{x} \in \mathbb{R}^d} f_{\mathbf{X};n}(\mathbf{x}; h_n). \tag{3.4.1}$$

Since $K(\cdot)$ is continuous and vanishing at infinity, the choice of $\widehat{\boldsymbol{\Theta}}_{n,h_n}$ as a random variable satisfying (3.4.1) can be made with the help of an order on $\mathbb{R}^d$. The definition

$$\widehat{\boldsymbol{\Theta}}_{n,h_n} = \inf \left\{ \mathbf{y} \in \mathbb{R}^d \ \text{ such that } \ f_{\mathbf{X};n}(\mathbf{y}; h_n) = \sup_{\mathbf{x}\in\mathbb{R}^d} f_{\mathbf{X};n}(\mathbf{x}; h_n) \right\},$$

where the infimum is taken with respect to the lexicographic order on $\mathbb{R}^d$, ensures the measurability of the kernel mode estimator. Let us now assume $\boldsymbol{\Theta}$ is nondegenerate, that is, $D^2 f(\boldsymbol{\Theta})$ (the second order differential at the point $\boldsymbol{\Theta}$) is nonsingular. We denote by $\nabla \ell(\cdot)$ the gradient of the function $\ell(\cdot)$. By definition of $\widehat{\boldsymbol{\Theta}}_{n,h_n}$, we have

$$\nabla f_{\mathbf{X};n}(\widehat{\boldsymbol{\Theta}}_{n,h_n}; h_n) = 0,$$

so that

$$\nabla f_{\mathbf{X};n}(\widehat{\boldsymbol{\Theta}}_{n,h_n}; h_n) - \nabla f_{\mathbf{X};n}(\boldsymbol{\Theta}; h_n) = -\nabla f_{\mathbf{X};n}(\boldsymbol{\Theta}; h_n). \tag{3.4.2}$$

For each $i = 1, \ldots, d$, Taylor's expansion applied to the real-valued application $\frac{\partial f_{\mathbf{X};n}(\cdot;h_n)}{\partial x_i}$ implies the existence of $\boldsymbol{\xi}_n(i) = (\xi_{n;1}(i), \ldots, \xi_{n;d}(i))^\top$ such that

$$\begin{cases} \dfrac{\partial}{\partial x_i} f_{\mathbf{X};n}(\widehat{\boldsymbol{\Theta}}_{n,h_n}; h_n) - \dfrac{\partial}{\partial x_i} f_{\mathbf{X};n}(\boldsymbol{\Theta}; h_n) = \displaystyle\sum_{j=1}^{d} \dfrac{\partial^2}{\partial x_i \partial x_j} f_{\mathbf{X};n}(\boldsymbol{\xi}_n(i); h_n)(\widehat{\theta}_{n,j;h_n} - \theta_j), \\ |\xi_{n;j}(i) - \theta_j| \leq |\widehat{\theta}_{n,j;h_n} - \theta_j|, \ \ \forall j = 1, \ldots, d. \end{cases}$$

Define the $d \times d$ matrix $H_n = (H_{n,i,j})_{1\leq i,j\leq d}$ by setting

$$H_{n,i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f_{\mathbf{X};n}(\boldsymbol{\xi}_n(i); h_n).$$

Equation (3.4.2) can then be rewritten as

$$H_n(\widehat{\boldsymbol{\Theta}}_{n,h_n} - \boldsymbol{\Theta}) = -\nabla f_{\mathbf{X};n}(\boldsymbol{\Theta}; h_n). \tag{3.4.3}$$

Application of Corollary 3.2.0.3 ensures that

$$\lim_{n\to\infty} \sup_{\mathbf{x}\in\mathbb{R}^d} \left| \frac{\partial^2}{\partial x_i \partial x_j} f_{\mathbf{X};n}(\mathbf{x}; h_n) - \mathbb{E}\left( \frac{\partial^2}{\partial x_i \partial x_j} f_{\mathbf{X};n}(\mathbf{x}; h_n) \right) \right| = 0, a.s.$$

Moreover, classical computations give the uniform convergence of $\mathbb{E}\Big(\frac{\partial^2}{\partial x_i \partial x_j}$ $f_{\mathbf{X};n}(\mathbf{x}; h_n))$ to $\frac{\partial^2}{\partial x_i \partial x_j} f_{\mathbf{X}}(\mathbf{x})$ in a neighborhood of $\boldsymbol{\Theta}$. Since

$$\lim_{n\to\infty} \widehat{\boldsymbol{\Theta}}_{n,h_n} = \boldsymbol{\Theta} \ \ a.s.,$$

we thus obtain

$$\lim_{n\to\infty} H_n = D^2 f(\boldsymbol{\Theta}).$$

In view of (3.4.3), it follows that the convergence rate of $\widehat{\boldsymbol{\Theta}}_{n,h_n} - \boldsymbol{\Theta}$ is given by that of $[D^2 f(\boldsymbol{\Theta})]\nabla f_{\mathbf{X};n}(\boldsymbol{\Theta}; h_n)$ refer to Mokkadem and Pelletier (2003). Under some regularity assumptions and making use of Corollary 3.2.0.3, one can show that

$$\limsup_{n\to\infty} \sup_{D_3(\log n/n)^{\gamma}\leq h\leq h_0} \frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}}\left|\widehat{\boldsymbol{\Theta}}_{n,h_n}-\boldsymbol{\Theta}\right|}{\sqrt{(\log(1/h)\vee\log\log n)}}<\infty, \tag{3.4.4}$$

## 3.5 Shannon's entropy

The differential (or Shannon) entropy of $f(\cdot)$ is defined to be

$$H(f) \;:=\; -\int_{\mathbb{R}^d} f(\mathbf{x})\log\left(f(\mathbf{x})\right)d\mathbf{x} \tag{3.5.1}$$

$$:=\; -\int_{\mathbb{R}^d} \log\left(f(\mathbf{x})\right)d\mathbb{F}(\mathbf{x}), \tag{3.5.2}$$

whenever this integral is meaningful, and where, for $\mathbf{x}=(x_1,\ldots,x_d)$, $d\mathbf{x}$ denotes Lebesgue measure in $\mathbb{R}^d$. We will use the convention that $0\log(0)=0$ since $u\log(u)\to 0$ as $u\to 0$. The concept of differential entropy was originally introduced in Shannon's paper Shannon (1948). Since this early epoch, the notion of entropy has been the subject of great theoretical and applied interest. We refer to (Cover, 2006, Chapter 8.) for a comprehensive overview of differential entropy and their mathematical properties. Entropy concepts and principles play an fundamental role in many applications, such as quantization theory Rényi (1959), statistical decision theory Kullback (1959), and contingency table analysis Gokhale and Kullback (1978). Csiszár (1962) introduced the concept of convergence in entropy and showed that the latter convergence concept implies convergence in $\mathcal{L}_1$. This property indicates that entropy is a useful concept to measure "*closeness in distribution*", and also justifies heuristically the usage of sample entropy as test statistics when designing entropy-based tests of goodness-of-fit. This line of research has been pursued by Vasicek (1976); Dudewicz and Van Der Meulen (1981); Ebrahimi *et al.* (1992) [including the references therein]. The idea here is that many families of distributions are characterized by maximization of entropy subject to constraints (see, e.g., Jaynes (1957) and Lazo and Rathie (1978)). Given $f_{\mathbf{X},n}(\cdot,h)$ in (3.1), we estimate $H(f)$ using the representation (3.5.1), by setting

$$H_{n,h_n}(f):=-\int_{A_n} f_{\mathbf{X},n}(\mathbf{x},h_n)\log\big(f_{\mathbf{X};n}(\mathbf{x};h_n)\big)d\mathbf{x}, \tag{3.5.3}$$

where

$$\mathbf{A}_n:=\{\mathbf{x}\in\mathbb{R}^d : f_{\mathbf{X};n}(\mathbf{x};h_n)\geq\gamma_n\},$$

and $\gamma_n \downarrow 0$ is a sequence of positive constant. To prove the strong consistency of $H_{n,h_n}(f)$, we shall consider another, but more appropriate and more computationally convenient, centering factor than the expectation $\mathbb{E}H_{n,h_n}(f)$, which is delicate to handle. This is given by

$$\widehat{\mathbb{E}}H_{n;h_n}(f) := -\int_{A_n} \mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)\log\big(\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)\big)d\mathbf{x}.$$

We first decompose $H_{n;h_n}(f) - \widehat{\mathbb{E}}H_{n;h_n}(f)$ into the sum of two components, by writing

$$\begin{aligned}
&H_{n,h_n}(f) - \widehat{\mathbb{E}}H_{n,h_n}(f)\\
&\quad= -\int_{A_n} f_{\mathbf{X};n}(\mathbf{x};h_n)\log\big(f_{\mathbf{X};n}(\mathbf{x};h_n)\big)d\mathbf{x}\\
&\qquad+\int_{A_n} \mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)\log\big(\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)\big)d\mathbf{x}\\
&\quad= -\int_{A_n}\{\log f_{\mathbf{X};n}(\mathbf{x};h_n) - \log\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)\}\,\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)d\mathbf{x}\\
&\qquad-\int_{A_n}\{f_{\mathbf{X};n}(\mathbf{x};h_n) - \mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)\}\log f_{\mathbf{X};n}(\mathbf{x};h_n)d\mathbf{x}\\
&\quad:= \mathbf{\Delta}_{1,n,h_n} + \mathbf{\Delta}_{2,n,h_n}.
\end{aligned}\tag{3.5.4}$$

We observe that for all $z > 0$,

$$|\log z| \leq \left|\frac{1}{z} - 1\right| + |z-1|\,.$$

Therefore, for any $\mathrm{x} \in \mathbf{A}_n$, we get

$$\begin{aligned}
|\log f_{\mathbf{X};n}(\mathbf{x};h_n) - \log\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)| &= \left|\log\frac{f_{\mathbf{X};n}(\mathbf{x};h_n)}{\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)}\right|\\
\leq \left|\frac{\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)}{f_{\mathbf{X};n}(\mathbf{x};h_n)} - 1\right| + \left|\frac{f_{\mathbf{X};n}(\mathbf{x};h_n)}{\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)} - 1\right|&\\
= \frac{|\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n) - f_{\mathbf{X};n}(\mathbf{x};h_n)|}{f_{\mathbf{X};n}(\mathbf{x};h_n)} + \frac{|f_{\mathbf{X};n}(\mathbf{x};h_n) - \mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)|}{\mathbb{E}f_{\mathbf{X};n}(\mathbf{x};h_n)}.&
\end{aligned}$$

Under conditions of Corollary 3.1.0.1, in addition we assume, $cn^{-1}\gamma_n^{-4}(\log n) \leq h_n$, then we can prove the following result, which is an extension of Theorem 2.1 of Bouzebda and Elhattab (2011), there exists a positive constant $\Upsilon$, such that

$$\limsup_{n\to\infty}\sup_{h_n\leq h\leq 1}\frac{\sqrt{nh^{2d-d_{\mathrm{vol}}+\varepsilon}\gamma_n^4}|H_{n,h}(f) - \widehat{\mathbb{E}}H_{n,h}(f)|}{\sqrt{(\log(1/h)\vee\log\log n)}} \leq \infty \;\; a.s.$$

## 3.6 Additive models

Notice that the parametric regression models provide useful tools for analyzing practical data when the models are correctly specified, but may suffer from large modeling biases when the structures of the models are misspecified, which is the case in many practical problems. As an alternative, nonparametric smoothing methods ease the concerns on modeling biases. However, it is well known that unrestricted multivariate nonparametric regression models are subject to the *curse of dimensionality*, in multivariate settings, and fail to take advantage of the flexibility structure in modeling phenomena with *moderate* set of data, see Stone (1985, 1986), Fan and Gijbels (1996) and Härdle (1990) among others. We assume that the relation between the response variables and the covariates can be represented by the following relation

$$\Psi(Y) = m_0 + m_\Psi(\mathbf{X}) + \varepsilon, \tag{3.6.1}$$

where $m(\cdot)$ is the nonlinear part of the model and $\varepsilon$ is the modeling error and $m_0$ is a constant. The papers by Stone (1985, 1986) proposed the additive model regression, which allows easier interpretation of the contribution of each explanatory variable and reduction of the computational requirement. Hence, additive regression models circumvent the *curse of dimensionality* that afflicts the estimation of fully nonparametric regression models, the interested reader may refer to Fan and Gijbels (1996), Härdle (1990). To reduce the dimension impact an additive structure to the nonparametric function $m_\psi(\cdot)$, that is

$$\begin{aligned} m_\Psi(\mathbf{x}) &= \mathbb{E}\left(\Psi(Y) \mid \mathbf{X} = \mathbf{x}\right), \ \forall\, \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \\ &= \mu + \sum_{l=1}^{d} m_l(x_l) := m_{\Psi,add}(\mathbf{x}), \end{aligned} \tag{3.6.2}$$

where $x_l$ is the $l$-th component of the vector x. To avoid unnecessary complexity, we assume that $\mu = 0$. For the identifiability purpose of the additive component functions $m_l(\cdot)$, we impose the constraints

$$\mathbb{E}m_l(X_l) = 0 \ \text{ for } \ 1 \leq l \leq d.$$

Notice that the backfitting algorithm of Breiman and Friedman (1985), Buja *et al.* (1989) and Hastie and Tibshirani (1990) is widely used to estimate the one - dimensional components $m_l(\cdot)$ and regression function $m(\cdot)$. The backfitting idea is to project the data onto the space of functions which are additive. This projection is done via least squares, where the least squares problem is solved with the Gauss-Seidel algorithm. Notice that the additive model has now become a widely used multivariate smoothing technique, in large part due to the extensive discussion in Hastie and Tibshirani (1990), where the authors give a good overview and analyze estimation techniques based on backfitting, and the availability of fitting routines in S-Plus, described in Chambers *et al.* (1990). It should be remarked that important progress has also been made by Mammen *et al.* (1999) or Opsomer *et al.* (1997) in the asymptotic theory of backfitting. Mammen *et al.* (2012) provided an overview over smooth backfitting type estimators in additive models and also discussed extensions to varying coefficient models, additive models with missing observations, and the case of nonstationary covariates. Auestad and Tjø stheim (1991), Tjøstheim and Auestad (1994) and Linton and Nielsen (1995) proposed a method based on marginal integration of the mean function $m(\cdot)$ for estimating the additive components. Their analysis is restricted to the case of dimension $d = 2$, Chen *et al.* (1996) tried to extend this result to arbitrary $d$, we may refer also to Newey (1994), Bouzebda and Chokri (2014), Bouzebda *et al.* (2016) and Bouzebda and Didi (2017) for more references. One advantage of the integration method is that its statistical properties are easier to describe; specifically, one can easily prove central limit theorems and give explicit expressions for the asymptotic bias and variance of the estimators. There is a main disadvantage of the integration estimator which is perhaps even more time consuming to compute than the backfitting estimator. Motivated by all these properties, our approach will be based on the marginal integration method. Let $q_1(\cdot), \ldots, q_d(\cdot)$ be $d$ density functions defined in $\mathbb{R}$ with some compact support included in $\mathcal{C}$, where

$$\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_d$$

is a compact set of $\mathbb{R}^d$, and $\mathcal{C}_1, \ldots, \mathcal{C}_d$ be a compact intervals of $\mathbb{R}$. Setting

$$q(\mathbf{x}) = \prod_{l=1}^{d} q_l(x_l) \text{ and } q_{-l}(\mathbf{x}_{-l}) = \prod_{j=1, j\neq l}^{d} q_j(x_j).$$

We then obtain

$$\eta_l(x_l) = \int_{\mathbb{R}^{d-1}} m_\Psi(\mathbf{x}) d\mathbf{x}_{-l} - \int_{\mathbb{R}^d} m_\Psi(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}, \quad \text{for} \quad l = 1, \ldots, d, \tag{3.6.3}$$

in such a way that the following two equalities hold,

$$\eta_l(x_l) = m_l(x_l) - \int_{\mathbb{R}} m_l(z) q_l(z) dz, \quad \text{for} \quad l = 1, \ldots, d, \tag{3.6.4}$$

$$m_{\Psi,add}(\mathbf{x}) = \sum_{l=1}^{d} \eta_l(x_l) + \int_{\mathbb{R}^d} m_\Psi(\mathbf{z}) q(\mathbf{z}) d\mathbf{z}. \tag{3.6.5}$$

Finally, we will assume tacitly the following assumptions on the density functions $q_\ell(\cdot)$, for $1 \leq \ell \leq d$.

(Q.1) $q_\ell(\cdot)$ is bounded and continuous, for all $1 \leq \ell \leq d$;

(Q.2) $q_\ell(\cdot)$ has $k+1$ continuous and bounded derivatives.

In view of (3.6.4) and (3.6.5), we note that $\eta_l(\cdot)$ and $m_l(\cdot)$ are equal up to an additional constant. Therefore, $\eta_l(\cdot)$ is also an additive component, fulfilling a different identifiability condition. From (2.3.9) and (3.6.3), a natural estimate of this $l$-th component is given by

$$\widehat{\eta}_l(x_l;h) = \int_{\mathbb{R}^{d-1}} m_{\Psi;n}(\mathbf{x};h)q_{-l}(\mathbf{x}_{-l})d\mathbf{x}_{-l} - \int_{\mathbb{R}^d} m_{\Psi;n}(\mathbf{x};h)q(\mathbf{x})d\mathbf{x}, \quad \text{for} \quad l = 1,\ldots,d, \tag{3.6.6}$$

from which we deduce the estimate $\widehat{m}_{\Psi,n,add}(\cdot)$ of the additive regression function,

$$\widehat{m}_{\Psi,n,add}(\mathbf{x};h) \quad = \quad \sum_{l=1}^{d} \widehat{\eta}_l(x_l;h) + \int_{\mathbb{R}^d} m_{\Psi,n}(\mathbf{x})q(\mathbf{x})d\mathbf{x}. \tag{3.6.7}$$

Notice that we have, for $l = 1,\ldots,d$,

$$\begin{aligned}\widehat{\eta}_l(x_l;h) - \widehat{\mathbb{E}}\widehat{\eta}_l(x_l;h) \quad = \quad & \int_{\mathbb{R}^{d-1}} \left\{ m_{\Psi;n}(\mathbf{x};h) - \widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h) \right\} q_{-l}(\mathbf{x}_{-l})d\mathbf{x}_{-l} \\ & - \int_{\mathbb{R}^d} \left\{ m_{\Psi;n}(\mathbf{x};h) - \widehat{\mathbb{E}}m_{\Psi;n}(\mathbf{x};h) \right\} d\mathbf{x}. \end{aligned} \tag{3.6.8}$$

Making use of Corollary 3.1.0.5 and similar arguments to those used in Bouzebda *et al.* (2016) in the proof of Theorem 3.1, one can show that the relation (3.6.8) can be used to show that

$$\limsup_{n\to\infty} \sup_{x_1 \in I \subset \mathbb{R}} \sqrt{\frac{nh}{2\log(1/h)}} \left| \widehat{\eta}_1(x_1;h) - \widehat{\mathbb{E}}\widehat{\eta}_1(x_1;h) \right| < C_9 \quad a.s., \tag{3.6.9}$$

where $C_9 > 0$. By the same reasoning as in Theorem 3.2 in Bouzebda *et al.* (2016), one can show that

$$\limsup_{n\to\infty} \sup_{\mathbf{x}\in I^d \subset \mathbb{R}^d} \sqrt{\frac{nh}{2\log(1/h)}} \left| \widehat{m}_{\Psi,n,add}(\mathbf{x};h) - \widehat{\mathbb{E}}\widehat{m}_{\Psi,n,add}(\mathbf{x};h) \right| < C_{10} \quad a.s., \tag{3.6.10}$$

where $C_{10} > 0$ and

$$\widehat{\mathbb{E}}\widehat{m}_{\Psi,n,add}(\mathbf{x};h) \quad = \quad \sum_{l=1}^{d} \widehat{\mathbb{E}}\widehat{\eta}_l(x_l;h) + \int_{\mathbb{R}^d} \widehat{\mathbb{E}}m_{\Psi,n}(\mathbf{x})q(\mathbf{x})d\mathbf{x}. \tag{3.6.11}$$

We will not discuss further (3.6.9) and (3.6.10), and leave its study open for future research. It will be of interest to consider the exact constants in the preceding equations, the proof of such a statement, however, should require a different methodology than that used in this thesis, and we leave this problem for future work.

**Remark 3.6.0.1** *The extension of results to te the functional setting is difficult since the only available papers for the "volume dimension" are only given in the finite dimensional framework. The subject is itself the subject of a long paper to be investigated. Even if this is out of the scope of the present paper, we give some idea about the dimensionality reduction as suggest by the referee. To be more precise, it well known that the estimation problems of a regression function are especially hard in the case when the dimension of the explanatory* $\mathbf{X}$ *is large. It worth noticing that one consequence of this is that the optimal minimax rate of convergence* $n^{-2k/(2k+d)}$ *for the estimation of a* $k$ *times differentiable regression function converges to zero rather slowly if the dimension d of* $\mathbf{X}$ *is large compared to* $k$*. To circumvent the so-called curse of dimensionality, the only way is to impose additional assumptions on the regression functions. The simplest way is to consider the linear models but this rather restrictive parametric assumption can be extended in several ways. An idea is to consider the additif models to simplify the problem of regression estimation by fitting only functions to the data which have the same additive structure. In projection pursuit one generalizes this further by assuming that the regression function is a sum of univariate functions applied to projections of* $\mathbf{x}$ *onto various directions, we note that this includes the single index models as particular cases, the interested reader may refer to (Györfi* et al.*, 2002, Chapter 22) for more rigorous developments of such techniques. Other ways are to be investigated are the semi-parametric models, considered like intermediary models between linear and nonparametric ones, aiming to combine the flexibility of nonparametric approaches together with the interpretability of the parametric ones, for details on these methods for functional data, one can refer to (Ling and Vieu, 2018, Section 4.2) and the reference therein. Notice that there is some recent advances for the uniform convergence rate in the functional framework, we may refer to Kara-Zaitri* et al. *(2017), Ling* et al. *(2019) and Bouzebda and Nemouchi (2020) among many other references.*

## 3.7 The bandwidth selection criterion

The selection of the bandwidth, however, is more problematic. It is worth noticing that the choice of the bandwidth is crucial to obtain a good rate of consistency, for example, it has a big influence on the size of the estimate's bias. In general, we are interested in the selection of bandwidth that produces an estimator which has a good balance between the bias and the variance of the considered estimators. It is then more appropriate to consider the bandwidth varying according to the criteria applied and to the available data and location which cannot be achieved by using the classical methods. The interested reader my refer to Mason (2012) for more details and discussion on the subject.

Many methods have been established and developed to construct, in asymptotically optimal ways, bandwidth selection rules for nonparametric kernel estimators especially for Nadaraya-Watson regression estimator we quote among them Hall (1984), Hardle and Marron (1985), Tsybakov (1987), and Rachdi and Vieu (2007). This parameter has to be selected suitably, either in the standard finite dimensional case, or in the infinite dimensional framework for insuring good practical performances. Let us define the leave-out-$(\mathbf{X}_i, \mathbf{Y}_i)$ estimator for regression function

$$m_{\Psi,n,j}(\mathbf{x};h) = \frac{\sum_{i=1,i\neq j}^{n} \Psi(\mathbf{Y}_i) K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}{\sum_{i=1,i\neq j}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}. \tag{3.7.1}$$

In order to minimize the quadratic loss function, we introduce the following criterion, we have for some (known) non-negative weight function $\mathcal{W}(\cdot)$ :

$$CV\left(\Psi,h\right) := \frac{1}{n}\sum_{j=1}^{n}\left(\Psi\left(\mathbf{Y}_j\right) - m_{\Psi,n,j}(\mathbf{x};h)\right)^2 \mathcal{W}\left(\mathbf{X}_j\right). \tag{3.7.2}$$

Following the ideas developed by Rachdi and Vieu (2007), a natural way for choosing the bandwidth is to minimize the precedent criterion, so let's choose $\widehat{h}_n \in [a_n, b_n]$ minimizing among $h \in [a_n, b_n]$ :

$$\sup_{\Psi\in\mathcal{F}_q} CV\left(\Psi,h\right),$$

we can conclude, by Corollary 3.1.0.5, that :

$$\sup_{\Psi\in\mathcal{F}_q}\sup_{\mathbf{x}\in\mathbf{I}}\left|m_{\Psi,n,j}(\mathbf{x};\widehat{h}_n) - \widehat{E}(m_{\Psi,n,j}(\mathbf{x};\widehat{h}_n))\right| \longrightarrow 0 \qquad \text{p.s.}$$

The main interest of our results is the possibility to derive the asymptotic properties of our estimate even if the bandwidth parameter is a random variable, like in the last equation. One can replace (3.7.2) by

$$CV\left(\Psi,h\right) := \frac{1}{n}\sum_{j=1}^{n}\left(\Psi\left(\mathbf{Y}_j\right) - m_{\Psi,n,j}(\mathbf{x};h)\right)^2 \widehat{\mathcal{W}}\left(\mathbf{X}_j,\mathbf{x}\right). \tag{3.7.3}$$

In practice, one takes for $j = 1, \ldots, n$, the uniform global weights $\mathcal{W}(\mathbf{X}_j) = 1$, and the local weights

$$\widehat{W}(\mathbf{X}_j, \mathbf{x}) = \begin{cases} 1 & \text{if} \quad \|\mathbf{X}_j - \mathbf{x}\| \leq h, \\ 0 & \quad\text{otherwise}. \end{cases}$$

**Remark 3.7.0.1** *Deheuvels and Mason (2004) consider local plug-in type estimators $\widehat{h}_n = \widehat{h}_n(x)$, which satisfy,*

$$\mathbb{P}\left(a_n \leq \widehat{h}_n(x) \leq b_n : x \in \mathbb{R}\right) \to 1,$$

*with $a_n = c_1 h_n$ and $b_n = c_2 h_n$, where $0 < c_1 \leq c_2 < \infty$, or fulfil, for any $\varepsilon > 0$*

$$\mathbb{P}\left(\sup_{x \in \mathbf{I}} \left|\frac{\widehat{h}_n(x)}{h_n} - \eta(x)\right| > \varepsilon\right) \to 0, \tag{3.7.4}$$

*where $\eta(\cdot)$ is an appropriate continuous function on $\mathbb{R}$ and $I = [a, b] \subset \mathbb{R}$, for $a < b$. We refer to their Example 2.1 p. 246, where they show subject to smoothness conditions that the optimal $\widehat{h}_n(x)$ satisfies (3.7.4) with $h_n = n^{-1/5}$, for $d = 1$, in terms of asymptotic mean square error for estimating the density function $f(\cdot)$ or regression function $m_\Psi(\cdot)$. Following their methods, it will be interesting to derive our results for local plug-in estimators $\widehat{h}_n(x)$, where the convergence is either in probability or with probability 1 depending on conditions on $\widehat{h}_n(x)$. We omit the corresponding details here.*

# Chapter 4

# Uniform in bandwidth consistency for nonparamteric I.P.C.W. estimators of the regression function in censored case

## Introduction

Consider a triple $(Y, C, \mathbf{X})$ of random variables defined in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$. Here $Y$ is the variable of interest, $C$ a censoring variable and $\mathbf{X}$ a concomitant variable. Throughout, we will use Maillot and Viallon (2009) notation and we work with a sample $\{(Y_i, C_i, \mathbf{X}_i)_{1 \leq i \leq n}\}$ of independent and identically distributed replication of $(Y, C, \mathbf{X})$, $n \geq 1$. Actually, in the right censorship model, the pairs $(Y_i, C_i)$, $1 \leq i \leq n$, are not directly observed and the corresponding information is given by $Z_i := \min\{Y_i, C_i\}$ and $\delta_i := \mathbb{1}\{Y_i \leq C_i\}$, $1 \leq i \leq n$. Accordingly, the observed sample is

$$\mathcal{D}_n = \{(Z_i, \delta_i, \mathbf{X}_i), i = 1, \ldots, n\}.$$

Survival data in clinical trials or failure time data in reliability studies, for example, are often subject to such censoring. To be more specific, many statistical experiments result in incomplete samples, even under well-controlled conditions. For example, clinical data for surviving most types of disease are usually censored by other competing risks to life which result in death. In the sequel, we impose the following assumptions upon the distribution of $(\mathbf{X}, Y)$. Denote by $\mathbf{I}$

a given compact set in $\mathbb{R}^d$ with nonempty interior and set, for any $\alpha > 0$,

$$I_\alpha = \{\mathbf{x} : \inf_{\mathbf{u}\in\mathbf{I}} \|\mathbf{x} - \mathbf{u}\| \leq \alpha\}.$$

We will assume that, for a given $\alpha > 0$, $(\mathbf{X}, Y)$ [resp. $\mathbf{X}$] has a density function $f_{\mathbf{X},Y}$ [resp. $f_{\mathbf{X}}$] with respect to the Lebesgue measure on $I_\alpha \times \mathbb{R}$ [resp. $I_\alpha$]. For $-\infty < t < \infty$, set

$$F_Y(t) = \mathbb{P}(Y \leq t), \;\; G(t) = \mathbb{P}(C \leq t), \;\; \text{and} \;\; H(t) = \mathbb{P}(Z \leq t),$$

the right-continuous distribution functions of $Y$, $C$ and $Z$ respectively. For any right-continuous distribution function $L$ defined on $\mathbb{R}$, denote by

$$T_L = \sup\{t \in \mathbb{R} : L(t) < 1\}$$

the upper point of the corresponding distribution. Now consider a pointwise measurable class $\mathcal{F}$ of real measurable functions defined on $\mathbb{R}$, and assume that $\mathcal{F}$ is of VC-type.

## 4.1 Definition of the I.P.C.W estimators

In this thesis, we will mostly focus on the regression function of $\psi(Y)$ evaluated at $\mathbf{X} = \mathrm{x}$, for $\psi \in \mathcal{F}$ and $\mathrm{x} \in I_\alpha$, given by

$$m_\psi(\mathbf{x}) = \mathbb{E}(\psi(Y) \mid \mathbf{X} = \mathbf{x}),$$

when $Y$ is right-censored. To estimate $m_\psi(\cdot)$,we make use of the Inverse Probability of Censoring Weighted (I.P.C.W.) estimators which have recently gained popularity in the censored data literature (see Kohler *et al.* (2002), Carbonez *et al.* (1995), Brunel and Comte (2006)). The key idea of I.P.C.W. estimators is as follows. Introduce the real-valued function $\Phi_\psi(\cdot,\cdot)$ defined on $\mathbb{R}^2$ by

$$\Phi_\psi(y, c) = \frac{\mathbb{1}\{y \leq c\}\psi(y \wedge c)}{1 - G(y \wedge c)}. \tag{4.1.1}$$

Assuming the function $G(\cdot)$ to be known, first note that $\Phi_\psi(Y_i, C_i) = \delta_i\psi(Z_i)/(1 - G(Z_i))$ is observed for every $1 \leq i \leq n$. Moreover, under the Assumption **(I)** below,

**(I)** $C$ and $(Y, \mathbf{X})$ are independent.

We have

$$\begin{aligned} m_{\Phi_\psi}(\mathbf{x}) &:= \mathbb{E}(\Phi_\psi(Y,C) \mid \mathbf{X}=\mathbf{x}) \\ &= \mathbb{E}\left\{\frac{\mathbb{1}\{Y \le C\}\psi(Z)}{1-G(Z)} \mid \mathbf{X}=\mathbf{x}\right\} \\ &= \mathbb{E}\left\{\frac{\psi(Y)}{1-G(Y)}\mathbb{E}(\mathbb{1}\{Y \le C\} \mid \mathbf{X}, Y) \mid \mathbf{X}=\mathbf{x}\right\} \\ &= m_\psi(\mathbf{x}). \end{aligned} \tag{4.1.2}$$

Therefore, any estimate of $m_{\Phi_\psi}(\cdot)$, which can be built on fully observed data, turns out to be an estimate for $m_\psi(\cdot)$ too. Thanks to this property, most statistical procedures known to provide estimates of the regression function in the uncensored case can be naturally extended to the censored case. For instance, kernel-type estimates are particularly easy to construct. Set, for $\mathbf{x} \in \mathbf{I}$, $h \ge l_n$, $1 \le i \le n$,

$$\overline{\omega}_{n,K,h,i}(\mathbf{x}) := K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right). \tag{4.1.3}$$

We assume that $h$ satisfies **(H.1)**. In view of (4.1.1), (4.1.2), and (4.1.3), whenever $G(\cdot)$ is known, a kernel estimator of $m_\psi(x)$ is given by

$$\widetilde{m}_{\psi,n,h}(\mathbf{x}) = \sum_{i=1}^{n} \overline{\omega}_{n,K,h,i}(\mathbf{x}) \frac{\delta_i \psi(Z_i)}{1-G(Z_i)}. \tag{4.1.4}$$

The function $G(\cdot)$ is generally unknown and has to be estimated. We will denote by $G_n^*(\cdot)$ the Kaplan-Meier estimator of the function $G(\cdot)$ Kaplan and Meier (1958). Namely, adopting the conventions

$$\prod_{\emptyset} = 1$$

and $0^0 = 1$ and setting

$$N_n(u) = \sum_{i=1}^{n} \mathbb{1}\{Z_i \ge u\},$$

we have

$$G_n^*(u) = 1 - \prod_{i:Z_i \le u} \left\{\frac{N_n(Z_i)-1}{N_n(Z_i)}\right\}^{(1-\delta_i)}, \quad \text{for} \quad u \in \mathbb{R}.$$

Given these notation, we will investigate the following estimator of $m_\psi(\mathbf{x})$

$$\widetilde{m}^*_{\psi,n,h}(\mathbf{x}) = \sum_{i=1}^{n} \overline{\omega}_{n,K,h,i}(\mathbf{x}) \frac{\delta_i \psi(Z_i)}{1-G_n^*(Z_i)}, \tag{4.1.5}$$

refer to Kohler *et al.* (2002) and Maillot and Viallon (2009). Adopting the convention $0/0 = 0$, this quantity is well defined, since $G_n^*(Z_i) = 1$ if and only if $Z_i = Z_{(n)}$ and $\delta_{(n)} = 0$, where $Z_{(k)}$ is the $k$th ordered statistic associated with the sample $(Z_1, \ldots, Z_n)$ for $k = 1, \ldots, n$ and $\delta_{(k)}$ is the $\delta_j$ corresponding to $Z_k = Z_j$. When the variable of interest is right-censored, functionals of the (conditional) law can generally not be estimated on the complete support (see Brunel and Comte (2006)).

## 4.2 Assumptions

In order to obtain our results, we will work under the following assumptions.

**(A.1)** $\mathcal{F} = \{\psi := \psi_1 \mathbb{1}\{(-\infty, \tau)\}, \psi_1 \in \mathcal{F}_1\}$, where $\tau < T_H$ et $\mathcal{F}_1$ is a pointwise measurable class of real measurable functions defined on $\mathbb{R}$ and of type VC.

**(A.2)** The class of functions $\mathcal{F}$ has a measurable and uniformly bounded envelope function $\Upsilon$ with,

$$\Upsilon(y) \geq \sup_{\psi \in \mathcal{F}} \mid \psi(y) \mid, \quad y \leq T_H.$$

**(A.3)** The class of functions $\mathcal{M} := \left\{ \frac{m_\psi}{f_{\mathbf{X}}}, \psi \in \mathcal{F} \right\}$ is relatively compact with respect to the sup- norm topology on $I_\alpha$.

In what follows, we will study the uniform convergence of $\widetilde{m}^*_{\psi,n,h}(\mathbf{x})$ centered by the following centering factor

$$\widehat{\mathbb{E}} m_{\psi;n}(\mathbf{x}; h) = \frac{\mathbb{E}\left(\psi(Y) K\left(\dfrac{\mathbf{x} - \mathbf{X}}{h}\right)\right)}{\mathbb{E}\left(K\left(\dfrac{\mathbf{x} - \mathbf{X}}{h}\right)\right)}.$$

This choice is justified by the fact that, under hypothesis **(I)** we have

$$\begin{aligned} \mathbb{E}\left\{\Phi_\psi(Y, C) K\left(\frac{\mathbf{x} - \mathbf{X}}{h}\right)\right\} &= \mathbb{E}\left\{\frac{\mathbb{1}\{Y \leq C\}\psi(Z)}{1 - G(Z)} K\left(\frac{\mathbf{x} - \mathbf{X}}{h}\right)\right\} \qquad (4.2.1) \\ &= \mathbb{E}\left\{\frac{\psi(Y) K\left(\dfrac{\mathbf{x} - \mathbf{X}}{h}\right)}{1 - G(Y)} \mathbb{E}[\mathbb{1}\{Y \leq C\} \mid \mathbf{X}, \mathbf{Y}]\right\} \\ &= \mathbb{E}\left\{\psi(Y) K\left(\frac{\mathbf{x} - \mathbf{X}}{h}\right)\right\}. \end{aligned}$$

## 4.3 Results

We have now all the ingredients to state the result corresponding to the censored case.

**Theorem 4.3.0.1** *Under assumptions* **(A.1-3)**, **(I)**, *Assumption 1, assume that $h$ satisfies* **(H.1-H.3)** *and for any kernel $K(\cdot)$ satisfying Assumptions 3 and 4, with probability at least $1-\delta$,*

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\left|\widetilde{m}^*_{\psi,n,h}(\mathbf{x})-\widehat{\mathbb{E}}(m_{\psi,n}(\mathbf{x};h))\right|\leq C'\sqrt{\frac{\log\left(1/l_n\right)_+ +\log\left(2/\delta\right)}{nl_n^{2d-d_{\text{vol}}+\varepsilon}}}. \tag{4.3.1}$$

## Proof of Theorem 4.3.0.1

In the following proposition we show that Theorem 4.3.0.1 naturally follows from Corollary 3.1.0.5. We first establish the version of Theorem 4.3.0.1 corresponding to the case where $G(\cdot)$ is known (i.e., with $\widetilde{m}^*_{\psi,n,h}$ replaced by $\widetilde{m}_{\psi,n,h}$). To complete the proof of Theorem 4.3.0.1, the consistency of the Kaplan-Meier estimator will be helpful (see Lemma 4.3.0.3 below)

**Proposition 4.3.0.2** *Under assumptions* **(A.1- 3)**, **(I)**, *Assumption 1, assume that $h$ satisfies* **(H.1-3)** *and for any kernel $K(\cdot)$ satisfying Assumptions 3 and 4, with probability at least $1-\delta$*

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\psi\in\mathcal{F}}\left|\widetilde{m}_{\psi,n,h}(\mathbf{x})-\widehat{\mathbb{E}}m_{\psi}(\mathbf{x};h)\right|\leq C\sqrt{\frac{\log\left(1/l_n\right)_+ +\log\left(2/\delta\right)}{nl_n^{2d-d_{\text{vol}}+\varepsilon}}} \tag{4.3.2}$$

**Proof.** Recalling the definition 4.1.1 of $\Phi_\psi$

$$\Phi_\psi(y,c)=\frac{\mathbb{1}\{y\leq c\}\psi(y\wedge c)}{1-G(y\wedge c)}. \tag{4.3.3}$$

it is obvious that $\Phi_\psi$ is uniformly bounded, in $(y,c)\in\mathbb{R}^2$ and $\psi\in\mathcal{F}$, since $\mathcal{F}$ is uniformly bounded, $\psi(t)=0$ for all $t>\tau$ and $G(\tau)<1$. This property, when combined with the VC property of $\mathcal{F}_1$, ensures that the class of function

$$\mathcal{F}_\Phi:=\{\Phi_\psi:\psi\in\mathcal{F}\}$$

verifies (F.i), (F.iii). Similarly, it can be shown that $\mathcal{F}_\Phi$ is a pointwise measurable class of functions (F.ii). Moreover, by **(A.3)** and (4.1.2), the class

$$\mathcal{M}_\Phi:=\{m_{\Phi_\psi}\mid f_{\mathbf{X}},\psi\in\mathcal{F}_1\}$$

is almost surely relatively compact with respect to the sup- norm topology on $I_\alpha$. So we can apply Corollary 3.1.0.5 with $\mathbf{Y}=(Y,C)$ and $\Psi=\Phi_\psi$. The result of Proposition 4.3.0.2 is straightforward. ■

To complete the demonstration of Theorem 4.3.0.1, we will use the result of the next approximation Lemma 4.3.0.3.

**Lemma 4.3.0.3** *Under assumptions of Theorem 4.3.0.1, we have with probability one,*

$$\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\psi\in\mathcal{F}}\left|\widetilde{m}_{\psi,n,h}(\mathbf{x})-\widetilde{m}^*_{\psi,n,h}(\mathbf{x})\right| = o\left(\sqrt{\frac{\log(1/h)}{nh^d}}\right) \quad \textit{as} \quad n\rightarrow\infty. \tag{4.3.4}$$

**Proof.**

Notice that we have

$$\begin{aligned}
&\sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\psi\in\mathcal{F}}\left|\widetilde{m}_{\psi,n,h}(\mathbf{x})-\widetilde{m}^*_{\psi,n,h}(\mathbf{x})\right| \\
&\quad = \sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sup_{\psi\in\mathcal{F}}\left|\sum_{i=1}^{n}\overline{\omega}_{n,K,h,i}(\mathbf{x})\delta_i\psi(Z_i)\left(\frac{1}{1-G(Z_i)}-\frac{1}{1-G^*_n(Z_i)}\right)\right| \\
&\quad \leq \sup_{h\geq l_n}\sup_{\mathbf{x}\in\mathbf{I}}\sum_{i=1}^{n}|\overline{\omega}_{n,K,h,i}(\mathbf{x})|\sup_{t\leq\tau}\frac{\sup_{\psi\in\mathcal{F}}|\psi(t)|}{[1-G^*_n(\tau)][1-G(\tau)]}\sup_{t\leq\tau}|G^*_n(t)-G(t)|.
\end{aligned} \tag{4.3.5}$$

Since

$$\sup_{\psi\in\mathcal{F}}|\psi(t)|<\infty,$$

the kernel $K(\cdot)$ is uniformly bounded and

$$\tau<T_H=T_F\leq T_G,$$

the law of iterated logarithm for $G^*_n(\cdot)$ established in Földes and Rejtő (1981) ensures that

$$\sup_{t\leq\tau}|G^*_n-G(t)|=O\left(\sqrt{\frac{\log\log n}{n}}\right) \qquad \text{almost surely as} \quad n\rightarrow\infty.$$

By combining the results of Proposition 4.3.0.2 and Lemma 4.3.0.3, the result of the Theorem 4.3.0.1 is immediate by noting that, under the conditions **(H.1-3)**, we have, for $n$ sufficiently large,

$$\sup_{t\leq\tau}|G^*_n-G(t)|=o\left(\sqrt{\frac{\log(1/h)}{nh^d}}\right) \qquad \text{almost surely as} \quad n\rightarrow\infty.$$

Hence the proof is complete. ■

**Kernel estimator of the conditional distribution function in the censored case**

We will show how Theorem 4.3.0.1 can be used **(A.2)** to establish the uniform in bandwidth consistency for an estimator for the conditional distribution function. Towards this aim, we introduce the following quantities:

$$\widetilde{F}^*_{h,n}(t \mid \mathbf{x}) := \sum_{i=1}^{n} \overline{\omega}_{n,K,h,i}(\mathbf{x}) \frac{\delta_i \mathbb{1}\{Z_i \leq t\}}{1 - G_n^*(Z_i)} \text{ and}$$

$$\widehat{\mathbb{E}}F_{h,n}(t \mid \mathbf{x}) = \frac{\mathbb{E}\left[\mathbb{1}\{Y \leq t\} K\left((\mathbf{x} - \mathbf{X})/h\right)\right]}{\mathbb{E}K\left((\mathbf{x} - \mathbf{X})/h\right)}.$$

Rates of convergence for $\widetilde{F}^*_{h,n}(t \mid \mathbf{x})$ can be obtained under weaker conditions, when restricting ourselves to $t \in [-\infty, \tau_0]$ with $\tau_0 < T_H$. On the other hand, it is noteworthy that **(A.2-3)** are automatically fulfilled for the particular choice

$$\mathcal{F} = \{\mathbb{1}(-\infty, t] : t \leq T_H\}.$$

This property will enable us to easily describe uniform consistency for estimators of the conditional distribution function $F(t \mid \cdot) := \mathbb{P}(Y \leq t \mid \mathbf{X} = \cdot)$ over $t \in (-\infty, T_H)$.

**Corollary 4.3.0.4** *Under assumptions* ***(A.1)****,* ***(I)*** *and Assumption 1, assume that $h$ satisfies* ***(H.1-3)*** *and for any kernel $K(\cdot)$ satisfying Assumptions 3 and 4, we have, for all $\tau_0 < T_H$, with probability at least $1 - \delta$,*

$$\sup_{h \geq l_n} \sup_{\mathbf{x} \in \mathbb{X}} \sup_{t \leq \tau_0} \left| \widetilde{F}^*_{h,n}(t \mid \mathbf{x}) - \widehat{\mathbb{E}}F_{h,n}(t \mid \mathbf{x}) \right| \leq C \sqrt{\frac{\log\left(1/l_n\right)_+ + \log\left(2/\delta\right)}{n l_n^{2d - d_{\mathrm{vol}} + \varepsilon}}}. \tag{4.3.6}$$

**Proof.** Corollary 4.3.0.4 being a direct consequence of Theorem 4.3.0.1 with

$$\mathcal{F} = \{\mathbb{1}\{[0, t]\} : t < \tau_0 < T_H\},$$

details of its proof are omitted. ■

### Kernel estimator of the conditional density function in censored case

To establish the uniform in bandwidth consistency for the estimates of the conditional density, we denote by $h^{'}$ an additional bandwidth. As for the conditional density $f(t \mid \mathbf{x}) := f_{\mathbf{X},Y}(\mathbf{x}, t)/f_{\mathbf{X}}(\mathbf{x})$, we consider the following estimator:

$$\widetilde{f}^*_{h,h^{'},n}(t \mid \mathbf{x}) := \sum_{i=1}^{n} \overline{\omega}_{n,K,h,i}(\mathbf{x}) \frac{\delta_i \mathbb{1}\{Z_i \in [t - \frac{h^{'}}{2}; t + \frac{h^{'}}{2}]\}}{h^{'}[1 - G_n^*(Z_i)]},$$

and the corresponding centering term,

$$\widehat{\mathbb{E}}f_{h,h^{'},n}(t \mid \mathbf{x}) := \frac{\mathbb{E}\left[\mathbb{1}\{Y \in [t - \frac{h^{'}}{2}; t + \frac{h^{'}}{2}]\} K\left((\mathbf{x} - \mathbf{X})/h\right)\right]}{h^{'}\mathbb{E}K\left((\mathbf{x} - \mathbf{X})/h\right)}.$$

**Corollary 4.3.0.5** *Under assumptions **(A.1)**, **(I)**, Assumption 1, assume that $h$ satisfies **(H.1-3)** and for any kernel $K(\cdot)$ satisfying Assumptions 3 and 4, we have, for all $\tau_0 < T_H$ and for $\{h_n^{'}\}_{n\geq 1}$ a sequence of positive constants such that $h_n^{'} \geq \sqrt{\frac{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}{\log(1/l_n)_+ + \log(2/\delta)}}$, with probability at least $1-\delta$,*

$$\sup_{h\geq l_n}\sup_{h'\geq h_n'}\sup_{\mathbf{x}\in\mathbb{X}}\sup_{t\leq\tau_0}\left|\widetilde{f}^*_{h,h',n}(t\mid\mathbf{x}) - \widehat{\mathbb{E}}f_{h,h',n}(t\mid\mathbf{x})\right| \leq C'\sqrt{\frac{\log(1/l_n)_+ + \log(2/\delta)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{4.3.7}$$

**Proof.**

Since

$$h_n^{'} \geq \sqrt{\frac{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}{\log(1/l_n)_+ + \log(2/\delta)}},$$

we have for any $s > 0$,

$$\begin{aligned}&\sup_{h\geq l_n}\sup_{h'\geq h_n'}\sup_{\mathbf{x}\in\mathbb{X}}\sup_{t\leq\tau_0}\left|\widetilde{f}^*_{h,h',n}(t\mid\mathbf{x}) - \mathbb{E}f_{h,h',n}(t\mid\mathbf{x})\right|\\ &\leq \frac{1}{h_n^{'}}\sup_{h\geq l_n}\sup_{h'\leq s}\sup_{\mathbf{x}\in\mathbb{X}}\sup_{t\leq\tau_0} h^{'}\left|\widetilde{f}^*_{h,h',n}(t\mid\mathbf{x}) - \mathbb{E}f_{h,h',n}(t\mid\mathbf{x})\right|.\end{aligned}$$

Set $s\in(0, T_H-\tau_0)$. By applying Theorem 4.3.0.1 with

$$\mathcal{F}_s = \left\{\mathbb{1}\left\{\left[t-\frac{h^{'}}{2}, t+\frac{h^{'}}{2}\right]\right\}, t<\tau_0<T_H\right\},$$

we have with probability at least $1-\delta$,

$$\sup_{h\geq l_n}\sup_{h'\geq h_n'}\sup_{\mathbf{x}\in\mathbb{X}}\sup_{t\leq\tau_0}\left|\widetilde{f}^*_{h,h',n}(t\mid\mathbf{x}) - \mathbb{E}f_{h,h',n}(t\mid\mathbf{x})\right| \leq C'\sqrt{\frac{\log(1/l_n)_+ + \log(2/\delta)}{nl_n^{2d-d_{\mathrm{vol}}+\varepsilon}}}. \tag{4.3.8}$$

Hence the proof is complete. ■

# Conclusions and perspectives

In this thesis, we have used general methods based upon empirical process techniques to prove uniform in bandwidth consistency for kernel-type function estimators. We have considered an extended setting when dimension can be much lower than the ambient dimension. In addition, our work complements the paper Kim *et al.* (2018) by considering other examples of kernel estimates. Examples include regression function, conditional distribution, mode and Shannon's entropy. We have investigated the general kernel estimators in the framework of censored data. The results allow data-driven local bandwidths.
To complete this thesis we raise some perspectives that may be the object of future works:

- Our results are especially useful to establish uniform consistency of data-driven bandwidth kernel-type function estimators. The interest of doing so would be to extend our work to $k$-nearest neighbours estimators. Presently it is beyond reasonable hope to achieve this program without new technical arguments.

- An other direction of research is to consider the projection pursuit regression and projection pursuit conditional distribution, which need an extension and generalization of the methods used in the present thesis. If we assume that the regression function $m_{\Psi}(\cdot)$ is smooth enough, that is $p+1$ times differentiable at a fixed $\mathrm{x}_0$, it will be better to use the local polynomial regression techniques, refer to Fan and Gijbels (1996), to obtain a more appropriate estimate at $\mathrm{x}_0$ than that given by the Nadaraya-Watson estimator. The uniform consistency of such estimators will be treated in future investigation.

- We study this model with other types of censored data (left censorship, double censorship, mixed censorship). To do this, we define a general censoring framework that encompasses all censoring models and we show uniform in bandwidth consistency for kernel-type density and regression function estimators.

- As a generalization of Nadaraya-Watson estimates of a regression function and using the same concept of the present work we use a conditional $U$-statistics and we apply the methods developed in Dony and Mason (2008) and Bouzebda and Nemouchi (2020) to

$$m(\mathbf{t}) := \mathbb{E}[\varphi(Y_1, \ldots, Y_m)|(X_1, \ldots, X_m) = \mathbf{t}], \text{ for } \mathbf{t} \in \mathbb{R}^{dm}.$$

  to establish uniform in **t** and in bandwidth consistency (i.e., $h_n$, $h_n \in [a_n, b_n]$ where $0 < a_n < b_n \to 0$ at some specific rate). This work is in progress.

- In the traditional kernel methods for curve estimation, it has been widely regarded that the performance of the kernel methods depends largely on the smoothing bandwidth, and depends very little on the form of the kernel. Most kernels used are symmetric kernels and, once chosen, are fixed. This may be efficient for estimating curves with unbounded supports, but not for curves which have compact support or subset of the whole real line and are discontinuous at boundary points. A great advantage of the wavelet methods in statistics is to provide adaptive procedures in the sense that they automatically adapt to the regularity of the object to be estimated. Another remarkable advantage of the wavelet procedures is that they can be very easily used. For future investigations, we will extend our results to the wavelet estimators.

- We have treated the uniform convergence in both cases when the class of functions is bounded or unbounded satisfying some moment conditions. It would be of interest to complete our investigation by considering the weak dependence. A natural question arises is, how to relax the dependence assumption on the sequence of r.v. to cope with more general framework by considering the weak dependence or by assuming only the ergodicity.

# Appendix A

## A.1 Uniform convergence on a function class

As discussed in chapter 2 we combine Talagrand inequality (Bousquet (2002), Steinwart and Christmann (2008)) and (Sriperumbudur and Steinwart (2012)) VC type bound to obtain our result. We derive a uniform convergence for a more general class of functions (for more details see Kim *et al.* (2018)).

**Theorem A.1.0.1** *Let $(\mathbb{R}^d, \mathbb{P})$ be a probability space and let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. from $\mathbb{P}$. Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that is uniformly bounded VC-class with dimension $\nu$, i.e., there exists positive numbers $A$, $B$ such that, for all $f \in \mathcal{F}$, $\|f\|_\infty \leq B$, and for every probability measure $\mathbb{Q}$ on $\mathbb{R}^d$ and for every $\epsilon \in (0, B)$, the covering number $\mathcal{N}(\mathcal{F}, L_2(\mathbb{Q}), \epsilon)$ satisfies*

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

*Let $\sigma > 0$ with $\mathbb{E}_\mathbb{P} f^2 \leq \sigma^2$ for all $f \in \mathcal{F}$. Then there exists a universal constant $C$ not depending on any parameters such that*

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}_i) - \mathbb{E}[f(\mathbf{X})] \right|$$

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

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

Let $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d from a probability space $(\mathcal{H}, \mathcal{A}, \mathbb{P})$ with comment distribution. Let $\mathcal{G}$ be a pointwise measurable class a real valued functions defined on

$\mathcal{H}$. Further let $\varepsilon_1, \ldots, \varepsilon_n$ be a sequence of independent Rademacher random variables independent of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Let $G(\cdot)$ be a finite-valued measurable function satisfying for all $\mathbf{x} \in \mathcal{H}$,

$$G(\mathbf{x}) \geq \sup_{g \in \mathcal{G}} |g(\mathbf{x})|,$$

and for some $\nu > 0$, $C < \infty$

$$\mathcal{N}(\varepsilon, \mathcal{G}) \leq C\varepsilon^{-\nu}, 0 < \varepsilon < 1;$$

with

$$\mathcal{N}(\varepsilon, \mathcal{G}) = \sup_{\mathbb{Q}} \mathcal{N}(\varepsilon\sqrt{\mathbb{Q}(G^2)}, \mathcal{G}, d\mathbb{Q});$$

where the supremum is taken over all probability measures $\mathbb{Q}$ on $(\mathcal{H}, \mathcal{A})$ for which $0 < \int G^2 d\mathbb{Q} < \infty$ and $d\mathbb{Q}$ is the $L_2$-metric. In our proofs, we make frequent use of the following moment bound, established in Proposition 1 of Einmahl *et al.* (2005).

## A.2 Empirical processes tools

Recall the definitions introduced from empirical process theory in the second section. Keep in mind that the class $\mathcal{G}$ denotes a generic class of functions with envelope function $G(\cdot)$.

**Proposition A.2.0.1** *(Proposition 1 of (Einmahl* et al.*, 2005))*
*Let $\mathcal{G}$ be a pointwise measurable class of bounded functions such that for some constants $C, \nu \geq 1$ and $0 < \sigma \leq \beta$ and $G(\cdot)$ as above, the following conditions hold:*

**(i)** $\mathbb{E}\left[G(\mathbf{X})^2\right] \leq \beta^2$;

**(ii)** $\mathcal{N}(\varepsilon, \mathcal{G}) \leq C\varepsilon^{-\nu}, 0 < \varepsilon < 1$;

**(iii)** $\sigma_0^2 := \sup_{g \in \mathcal{G}} \mathbb{E}\left[g(\mathbf{X})^2\right] \leq \sigma^2$;

**(iv)**

$$\sup_{g \in \mathcal{G}} \|g\|_\infty \leq \frac{1}{4\sqrt{\nu}} \sqrt{n\sigma^2 / \log(C_1\beta/\sigma)}, \text{ where } C_1 = C^{1/\nu} \vee e.$$

*Then we have for some absolute constant $A$,*

$$\mathbb{E}\left\|\sum_{i=1}^{n} \varepsilon_i g(\mathbf{X}_i)\right\|_{\mathcal{G}} \leq A\sqrt{\nu n\sigma^2 \log(C_1\beta/\sigma)}. \tag{A.2.1}$$

**Corollary A.2.0.2** *Let $\mathcal{G}$ be as in Proposition A.2.0.1* ***(i)**-**(iii)**, and instead of* ***(iv)*** *assume that*

**(v)** $\sup_{g\in\mathcal{G}} \|g\|_\infty \le U$, *where* $\sigma_0 \le U \le C_2\sqrt{n}\beta$, *and* $C_2 = \frac{1}{4\sqrt{\nu\log C_1}}$.

*Then we have*

$$\mathbb{E}\left\|\sum_{i=1}^{n} \varepsilon_i g(\mathbf{X}_i)\right\|_{\mathcal{G}} \le A\{\sqrt{\nu n\sigma_0^2\log(C_1\beta/\sigma_0)} + 2\nu U\log(C_3 n(\beta/U)^2)\}. \tag{A.2.2}$$

## A.3 Talagrand's inequality

The following inequality, which is essentially due to Talagrand (1994) (see also Ledoux (1997)), is crucial for our work. Let $\alpha_n$ be the empirical process based on the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, that is, if $g : \mathcal{G} \to \mathbb{R}$, we have

$$\alpha_n(g) = \sum_{i=1}^{n}(g(\mathbf{X}_i) - \mathbb{E}g(\mathbf{X}))/\sqrt{n},$$

and set for any class $\mathcal{G}$ of such functions

$$\|\sqrt{n}\alpha_n\|_{\mathcal{G}} = \sup \|\sqrt{n}\alpha_n(g)\|.$$

**Theorem A.3.0.1 (?'s inequality)** *Let $\mathcal{G}$ be a pointwise measurable class of functions satisfying, for some $0 < M < \infty$,*

$$\|g\|_\infty \le M, \qquad g \in \mathcal{G}.$$

*Then we have for all $t > 0$,*

$$\mathbb{P}\left\{\max_{1\le m\le n}\|\sqrt{m}\alpha_m\|_{\mathcal{G}} \ge A_1\left(\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i g(\mathbf{X}_i)\right\|_{\mathcal{G}} + t\right)\right\} \le 2\left\{\exp\left(-\frac{A_2 t^2}{n\sigma_{\mathcal{G}}^2}\right) + \exp\left(-\frac{A_2 t}{M}\right)\right\},$$

*where*

$$\sigma_{\mathcal{G}^2} = \sup_{g\in\mathcal{G}} \mathrm{Var}(g(\mathbf{X}))$$

*and $A_1, A_2$ are universal constants.*

# Bibliography

Akaike, H. (1954). An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, **6**(2), 127–132.

Auestad, B. r. and Tjø stheim, D. (1991). Functional identification in nonlinear time series. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 493–507. Kluwer Acad. Publ., Dordrecht.

Bosq, D. (1987). La statistique non paramétrique des processus. *Rend. Sem. Mat. Univ. Politec. Torino*, **45**(2), 1–24 (1988).

Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, **334**(6), 495–500.

Bouzebda, S. (2012). On the strong approximation of bootstrapped empirical copula processes with applications. *Mathematical Methods of Statistics*, **21**(3), 153–188.

Bouzebda, S. and Chokri, K. (2014). Statistical tests in the partially linear additive regression models. *Statistical Methodology*, **19**, 4–24.

Bouzebda, S. and Didi, S. (2017). Additive regression model for stationary and ergodic continuous time processes. *Communications in Statistics-Theory and Methods*, **46**(5), 2454–2493.

Bouzebda, S. and El-hadjali, T. (2020). Uniform convergence rate of the kernel regression estimator adaptive to intrinsic dimension in presence of censored data. *J. Nonparametr. Stat.*, **32**(4), 864–914.

Bouzebda, S. and Elhattab, I. (2009). A strong consistency of a nonparametric estimate of entropy under random censorship. *C. R. Math. Acad. Sci. Paris*, **347**(13-14), 821–826.

Bouzebda, S. and Elhattab, I. (2011). Uniform-in-bandwidth consistency for kernel-type estimators of shannon's entropy. *Electronic journal of statistics*, **5**, 440–459.

Bouzebda, S. and Nemouchi, B. (2020). Uniform consistency and uniform in bandwidth consistency for nonparametric regression estimates and conditional $U$-statistics involving functional data. *J. Nonparametr. Stat.*, **32**(2), 452–509.

Bouzebda, S., Chokri, K., and Louani, D. (2016). Some uniform consistency results in the partially linear additive model components estimation. *Communications in Statistics-Theory and Methods*, **45**(5), 1278–1310.

Bouzebda, S., Elhattab, I., and Seck, C. T. (2018). Uniform in bandwidth consistency of nonparametric regression based on copula representation. *Statistics & Probability Letters*, **137**, 173–182.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, **80**(391), 580–598.

Brunel, E. and Comte, F. (2006). Nonparametric adaptive regression estimation in presence of censoring.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, **4**(421-424).

Carbonez, A., Györfi, L., and van der Meulen, E. C. (1995). Partitioning-estimates of a regression function under random censoring. *Statistics & Risk Modeling*, **13**(1), 21–38.

Chacón, J. E. and Duong, T. (2018). *Multivariate kernel smoothing and its applications*. CRC Press.

Chambers, J., Hastie, T., and Pregibon, D. (1990). Statistical models in s. In K. Momirović and V. Mildner, editors, *Compstat*, pages 317–321, Heidelberg. Physica-Verlag HD.

Chen, R., Härdle, W., Linton, O. B., and Severance-Lossin, E. (1996). Nonparametric estimation of additive separable regression models. In *Statistical Theory and Computational Aspects of Smoothing*, pages 247–265. Springer.

Chung, K.-L. (1949). An estimate concerning the kolmogroff limit distribution. *Transactions of the American Mathematical Society*, **67**(1), 36–50.

Collomb, G. (1981). Estimation non-paramétrique de la régression: revue bibliographique. *International Statistical Review/Revue Internationale de Statistique*, pages 75–93.

Cover, T. M. (2006). Elements of information theory.

Csiszár, I. (1962). Informationstheoretische konvergenzbegriffe im raum der wahrscheinlichkeitsverteilungen. *Publications of the Mathematical Institute, Hungarian Academy of Sciences, VII, Series A*, pages 137–157.

Csörgő, M. and Révész, P. (1981). *Strong approximations in probability and statistics*. Probability and Mathematical Statistics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London.

Deheuvels, P. (1974). Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. *CR Acad. Sci. Paris*, **278**, 1217–1220.

Deheuvels, P. (1991). Laws of the iterated logarithm for density estimators. In *Nonparametric Functional Estimation and Related Topics*, pages 19–29. Springer.

Deheuvels, P. (1992). Functional laws of the iterated logarithm for large increments of empirical and quantile processes. *Stochastic processes and their applications*, **43**(1), 133–163.

Deheuvels, P. (2000). Uniform limit laws for kernel density estimators on possibly unbounded intervals. In *Recent advances in reliability theory*, pages 477–492. Springer.

Deheuvels, P. (2011). One bootstrap suffices to generate sharp uniform bounds in functional estimation. *Kybernetika*, **47**(6), 855–865.

Deheuvels, P. and Mason, D. M. (1992). Functional laws of the iterated logarithm for the increments of empirical and quantile processes. *The Annals of Probability*, pages 1248–1287.

Deheuvels, P. and Mason, D. M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Statistical Inference for Stochastic Processes*, **7**(3), 225.

Devroye, L. (1987). *A course in density estimation*. Birkhauser Boston Inc.

Devroye, L. and Györfi, L. (1985). *Nonparametric density estimation*. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons, Inc., New York. The $L_1$ view.

Devroye, L. and Lugosi, G. (2001). Bcombinatorial methods in density estimation,ˆ springer. *Berlin Heidelberg New York*.

Donsker, M. D. (1951). An invariance principle for certain probability linit theorems. AMS.

Dony, J. and Mason, D. M. (2008). Uniform in bandwidth consistency of conditional $U$-statistics. *Bernoulli*, **14**(4), 1108–1133.

Dony, J., Einmahl, U., *et al.* (2009). Uniform in bandwidth consistency of kernel regression estimators at a fixed point. In *High dimensional probability V: The Luminy volume*, pages 308–325. Institute of Mathematical Statistics.

Dudewicz, E. J. and Van Der Meulen, E. C. (1981). Entropy-based tests of uniformity. *Journal of the American Statistical Association*, **76**(376), 967–974.

Dudley, R. M. (2014). *Uniform central limit theorems*, volume 142. Cambridge university press.

Ebrahimi, N., Habibullah, M., and Soofi, E. S. (1992). Testing exponentiality based on kullback-leibler information. *Journal of the Royal Statistical Society: Series B (Methodological)*, **54**(3), 739–748.

Eggermont, P. P. B., LaRiccia, V. N., and LaRiccia, V. (2001). *Maximum penalized likelihood estimation*, volume 1. Springer.

Einmahl, U. and Mason, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability*, **13**(1), 1–37.

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

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Farahmand, A. M., Szepesvári, C., and Audibert, J.-Y. (2007). Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 265–272.

Federer, H. (1959). Curvature measures. *Transactions of the American Mathematical Society*, **93**(3), 418–491.

Finkelstein, H. *et al.* (1971). The law of the iterated logarithm for empirical distribution. *The Annals of Mathematical Statistics*, **42**(2), 607–615.

Földes, A. and Rejtő, L. (1981). A lil type result for the product limit estimator. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **56**(1), 75–86.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Computer Science and Scientific Computing. Academic Press, Inc., Boston, MA, second edition.

Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2013). Nonparametric inference for density modes. *arXiv preprint arXiv:1312.7567*.

Giné, E. and Guillou, A. (2002). 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.

Giné, E., Koltchinskii, V., and Sakhanenko, L. (2004). Kernel density estimators: convergence in distribution for weighted sup-norms. *Probability theory and related fields*, **130**(2), 167–198.

Giné, E., Nickl, R., *et al.* (2009). Uniform limit theorems for wavelet density estimators. *The Annals of Probability*, **37**(4), 1605–1646.

Giné, E., Sang, H., *et al.* (2013). On the estimation of smooth densities by strict probability densities at optimal rates in sup-norm. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 128–149. Institute of Mathematical Statistics.

Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, **4**, 92–99.

Gokhale, D. V. and Kullback, S. (1978). *The information in contingency tables*, volume 23. M. dekker.

Györfi, L., Klober, M., Krzyzak, A., and Walls, H. (2002). A distribution free theory of nonparametric regression springer.

Hall, P. (1984). Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **67**(2), 175–196.

Härdle, W. (1990). *Applied nonparametric regression*. Number 19. Cambridge university press.

Hardle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, pages 1465–1481.

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models, volume 43 of. *Monographs on statistics and applied probability*, **15**.

Hein, M. and Audibert, J.-Y. (2005). Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22nd international conference on Machine learning*, pages 289–296.

Jaynes, E. T. (1957). Information theory and statistical mechanics. ii. *Physical review*, **108**(2), 171.

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

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**(282), 457–481.

Kara-Zaitri, L., Laksaci, A., Rachdi, M., and Vieu, P. (2017). Uniform in bandwidth consistency for various kernel estimators involving functional data. *Journal of Nonparametric Statistics*, **29**(1), 85–107.

Kégl, B. (2003). Intrinsic dimension estimation using packing numbers. In *Advances in neural information processing systems*, pages 697–704.

Kim, J., Shin, J., Rinaldo, A., and Wasserman, L. (2018). Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. *arXiv*, pages arXiv–1810.

Kim, J., Shin, J., Rinaldo, A., and Wasserman, L. (2019). Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *International Conference on Machine Learning*, pages 3398–3407. PMLR.

Kohler, M., Máthé, K., and Pintér, M. (2002). Prediction from randomly right censored data. *Journal of Multivariate Analysis*, **80**(1), 73–100.

Kosorok, M. (2008). Introduction to empirical processes and semiparametric inference springer: New york.

Kullback, S. (1959). Information theory and statistics. john riley and sons. *Inc. New York*.

Lazo, A. V. and Rathie, P. (1978). On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory*, **24**(1), 120–122.

Ledoux, M. (1997). On talagrand's deviation inequalities for product measures. *ESAIM: Probability and statistics*, **1**, 63–87.

Levina, E. and Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, **17**, 777–784.

Ling, N. and Vieu, P. (2018). Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics*, **52**(4), 934–949.

Ling, N., Meng, S., and Vieu, P. (2019). Uniform consistency rate of k nn regression estimation for functional time series data. *Journal of Nonparametric Statistics*, **31**(2), 451–468.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pages 93–100.

Maillot, B. and Viallon, V. (2009). Uniform limit laws of the logarithm for nonparametric estimators of the regression function in presence of censored data. *Mathematical Methods of Statistics*, **18**(2), 159–184.

Mammen, E., Linton, O., and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, pages 1443–1490.

Mammen, E., Park, B. U., and Schienle, M. (2012). Additive models: extensions and related models. Technical report, SFB 649 Discussion Paper.

Mason, D. M. (2012). Proving consistency of non-standard kernel estimators. *Statistical inference for stochastic processes*, **15**(2), 151–176.

Mason, D. M. and Swanepoel, J. W. (2011). A general result on the uniform in bandwidth consistency of kernel-type function estimators. *Test*, **20**(1), 72–94.

Mason, D. M., Swanepoel, J. W., *et al.* (2015). Uniform in bandwidth consistency of kernel estimators of the density of mixed data. *Electronic Journal of Statistics*, **9**(1), 1518–1539.

Mokkadem, A. and Pelletier, M. (2003). The law of the iterated logarithm for the multivariate kernel mode estimator. *ESAIM: Probability and Statistics*, **7**, 1–21.

Müller, H.-G. (1988). Nonparametric regression analysis of longitudinal data.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9**(1), 141–142.

Nadaraya, E. A. (1989). *Nonparametric estimation of probability densities and regression curves*. Springer.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, pages 233–253.

Noh, Y.-K., Sugiyama, M., Liu, S., Plessis, M. C. d., Park, F. C., and Lee, D. D. (2018). Bias reduction and metric learning for nearest-neighbor estimation of kullback-leibler divergence. *Neural Computation*, **30**(7), 1930–1960.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799.

Novo, S., Aneiros, G., and Vieu, P. (2019). Automatic and location-adaptive estimation in functional single-index regression. *Journal of Nonparametric Statistics*, **31**(2), 364–392.

Opsomer, J. D., Ruppert, D., *et al.* (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, **25**(1), 186–211.

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

Pollard, D. (1984). Convergence of. *Stochastic Processes, New Yodc Springer-Veflag*.

Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR.

Prakasa Rao, B. L. S. (1983). *Nonparametric functional estimation*. Probability and Mathematical Statistics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York.

Rachdi, M. and Vieu, P. (2007). Nonparametric regression for functional data: automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*, **137**(9), 2784–2801.

Rényi, A. (1959). On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, **10**(1-2), 193–215.

Rosenb1att, M. (1956). Dremarks on some nonparametric estimates of a density functiond. *Annals of Mathematical Statistics*, **27**, 832J837.

Scott, D. W. (2015). *Multivariate density estimation*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition. Theory, practice, and visualization.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, **27**(3), 379–423.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Singh, R. S. (1977). Applications of estimators of a density and its derivatives to certain statistical problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(3), 357–363.

Sriperumbudur, B. and Steinwart, I. (2012). Consistency and rates for clustering with dbscan. In *Artificial Intelligence and Statistics*, pages 1090–1098.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Steinwart, I., Sriperumbudur, B. K., and Thomann, P. (2017). Adaptive clustering using kernel density estimators. *arXiv preprint arXiv:1708.05254*.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, pages 590–606.

Stute, W. (1982). A law of the logarithm for kernel density estimators. *Ann. Probab.*, **10**(2), 414–422.

Stute, W. (1984). The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab.*, **12**(2), 361–379.

Stute, W. (1986a). Conditional empirical processes. *The Annals of Statistics*, pages 638–647.

Stute, W. (1986b). Conditional empirical processes. *Ann. Statist.*, **14**(2), 638–647.

Stute, W. (1986c). On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.*, **14**(3), 891–901.

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76.

Tapia, R. A. and Thompson, J. R. (1978). Nonparametric probability density estimation.

Tjøstheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association*, **89**(428), 1398–1409.

Tsybakov, A. (1987). On the choice of bandwidth in nonparametric kernel regression. *Teor. Veroyatnost. i Primenen*, **32**, 153–159.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society: Series B (Methodological)*, **38**(1), 54–59.

Wand, M. P. and Jones, M. C. (1995). Kernel smoothing, volume 60 of. *Monographs on statistics and applied probability*.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Wertz, W. (1978). *Statistical density estimation: a survey*. Number 13. Vandenhoeck & Ruprecht.

Ziegler, K. (2002). On nonparametric kernel estimation of the mode of the regression function in the random design model. *Journal of Nonparametric Statistics*, **14**(6), 749–774.

Ziegler, K. (2003). On the asymptotic normality of kernel regression estimators of the mode in the nonparametric random design model. *Journal of Statistical Planning and Inference*, **115**(1), 123–144.

## Thesis abstract

In this thesis, we are concerned with the uniform in bandwidth consistency of kernel-type estimators of the regression function derived by modern empirical process theory, under weaker conditions on the kernel than previously used in the literature. Our theorems allow data-driven local bandwidths for these statistics. We extend existing uniform bounds on kernel type-estimator and making it adaptive to the intrinsic dimension of the underlying distribution, which will be characterising by the so-called intrinsic dimension. The thesis is divided in three main parts, we describe as follows. The first part is devoted to general empirical processes indexed by classes of functions. The results are obtained for uniformly bounded classes of functions or unbounded with envelope functions satisfying some moment conditions. The purpose of the second part is the statistical applications to illustrate the usefullness of the main contribution. Applications include the uniform in bandwidth consistency of the kernel type estimators for density, regression, the conditional distribution, multivariate mode, Shannon's entropy, derivatives of density and regression functions. The third part is devoted to the uniform in bandwidth consistency for non-parametric inverse probability of censoring weighted (I.P.C.W.) estimators of the regression function under random censorship. These new results are applied for the non-parametric conditional density and conditional distribution functions.

**Keywords :** Conditional empirical processes; VC-classes; Kernel-type estimators, density function; regression function; censored data.

## Résumé de la thèse

Dans cette thèse, nous nous intéressons à la convergences uniforme en terme de fenêtre de l'estimateur à noyau de la fonction de régression, en utilisant la théorie moderne des processus empiriques, sous des conditions très générales sur le noyau par rapport à la littérature existante. Nous obtenons des résulats de convergence uniforme en terme de fenêtre adaptative à la dimension intrinsèque pour une famille large d'estimateurs à noyau. La thèse est divisée en trois parties principales, que nous décrivons comme suit. La première partie est consacrée aux processus empiriques généraux indexés par des classes de fonctions. Les résultats sont obtenus pour des classes de fonctions uniformément bornées, ou non bornées avec des fonctions enveloppes satisfaisant certaines conditions de moments. La deuxième partie a pour objet les applications statistiques permettant d'illustrer l'utilité de la contribution principale de cette thèse. Les applications comprennent la consistance uniforme en terme de fenêtre des estimateurs de type noyau de la densité, la régression, la distribution conditionnelle, le mode multivarié, l'entropie de Shannon, les dérivés des fonctions de densité et de régression et les modèles additifs. La troisième partie est consacrée à la consistance uniforme en terme de fenêtre de l'estimateur non paramétriques du type "inverse probability of censoring weighted" (I.P.C.W.) de la fonction de régression dans le cadre de la censure à droite. Ces nouveaux résultats sont appliqués aux fonctions, de densité conditionnelle et de distribution conditionnelle, non paramétriques.

**Mots clés :** Processus empiriques conditionnels ; classes VC ; l'estimateur à noyau , fonction de densité ; fonction de régression ; données censurées.

# ملخص المذكرة

في هذه الأطروحة سنهتم بالتقارب المنتظم على عرض النطاق لمقدر دالة الانحدار بطريقة النواة، باستخدام النظرية الحديثة للعمليات التجريبية، تحت شروط جد عامة على النواة مقارنة بما كان مستخدما سابقا نحصل على نتائج حول التقارب المنتظم من حيث عرض النطاق المتكيفة مع البعد الجوهري لعائلة كبيرة من مقدرات النواة. تنقسم الأطروحة إلى ثلاثة أجزاء رئيسية نصفها على النحو التالي. الجزء الأول مخصصا للعمليات التجريبية العامة المفهرسة حسب فئات الدوال.يتم الحصول على النتائج لفئات الدوال المحدودة بانتظام، أو غير المحدودة ذات دوال المغلف والتي تلبي شروط لحظية معينة. يركز الجزء الثاني على التطبيقات الإحصائية لتوضيح فائدة المساهمة الرئيسية لهذه الأطروحة. تتضمن التطبيقات تناسقا منتظم من حيث عرض النطاق للمقدرات من نوع النواة لكل من دالة الكثافة، الانحدار، التوزيع الشرطي، الوضع متعدد المتغيرات ، إنتروبيا شانون ، مشتقات دوال الكثافة والانحدار والنماذج المضافة. الجزء الثالث مخصص لدراسة التناسق المنتظم على عرض النطاق للاحتمالية العكسية غير الوسيطية للمقدر المحجوب الموزون لدالة الانحدار تحت حجب عشوائي. يتم تطبيق هذه النتائج الجديدة على دوال الكثافة الشرطية والانحدار الشرطي غير الوسيطية.

**الكلمات المفتاحية** العمليات التجريبية الشرطية، الفئات من نوع $VC$ ، المقدرات من نوع النواة، دالة الكثافة، دالة الانحدار، المعطيات الخاضعة للحجب.