Asymptotic properties of a Nadaraya-Watson type estimator for regression functions of infinite order*

Seok Young Hong† Oliver Linton‡

University of Cambridge

21 October 2016

Abstract

We consider a class of nonparametric time series regression models in which the regressor takes values in a sequence space and the data are stationary and weakly dependent. Technical challenges that hampered theoretical advances in these models include the lack of associated Lebesgue density and difficulties with regard to the choice of dependence structure of the data generating process in the dynamic regression framework. We propose an infinite dimensional Nadaraya-Watson type estimator with a bandwidth sequence that shrinks the effects of long lags. We investigate its asymptotic properties in detail under both static and dynamic regressions contexts, aiming to answer the open questions left by Linton and Sancetta (2009). First we show pointwise consistency of the estimator under a set of mild regularity conditions. We establish a CLT for the estimator at a point under stronger conditions as well for a feasibly studentized version of the estimator, thereby allowing pointwise inference to be conducted. We establish the uniform consistency over a compact set of logarithmically increasing dimension. We specify the explicit rates of convergence in terms of the Lambert W function, and show that the optimal rate that balances the upper bound of squared bias and variance is of logarithmic order, the precise rate depending on the smoothness of the regression function and the dependence of the data.
Keywords: Functional Regression; Nadaraya-Watson estimator; Curse of infinite dimensionality; Near Epoch Dependence.

---

*Part of this paper was written while the first author visited Institut de Mathématiques, Toulouse in November 2013. We thank Philippe Vieu for his kind hospitality and many helpful advice since then throughout writing this paper. We also thank John Aston, Jeroen Dalderop, Paul Doukhan, Jiti Gao, Hayden Lee, Mikhail Lifshits, Alexei Onatski, and Richard Nickl for their helpful comments. Special thanks to Alexey Rudenko for providing an original Russian photocopy of Sytaya (1974), and to Hyungjin Lee for translating the paper. Financial support from the European Research Council (ERC-2008AdG-NAMSEF) is gratefully acknowledged.

†Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, Faculty of Mathematics, University of Cambridge, United Kingdom, email: syh30@cam.ac.uk

‡Faculty of Economics, University of Cambridge, United Kingdom, e-mail: obl20@cam.ac.uk

1 Introduction

Nonparametric modelling is a common method for analyzing time series; see for example Härdle (1990), Bosq (1996), or Fan and Yao (2003) for a comprehensive review. A major advantage of this approach is that the relationship between the explanatory variables under study, denoted by $X = (X_1, \ldots, X_d)^\intercal$, and the response, say Y, can be modelled without assuming any restrictive parametric or linear structures. One issue with allowing for this extended flexibility is known as the *curse of dimensionality*; Stone (1980, 1982) showed that given a fixed measure of smoothness β allowed on the regression function the best achievable convergence rate (in minimax sense) $n^{-\beta/(2\beta+d)}$ deteriorates dramatically as the dimension/order d increases.

In a time series context it is often reasonable and advantageous to model the dependence upon the infinite past. For example, the AR(d) and ARX(d) models with $d = \infty$ naturally extends those classical linear models, and enables the influence of *all past information* to be taken into account in the modelling procedure, thereby allowing for maximal flexibility with regard to the dynamic structure. It can also be very useful for several semiparametric applications, and for testing the martingale hypothesis or the efficient market hypothesis in economics, where the conditional mean given all past information $E(Y_t|\mathcal{F}_{t-1})$ is the object of main interest. It is desirable to be able to nest the linear AR(∞) models, and this is one of our aims. Not restricting the number of conditioning variables also has an advantage of avoiding the statistician's *a priori* choice of the order d based on some order determination principles whose validity is often subject to question in practical situations. For these reasons, we are motivated to study a class of nonparametric time series regression models of infinite order that covers both static and dynamic regressions cases, where the latter includes the autoregression framework as a special case.

A general class of nonlinear AR(∞) models has been studied by Doukhan and Wintenberger (2008) who showed the existence of a stationary solution, see also Wu (2011). Pagan and Ullah (1988) proposed studying the nonparametric regression case where $d \to \infty$ in the context of an econometric analysis of risk models. Linton and Sancetta (2009) tackled the estimation problem in the context of an autoregressive model and established uniform almost sure consistency with respect to stationary ergodic sample observations. There is a vast literature on functional data (typical examples include curves and images), which are infinite-dimensional in nature. Masry (2005) provided a rigorous treatment of nonparametric regression with dependent functional data in which X lies in a general semi-metric space, constructing the central limit theorem. Further, Mas (2012) derived the minimax rate of convergence for nonparametric estimation of the regression function on strictly independent and identically distributed covariates. Ferraty and Vieu (2006) detailed a number of extensions and overview of nonparametric approaches in functional statistics literature. Geenens (2011) gave an up-to-date accessible summary on nonparametric functional regression, and introduced the term *curse of infinite dimensionality*, which reflects evident difficulties in nonparametric estimation of infinite-dimensional objects due to extreme sparsity. We discuss in the next section the difference between the functional data framework and our discrete time framework.

Whereas nonparametric regression problems for vector regressors have been exhaustively studied in the literature, statistical theories for their infinite-dimensional extensions have not been fully established due to some technical challenges. An obvious difficulty stems from the fact that the usual notion of density $p(\cdot)$ does not exist; since there is no σ-finite Lebesgue measure in infinite-dimensional spaces, the Lebesgue density (with respect to the infinite product of probability measures) of the regressor cannot be defined via the Radon-Nikodym theorem. Consequently, standard asymptotic arguments for kernel estimators are no longer valid, for example, Bochner's lemma: under suitable regularity conditions, for $j = 1, 2$

$$\begin{aligned} \frac{1}{h^d} E\left[\mathcal{K}^j\left(\frac{x-X}{h}\right)\right] &= \int \mathcal{K}^j(u) p(x-uh)\, du \\ &\to p(x)\|\mathcal{K}\|_j^j \quad \text{as} \quad h \to 0 \end{aligned} \tag{1}$$

where $\mathcal{K}$ is a multivariate kernel function (see subsection 2.2 below). Hence, classical limiting theories in nonparametric literature cannot be readily extended.

In this paper, we consider an infinite dimensional analogue of the classical nonparametric regression approach. We propose a Nadaraya-Watson (i.e. local constant) type estimator and investigate its large sample properties. In particular, we show both pointwise and uniform consistency of the estimator and establish its asymptotic normality under both static and dynamic regression contexts with respect to α-mixing and near epoch dependent sample observations. Upon imposing some regularity conditions on the vector of bandwidth sequence, we derive the rate of convergence via specifying the small deviation probabilities, and confirm the existence of the curse of infinite dimensionality. Our pointwise rate is consistent with the pointwise rate established by Mas (2012) in the case of independent and identically distributed regressors.

For notations, we define $a_n \simeq b_n$ by $a_n = b_n + o(1)$, and $c_n \sim d_n$ by equivalence of order between the two sequences c_n and d_n. Also, $f \preceq g$ means there exists some constant $c > 0$ such that $\lim_{n\to\infty} f(n)/g(n) \leq c$. The term 'stationarity' is taken to mean strict stationarity. Throughout, C (or C', C'') refers to some generic constant that may take different values in different places unless defined specifically otherwise.

2 Some Preliminaries

Consider the following regression model:

$$Y = m(X) + \varepsilon \tag{2}$$

where the regressor $X = (X_1, X_2, \ldots)^\intercal$ is a random element taking values in some sequence space S, the response Y is a real-valued variable, and the stochastic error ε is such that $E(\varepsilon|X) = 0$ a.s. The objective is to estimate the Borel function $m(\cdot) = E(Y|X = \cdot)$ based on n random samples observed from a strictly stationary data generating process $\{(Y_t, X_t) \in \mathbb{R} \times S\}_{t\in\mathbb{Z}}$ having some weak dependence structure (see section 2.1 below).

This setting is related to the usual framework adopted for functional data which has been widely studied by statisticians, see Ramsey and Silverman (2002). Recently, successful attempts have been made to develop theories for nonparametric inference in the functional statistics literature; Ferraty and Romain (2010) gives a comprehensive review. A major issue in this field of research lies in extending the statistical theories applicable to $\mathbb{R}^d$ to function spaces. In this literature, attention is usually on smooth functions that are approximated and reconstructed from finely discretised grids on some compact interval. In contrast, the setup in our model (2) can be viewed as looking at a countable number of discrete observations. Such a difference is reflected by the fact that the observed data is taken to be a discrete process $X = (X_s)$ with unbounded $s \in \mathbb{Z}^+$ so that $S = \{f|f : \mathbb{N} \to \mathbb{R}\}$, rather than $X = (X(s))$ with $s \in [0,T]^k$ so that $S = \{f|f : [0,T]^k \subset \mathbb{R}^k \to \mathbb{R}\}$, e.g. curves if $k = 1$, images if $k \geq 2$. The discrete nature of our setting has several fundamental distinctive features that allow us to look further into many specific practical applications, a notable example of which is stationary autoregressive modelling.

An immediate consequence of our framework is that the tuning parameter can be imposed on each and every dimension, allowing one to control the marginal influence of the regressors. For instance when it is sensible to postulate that the influence of distant covariates is getting monotonically downweighted, one may set the marginal bandwidths to increase in lags so as to impose higher amount of smoothing. Depending on the nature of the regressor, S may be taken as the space of all infinite real sequences $\mathbb{R}^\infty := \prod_{j=1}^{\infty} \mathbb{R}_j$ formed by taking Cartesian products of the reals, or its various linear subspaces such as ℓ_∞, ℓ_p, c. We propose to take $S = \mathbb{R}^\infty$ so as to refrain from imposing any prior restrictions with regard to the choice of the regressor; for example, taking S to be the space of bounded sequences excludes the possibility of the regressors with infinite supports (e.g. Gaussian process).

2.1 Dependence structure and leading examples

A distinctive characteristic of time series data is temporal dependence between observations. As in the usual multivariate framework, we need some suitable assumptions on the dependence structure between the samples in order to derive asymptotic theories and to obtain convergence rates of the estimator. In nonparametric time series literature, Rosenblatt (1956)'s α-mixing has been the *de facto* standard choice due to being the weakest among the class of mixing-type asymptotic independence conditions. To name a few earlier works, pointwise and uniform consistency of the local constant estimator were shown by Roussas (1990) and Andrews (1995), respectively, and the asymptotic normality was established by Fan and Masry (1992). The α-mixing condition has also been widely used in the context of dependent functional observations, see for instance Ferraty et al. (2010), Masry (2005), and Delsol (2009).

Definition 1. *A stochastic process $\{Z_t\}_{t=1}^{\infty}$ defined on some probability space $(\Omega, \mathcal{F}, P)$ is called α-mixing (cf. 'jointly' α-mixing if Z_t is $\mathbb{R}^d$-valued, with $d \in (1,\infty]$ if*

$$\alpha(r) := \sup_{A \in \mathcal{F}_{-\infty}^{t}, B \in \mathcal{F}_{r+t}^{\infty}} |P(A \cap B) - P(A)P(B)|$$

is asymptotically zero as $r \to \infty$, where $\mathcal{F}_a^b$ is the σ-algebra generated by $\{Z_s; a \leq s \leq b\}$. In particular, we say the process is algebraically (resp. exponentially) α-mixing if there exists some $c, k > 0$ such that $\alpha(r) \leq cr^{-k}$ (resp. if there exists some $\gamma, \varsigma > 0$ such that $\alpha(r) \leq \exp(-\varsigma r^\gamma)$).

The popularity of the α-mixing condition (NB. the modifier α- will occasionally be omitted if no confusion is likely) in the literature stems from the fact that its associated probability theories have been extensively developed, making asymptotic derivations simpler, see e.g. Doukhan (1994) or Rio (2000) for a comprehensive survey. However, several drawbacks have been pointed out in the literature. First, it is a rather strong technical condition that is hard to verify in practice. Moreover, it is well-known that even some basic processes are not mixing. e.g. AR(1) with Bernoulli innovations, Andrews (1984).

The primary limitation of mixing conditions that we should be aware of in our paper arises from the choice of the framework upon which our theories are developed. Recall that given the sample observations $\{Y_t, X_t\}_{t=1}^n$ the object of estimation is the conditional mean $E(Y_t|\mathcal{F})$, cf. (2), where the information set $\mathcal{F}$ is determined by the nature of the conditioning variables. There are two leading cases: the first case is the static regression where the information set is taken to mean $\sigma(X_{jt}; j = 1, 2, \ldots)$, the σ-algebra generated by the exogenous marginal regressors. The second case is the autoregression, where $X_{tj} = Y_{t-j}$ for all j, in which case $\mathcal{F} = \mathcal{F}_{t-1}$ represents $\sigma(Y_s; s \leq t-1)$, the σ-algebra generated by the sequence of the lags of the response $(Y_s)_{s \leq t-1}$. In fact, as for the latter framework we may consider a more general setup, i.e. a dynamic regression, where the information set is taken to be $\mathcal{F} = \sigma(X_{js}, Y_s;\ s \leq t-1)$ for some j. Details are formally given in Assumptions A below.

In the static regression case the usual joint α-mixing condition can be assumed on the sample data $\{Y_t, X_t\}$ as is usually done in the multivariate framework; since marginal regressors are observed at the same time t: $X_t = (X_{1t}, X_{2t}, \ldots)^\intercal$, assuming joint dependence does not require additional adjustments. Indeed, joint mixing implies both marginal component processes and any measurable function thereof are mixing[1]. In this paper, we do not necessarily require independence between component processes $\{X_{jt}\}$, $j = 1, 2, \ldots$; later we will show what happens to the asymptotic properties of the estimator when certain dependence is allowed (see Assumption C2).

Moving on to the dynamic regression setting, since the regressors are taken to be the lags of the response and/or a covariate, measurable functions of X_t depend on infinite time-lags and hence are *not* mixing in general[2]. Therefore an alternative set of dependence conditions is necessary to establish asymptotic theories for the second framework. We shall adopt the notion of near epoch dependence due to Ibragimov (1962) as for the dynamic regression setting and deal with two leading cases separately.

Definition 2. *A stochastic process $\{Z_t\}_{t=1}^\infty$ defined on some probability space $(\Omega, \mathcal{F}, P)$ is called near-epoch dependent or stable in L_2 with respect to a strictly stationary α-mixing process $\{\eta_t\}$ if the stability coefficients $v_2(r) := E|Z_t - Z_{t,(r)}|^2$ is asymptotically zero as $r \to \infty$, where $Z_{t,(r)} = \Psi_r(\eta_t, \ldots, \eta_{t-r+1})$* for some Borel function $\Psi_r : \mathbb{R}^r \to \mathbb{R}$.

A process that is *near epoch* dependent on a mixing sequence is influenced primarily by the "recent past" of the sequence and hence asymptotically resembles its dependence structure; see e.g. Billingsley (1968), Davidson (1994), or Lu (2001) for details. Following the usual convention, e.g. Bierens (1983), we shall take $\Psi_r(\eta_t, \ldots, \eta_{t-r+1}) \equiv E(Z_t|\eta_t, \ldots, \eta_{t-r+1})$. In section 3 it will be shown that under suitable conditions similar

[1] The converse is not necessarily true unless the marginal processes are independent to each other, see Bradley (2005, Section 5).

[2] Except for some very special cases; Davidson (1994, Theorem 14.9) gives a set of technical conditions under which a process with infinite (linear) temporal dependence is α-mixing.

asymptotic theories can be derived for both static and dynamic regression frameworks.

2.2 Local Weighting

In order to extend standard multivariate nonparametric theories to infinite dimension and to conduct nonparametric inference on the object of interest $m(\cdot)$, we first fix the notion of (i) local weighting and (ii) the measure of closeness between the objects. In the finite d-dimensional case, these are respectively done by the (multivariate) kernels $\mathcal{K} : (\mathbb{R}^d, \|.\|_E) \to [0, \infty)$ which control the way the weights are given based on the distance where $\|.\|_E$ is the Euclidean norm, and the bandwidth h that regulates the prescribed distance within which smoothing is done. The function $\mathcal{K}$ is called the product kernel or the spherical kernel depending on the way it is constructed from the univariate kernel $K(\cdot)$, i.e. $\mathcal{K}^P(u) = \prod_j^d K_j(u_j)$ or $\mathcal{K}^S(u) = K(\|u\|_E)$.

Motivated by the approach in functional statistics literature, we extend the latter scheme to construct the weighting function so as to account for the infinite-dimensional nature of our framework. For an element u of a normed sequence space, let

$$\mathcal{K}(u) := K(\|u\|), \tag{3}$$

where the univariate kernel K is a density function with non-negative support, for instance $K(\cdot) = \sqrt{2/\pi} \exp(- \cdot^2 /2)$. We now group the kernel functions into three subcategories depending on the way how they are generated. The first two, referred to as Type-I and Type-II kernels in Ferraty and Vieu (2006) generalize the usual 'window' kernels and monotonically decreasing kernels in finite dimension, respectively. Both types of kernels are continuous on compact support $[0, \lambda]$.

Definition 3. *A function $K : [0, \infty) \to [0, \infty)$ is called a kernel of type$-I$ if it integrates to 1, and if there exist real constants C_1, C_2 (with $0 < C_1 < C_2$) for which*

$$C_1 1_{[0,\lambda]}(u) \leq K(u) \leq C_2 1_{[0,\lambda]}(u), \tag{4}$$

where λ is some fixed positive real number. Also, a function $K : [0, \infty) \to [0, \infty)$ is called a kernel of type$-II$ if it satisfies (4) with $C_1 \equiv 0$, and is continuous on $[0, \lambda]$ and differentiable on $(0, \lambda)$ with the derivative K' that satisfies

$$C_3 \leq K'(u) \leq C_4$$

for some real constants C_3, C_4 such that $-\infty < C_3 < C_4 < 0$.

The definition above suggests that the uniform kernel on $[0, \lambda]$ is a type-I kernel, and the Epanechnikov, Biweight and Bartlett kernels belong to the class of Type-II kernels. Some of those with semi-infinite support, for example (one-sided) Gaussian, are covered by the last group, which we will call the Type-III kernels.

Definition 4. *A function $K : [0, \infty) \to [0, \infty)$ is a kernel of type$-III$ if it integrates to 1, and if it is of exponential type; that is, $K(r) \propto \exp(Cr^{\beta})$ for some β and C.*

2.3 Small deviation

The *small ball (or small deviation) probability* plays a crucial role in establishing asymptotic theories of this paper. Let S^* be a sequence space equipped with some norm $\|.\|$; then the small ball probability of an S^*-valued random element Z is a function defined as

$$\varphi_z(h) := P\left(\|z - Z\| \leq h\right). \tag{5}$$

We shall call the probability *centered* if $z = 0$ (in which case we write $\varphi(h)$), and *shifted* (with respect to some fixed point $z \in S^*$) if otherwise. The relation between the two quantities cannot be explicitly specified in general, and will be given in terms of the Radon-Nikodym derivative (See Assumption D1 below).

The name *small ball* stems from the fact that we are interested in its asymptotic behaviour as h, the bandwidth sequence in the context of nonparametric estimation, tends to zero. The function can be thought of as a measure for how much the observations are densely *packed* or *concentrated* around the fixed point z with respect to the associated norm and the reference distance h. From the definition it is straightforward to see that $\varphi_z(h) \to 0$ as $h \to 0$, and that $n\varphi_z(h)$ is an approximate count of the number of observations whose influence is taken into account in the smoothing procedure. When Z is a d-dimensional continuous random vector with density $p(\cdot) > 0$, it can be readily shown that the shifted small ball (w.r.t. the usual Euclidean norm) is given by

$$\varphi_z(h) = V_d h^d p(z) = O(h^d), \tag{6}$$

where $V_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of the d-dimensional unit sphere.

However suppose now Z takes values in an infinite-dimensional normed space; then it becomes difficult to specify the exact form of the small ball probability, and its behaviour varies depending heavily on the nature of the associated space and its topological structure. Due to non-equivalence of norms, it is intuitively clear that the "speed" at which $\varphi_z(h)$ converges to zero is affected by the choice of the norm $\|.\|$. Nonetheless, a rapid decay is expected in general irrespective of the choice of the norm due to extreme sparsity of data in infinite-dimensional spaces.

One possible example of S^* is $(\ell_r, \|.\|_r)$, the space of r-th power summable sequence equipped with the ℓ_r-norm; the centred small ball behaviour of sums of weighted i.i.d. random variables is widely studied in the literature, see for example Borovkov and Ruzankin (2008) and references therein. In this paper, we will focus our main attention on the case of $r = 2$ (and take $\|.\|$ to mean $\|.\|_2$ unless specified otherwise). Nevertheless, the results derived in this paper can be extended to the case of $r > 2$ as long as the regularity conditions are adjusted appropriately.

Writing the expected value of the kernel in terms of the small ball probability

$$E\mathcal{K}\left(\frac{z-Z}{h}\right) = EK\left(\frac{\|z-Z\|}{h}\right) = \int K(u)\ dP_{\|z-Z\|/h}(u) = \int K(u)\ d\varphi_z(uh), \tag{7}$$

we are able to bypass the difficulties mentioned in the introduction, and to establish the convergence of the integrals without explicitly requiring the existence of the density.

LEMMA 1. Ferraty and Vieu (2006, Lemma 4.3 & 4.4). *Suppose $\|.\|$ is some semi-norm defined on a functional space. If K is type-I, then it satisfies*

$$C_1^j \leq \frac{1}{\varphi_z(h\lambda)} \int_0^\lambda K^j(v)\ d\varphi_z(vh) \leq C_2^j, \quad j = 1, 2 \tag{8}$$

where $C_1, C_2 > 0$ are as defined in Definition 3. When the kernel K is type-II, if

$$\exists\ \varepsilon_0 > 0, C_5 > 0 \text{ s.t. } \forall \varepsilon < \varepsilon_0, \quad \int_0^{\varepsilon} \varphi_x(u)du > C_5 \varepsilon \varphi_x(\varepsilon) \tag{9}$$

then we have

$$C_6^j \le \frac{1}{\varphi_z(h\lambda)} \int_0^{\lambda} K^j(v)\ d\varphi_z(vh) \le C_7^j, \quad j = 1, 2 \tag{10}$$

where the constants $C_6 = -C_5 C_4$ and $C_7 = \sup_{s \in [0,\lambda]} K(s)$ are strictly positive.

Under the regularity conditions of Lemma 1, (8) and (10) hold for every $h > 0$, so it follows that for any kernels of type-I and II:

COROLLARY 1. *If the kernel K is either type-I or type-II, then for $j = 1, 2$ we have*

$$\frac{1}{\varphi_z(h\lambda)} E\left[\mathcal{K}^j\left(\frac{z - Z}{h}\right)\right] \longrightarrow \xi_j \quad \text{as} \quad h \to 0^+, \tag{11}$$

where ξ_1 and ξ_2 are some strictly positive real constants.

This result can be seen as an infinite-dimensional analogue of Bochner's lemma (1): i.e. for $Z \in \mathbb{R}^d$, $h^{-d} E\mathcal{K}((z - Z)/h) \to p(z) > 0$, and are fundamental in constructing asymptotic theories of kernel estimators. It is obvious that ξ_j is bounded below and above by C_1^j and C_2^j, respectively (or C_6^j and C_7^j depending on the choice of the kernel). With specific choices of kernels and regressors we may be able to specify the exact values of the constants in some certain cases. For example, it is straightforward to see that $\xi_1 = 1/\lambda$ and $\xi_2 = 1/\lambda^2$ when K is a uniform kernel.

REMARKS. (i) The latter half of Lemma 1 reveals the importance of condition (9) in constructing the asymptotics when the kernel is of type-II. Whereas the condition is widely assumed in functional statistics literature for that reason, Azais and Fort (2013) proved that it necessarily restricts the variable Z to be of finite dimension. In other words, whenever (9) is valid, the topology that governs the concentration properties of Z accounts effectively only for finite dimension. An example (cf. Section 13.3.3 of Ferraty and Vieu (2006)) includes the case where Z is associated with the semi-norm $\|y\| := (y_1, \ldots, y_p, 0, 0, \ldots)$ for some positive integer $p < \infty$, where $y \in \mathbb{R}^\infty$. Since this severely restricts the applicability of our paper, we shall not consider the case of Type-II kernels. (ii) A natural question one may then ask is whether (11) would hold for kernels with semi-infinite support such as the Type-III kernels. In finite $\mathbb{R}^d$-framework, it is well known that a set of assumptions including $\|u\|^d K(u) \to 0$ as $u \to \infty$ is sufficient for showing (1), see for instance Parzen (1962, Theorem 1A) and Pagan and Ullah (1999, Lemma 1). However, in the infinite-dimensional setting the answer is negative in most usual cases where the kernel is of exponential type (e.g. Gaussian kernel). Whereas the lower bound of the limit can be easily constructed via Chebyshev's inequality: with reference to Definition 4, writing $V = \|z - Z\|^\beta$ and $\delta = h^\beta$ we have

$$(0 <) \exp(-c_\delta \delta) \le [P(V \le \delta)]^{-1} E \exp(-c_\delta V), \tag{12}$$

the upper bound may not exist, and the rate at which the small ball probability decays to zero may dominate the speed at which the integral (7) converges to zero. This claim cannot be formally verified for all general cases because (as aforementioned) there is no unified result for the asymptotic behaviour of small deviation available. Nonetheless, the idea can be sketched in the common case where the asymptotics of the distribution function (i.e. small deviation) is of exponential order: $P(V \leq \delta) \sim \exp(-C\delta^{-\theta})$ as $\delta \to 0$ for some constants C and $\theta > 0$; by de Bruijn's exponential Tauberian theorem (see Bingham et al. (1987), Li (2012)) the necessary and sufficient condition for such a case is the following limiting behaviour of the Laplace transform near infinity:

$$E[\exp(-c_\delta V)] \sim \exp\Big(-C' \cdot c_\delta^{\theta/(1+\theta)}\Big) \quad \text{as} \quad c_\delta \to \infty$$

for some constant $C' > 0$. With $V = \|z - Z\|^2$, $\delta = h^2$, $c_\delta = 2^{-1}h^{-2}$ (which corresponds to the case of the Gaussian kernel) the difference in the order of convergence suggests that the RHS of (12) is unbounded, and that the limit (11) diverges. Due to this reason, we shall confine our attention to compactly supported kernels in this paper.

2.4 The Estimator

Consider the regression problem with $\mathbb{R}^\infty$-dimensional regressor $X = (X_1, X_2, \ldots)^\intercal$, where the regression operator is nonparametrically estimated with an estimator associated with a bandwidth matrix $H := diag(\underline{h}) = diag(h_1, h_2, \ldots)$. No further condition is explicitly assumed in the first instance except that we suppose a norm $\|.\|$ can be admitted to the *weighted regressor* by appropriately choosing the bandwidth sequence.

As an expositional example, let us assume $h_j = \phi_j h$ where $\{\phi_j^{-1}\}_j$ is some square-summable sequence (e.g. $\phi_j = j^p$ for some $p > 1/2$). Then by Kolmogorov's three-series theorem, the sequence of weighted regressor $\phi_j^{-1} X_j$ is square summable, w.p.1., provided that marginal regressors X_j' are independent with finite variance and satisfy

$$\sum_{j=0}^{\infty} E \min\left\{1, \phi_j^{-2} X_j^2\right\} < \infty, \tag{13}$$

so that $(\phi_1^{-1}X_1, \phi_2^{-1}X_2, \ldots)^\intercal =: Z$ is $(\ell_2, \|.\|_2)$-valued. In terms of the autoregressive framework the sequence ϕ_j can be interpreted as non-decreasing weights that represent the "relative influence" of the marginal regressors which diminishes as lags get apart.

Therefore, from now on we assume that Z is ℓ_2-valued and normed with $\|.\| = \|.\|_2$. Also, (with an abuse of notation) the usual definition of shifted small deviation is extended to account for the generalized support $[0, \lambda]$ and bandwidth vector $\underline{h} = (h_1, h_2, \ldots)^\intercal$:

$$\varphi_x(\underline{h}\lambda) := P\big(\|H^{-1}(x - X_t)\| \leq \lambda\big). \tag{14}$$

Equivalently, $\varphi_x(\underline{h}\lambda) = P(X_t \in \mathcal{E}(x, \underline{h}\lambda))$, where $\mathcal{E}$ is the infinite-dimensional hyper-ellipsoid centred at $x \in \mathbb{R}^\infty$, and λ is the constant defined in section 2.2. Clearly, $\varphi_x(\underline{h}\lambda) = \varphi_z(h\lambda)$, where $z := (\phi_1^{-1}x_1, \phi_2^{-1}x_2, \ldots)^\intercal$.

For later reference, we also define the joint small ball probability of the regressor vectors observed at different times t and s as the joint distribution

$$\psi_x(\underline{h}\lambda; t, s) := P\big((X_t, X_s) \in \mathcal{E}(x, \lambda\underline{h}) \times \mathcal{E}(x, \lambda\underline{h})\big). \tag{15}$$

We are finally in a position to introduce our estimator. We propose to estimate $m(x) = E(Y|X = x)$, $x \in \mathbb{R}^\infty$ with the following local constant type estimator:

$$\widehat{m}(x) := \frac{\sum_{t=1}^{n} \mathcal{K}\Big(H^{-1}(x - X_t)\Big) Y_t}{\sum_{t=1}^{n} \mathcal{K}\Big(H^{-1}(x - X_t)\Big)} \equiv \frac{\sum_{t=1}^{n} K\Big(\|H^{-1}(x - X_t)\|\Big) Y_t}{\sum_{t=1}^{n} K\Big(\|H^{-1}(x - X_t)\|\Big)}, \tag{16}$$

defined with respect to n-sample time series $\{Y_t, X_t\}_{t=1}^n$. Whereas in the 'cross-sectional case' we may observe an infinity of regressors, in the autoregression case we essentially observe only $\{Y_1, Y_2, \dots, Y_n\}$ rather than the full infinity, see Assumptions A below. Hence for practical applications, one may employ a truncation argument on the regressor (as will be done in section 3.4 - albeit with a different purpose) and let the effective dimension τ of the regressor X_t to increase in n.

The estimator can be viewed as an infinite-dimensional generalization of the standard multivariate local linear estimator, and is a special case of the one in Ferraty and Vieu (2002), Masry (2005) and references therein for functional data. In the following section we will examine some asymptotic properties of the estimator.

3 Asymptotic Properties

In this section we introduce the main results of our paper, deriving some large sample asymptotics of the proposed estimator (16). Consistency is shown in both pointwise and uniform sense and the asymptotic normality is established. Convergence rates are specified under sets of regularity conditions. All proofs are detailed in the appendix.

Following the discussions in section 2.1, we shall consider two different cases: (1) static regression and (2) dynamic regression. Below we specify two sets of dependence conditions, either of which will be assumed on the data generating process of the sample observations. Assumption A1 corresponds to the static regression case where we have exogenous regressors that are jointly observed in time in a weakly dependent manner. No restriction is needed as regards the dependence structure between the marginal regressors, although certain additional conditions can be potentially imposed at the later stage (see Assumptions E below). The second option A2 concerns the dynamic regressions framework; the notion of near epoch dependence is adopted to describe the dependence structure of the processes defined as functions of the response variables. The assumptions below suggest that there is a trade-off between the degree of mixing and the possible order of moments $2 + \delta$ allowed on the response variable.

Assumptions A

A1. *The marginal regressors $X_{1t}, X_{2t}, X_{3t}, \dots$ are exogenous variables, and the sample data $\{Y_t, X_t\}_{t=1}^n = \{Y_t, (X_{1t}, X_{2t}, \dots)\}_{t=1}^n$ is stationary and jointly arithmetically α-mixing with rate $k > (2\delta+4)/\delta$, where δ is as defined in Assumption B4 below.*

A2. *Each regressor is either a lag of the response variable Y_t or of a covariate V_t, i.e. $X_{jt} = Y_{t-j}$ or $X_{jt} = V_{t-j}$, $j \in \mathbb{N}$, and $\{Y_t, V_t\}_{t=1}^n$ is stationary and arithmetically α-mixing with rate $k > (2\delta + 4)/\delta$. Also, the process $K_t := K(\|H^{-1}(x - X_t)\|)$*

is near epoch dependent on (Y_t, V_t)*, and there exists some* $r = r_n \to \infty$ *such that the rate of stability for* K_t *denoted* $v_2(r_n) = v_2(r)$ *satisfies*

$$v_2(r)^{1/2}[\varphi_x(\underline{h}\lambda)]^{-(2\delta+3)/(2\delta+2)} n^{1/(2(\delta+1))} \to 0 \qquad as \;\; n \to \infty. \tag{17}$$

Remark. Our model under Assumption A2 can be viewed as a generalization of the NAARX model in Chen and Tsay (1993). The framework nests both the fully autoregressive framework in which $X_{jt} = Y_{t-j}$, $\forall j$ and the case where the regressor vector consists only of the lags of a covariate V_t.

3.1 Pointwise consistency

Pointwise consistency of the local constant estimator was first studied by Watson (1964) and Nadaraya (1964) with respect to i.i.d sample for the case of univariate regressor, i.e. $d = 1$. Their result was extended to the multivariate case (finite dimension) by Greblicki and Krzyzak (1980) and Devroye (1981). Robinson (1983) and Bierens (1983) were amongst the earliest papers that worked on consistency of the estimator with respect to dependent observations (both static regression and autoregression were allowed in their frameworks), followed by Roussas (1989), Fan (1990), and Phillips and Park (1998) to name a few out of numerous papers. The case of functional regressor was first studied by Ferraty and Vieu (2002), and references thereafter under various sets of regularity conditions.

In this section we show how these results can be extended specifically to our framework, proving pointwise weak consistency of the estimator (16) w.r.t. dependent sample satisfying either A1 or A2. A set of assumptions required for the theory is now introduced, and some introductory arguments are briefly sketched.

Assumptions B

- B1. *The regression operator* $m : \mathbb{R}^\infty \to \mathbb{R}$ *is continuous in some neighbourhood of* x
- B2. *The marginal bandwidths satisfy* $h_j = h_{j,n} \to 0$ *as* $n \to \infty$ *for all* $j = 1, 2, \ldots,$ *where* $diag(h_1, h_2, \ldots) = diag(\underline{h}) = H$ *is the bandwidth matrix, and the small ball probability obeys* $n\varphi_x(\underline{h}\lambda) \to \infty$ *for every point* $x \in \mathbb{R}^\infty$*, where* $\varphi_x(\underline{h}\lambda) := P(\|H^{-1}(x - X)\| \leq \lambda) \to 0$ *as* $n \to \infty$.
- B3. *The kernel* K *is type*$-I$
- B4. *The response* Y_t *satisfies* $E\big(|Y_t|^{2+\delta}\big) \leq C < \infty$ for some $C, \delta > 0$.
- B5. *The joint small ball probability (15) satisfies* $\psi_x(\underline{h}\lambda; i, j) \leq C\varphi_x(\lambda\underline{h})^2$, $\forall i \neq j$.
- B6. *The conditional expectation* $E\big(|Y_tY_s||(X_t, X_s)\big) \leq C < \infty$ *for all* t, s.

Remark. The continuity assumption B1 is necessary for asymptotic unbiasedness of the estimator. It will be shown that the estimator is unbiased at every point of continuity, and that the rate of convergence for the bias term can be specified upon imposing further smoothness condition on the regression operator, see later. Assumption B2 can be thought of as an extension of the usual bandwidth conditions that are assumed in finite-dimensional nonparametric literature, cf. (6). As discussed before, $n\varphi_x(\underline{h}\lambda)$ can be understood as an approximate number of observations that are "close enough" to x. Therefore, it is sensible to postulate that $n\varphi_x(\underline{h}\lambda) \to \infty$ as $n \to \infty$, meaning that the point x is visited many times by the sample of data as the size of sample grows to infinity. This is in line with the usual assumption $nh^d \to \infty$ when $X \in \mathbb{R}^d$, in which case the small ball probability is given by $\varphi_x(h) \propto h^d p_X(x)$ as noted in (6). Conditions B5 and B6 are imposed to control the asymptotics of the covariance terms. The validity of the former can be easily seen in the $\mathbb{R}^d$ frameworks; for relevant discussions, see Ferraty and Vieu (2006, Remark 11.2).

To sketch the idea, we write $K_t := K(\|H^{-1}(x - X_t)\|)$ for the sake of simplicity of presentation (note its dependence upon X_t), and express the estimator (16) as

$$\widehat{m}(x) := \frac{\sum_{t=1}^n K\big(\|H^{-1}(x - X_t)\|\big)Y_t}{\sum_{t=1}^n K\big(\|H^{-1}(x - X_t)\|\big)} = \frac{\frac{1}{n}\sum_{t=1}^n \frac{K_t}{EK_1}Y_t}{\frac{1}{n}\sum_{i=1}^n \frac{K_t}{EK_1}} = \frac{\widehat{m}_2(x)}{\widehat{m}_1(x)}. \tag{18}$$

We then employ the following decomposition:

$$\begin{aligned}\widehat{m}(x) - m(x) &= \frac{\widehat{m}_2(x)}{\widehat{m}_1(x)} - m(x) = \frac{\widehat{m}_2(x) - m(x)\widehat{m}_1(x)}{\widehat{m}_1(x)} \\ &= \frac{E\widehat{m}_2(x) - m(x)E\widehat{m}_1(x)}{\widehat{m}_1(x)} + \frac{[\widehat{m}_2(x) - E\widehat{m}_2(x)] - m(x)[\widehat{m}_1(x) - E\widehat{m}_1(x)]}{\widehat{m}_1(x)},\end{aligned} \tag{19}$$

where clearly $E\widehat{m}_1(x) = 1$. Below we show consistency by proving that the 'bias part' $E\widehat{m}_2(x) - m(x)$ and the 'variance part' $[\widehat{m}_2(x) - E\widehat{m}_2(x)] - m(x)[\widehat{m}_1(x) - 1]$ are negligible in large sample. As for the latter term, it suffices to show mean squared convergence of $\widehat{m}_2(x) - E\widehat{m}_2(x)$ to zero because $\widehat{m}_1(x) \to^P 1$ then readily follows.

THEOREM 1. *Suppose Assumptions B1-B5 hold. Then the estimator (16) with respect to the sample observations $\{Y_t, X_t^\intercal\}_{t=1}^n$ satisfying either A1 or A2 is weakly consistent for the regression operator $m(x)$. That is, as $n \to \infty$*

$$\widehat{m}(x) \xrightarrow{P} m(x). \tag{20}$$

REMARK. In the regression case (i.e. under Assumption A1), it is straightforward to see that both (20) and the uniform consistency result (to be constructed later in Theorem 4) can be strengthened to almost sure sense. We conjecture that the same extension can be made in the dynamic regressions framework described by Assumption A2[3]. However, we will not proceed to this direction in this paper and leave it for future studies. Note also that the result (20) trivially holds when the estimation is made with respect to i.i.d. data $\{(Y_t, X_t^\intercal);\ t \in \mathbb{Z}\}$. The arguments simply becomes less involved as the covariance term does not need to be considered anymore.

In the following section, it will be shown that the convergence rates can be specified and the asymptotic normality can be established by imposing a set of additional regularity conditions.

[3]In fact, it can be shown that it is indeed the case if Assumption B3 is replaced by $|Y_t| \leq C$.

3.2 Asymptotic Normality

Earlier studies on the limiting distribution of the standard Nadaraya-Watson estimator can be traced back to Schuster (1972) and Bierens (1987), where the case of univariate and multivariate regressors was considered, respectively. The case of dependent samples was studied in Robinson (1983) and Bierens (1983), Masry and Fan (1997) and by numerous others under various model setup and different regularity conditions.

General distributional theories for Nadaraya-Watson type estimators in a semi-metric space was established by Masry (2005, Theorem 4) and Delsol (2009). Our results are different from those in the existing literature in two perspectives. First, the difference of our framework from the functional literature discussed in the beginning of Section 2 allows us to develop asymptotic theories for our specific setting. Second, whereas the final results of many existing papers were given in terms of abstract functions, our results are presented with the explicit rate of convergence, thereby allowing potentially feasible applications.

The primary objective of this section is to outline this procedure in detail, and to introduce the main theories and some interesting consequences thereof. Both cases of the independent marginals (in other words, when the marginal regressors X_j are independent and identically distributed) and also a dependent framework are allowed. Specifically, we introduce how independence restriction can possibly be moderated to allow for some mild dependence structure. In particular, the second condition in Assumption C below specifies the extent to which certain cross-sectional dependence can be allowed on the marginal regressors in our theory while still allowing for specification of the exact form of the convergence rate of the estimator.

Assumptions C. *For every fixed t, the real-valued stochastic process formed by the marginal regressors $\{X_{jt}\}_{j=1}^{\infty}$ is either:*

C1. *independent and identically distributed over j with $EX_{jt}^4 \leq C < \infty$ $\forall j$*, or

C2. *stationary (over j), and admits a moving average representation:*

$$X_{jt} = \sum_{u=-\infty}^{\infty} a_u \epsilon_{j-u,t}, \tag{21}$$

where a_u is a square summable sequence, and $\{\epsilon_{jt}\}_j$ is an independent and identically distributed standard Gaussian sequence.

Remark. In either case marginal regressors are required to be identically distributed over j; an additional distributional assumption will be imposed in D2 below. Nonetheless, the possible degree of dependence allowed in C2 is very mild and general, since an equivalent condition of having the representation is simply the existence of the spectral density. Note that (21) includes the causal (one-sided) MA representation as a special case. If a stationary stochastic process $\{X_{jt}\}_j$ is α-mixing (over j), then it always has such a representation (i.e. $a_u = 0,\ \forall u < 0$) provided it is Gaussian. This is because any α-mixing process is regular[4] by definition, so is linearly regular

[4]In the sense of Ibragimov and Linnik (1971) and Davidson (1994, Part III)

when it is Gaussian, and hence (with stationarity) admits the Wold decomposition with independent Gaussian innovations by Corollary 17.3.1 of Ibragimov and Linnik (1971).

Note that each C1 and C2 is consistent with the case allowed in Assumption A1 and A2, respectively (because in the latter case the process $\{X_{jt}\}_j$ consists of temporal lags of the response variable and/or a covariate which form a mixing process by Assumption A2), although the dependence structure specified in C2 can be allowed also for the static case (i.e. A1). This suggests that there is absolutely no need to assume independence between marginal regressors in our model (2) under Gaussianity, and hence a wide flexibility is allowed in terms of the model setup. In particular, the convergence rates of our estimator will be shown to be invariant (upto some constant factor) to the choice between C1 and C2. Lastly, the requirement of finite fourth moment is imposed to ensure that the squared marginal regressors have finite second moments due to the reasons to be clarified below; obviously, when a lag of the response is included in the dynamic regression framework (Assumption A2), this forces $\delta \geq 2$ in Assumption B4.

We now introduce some main assumptions needed for distributional theories.

3.2.1 The 'bias component'

The first part concerns with the asymptotic 'bias', where Assumptions A is strengthened by imposing additional smoothness conditions and suitable bandwidth adjustments. They belong to a set of sufficient conditions under which the exact upper bound of the asymptotic bias can be specified. Note that alternatively, Fréchet differentiability type condition can be imposed, as was done in Mas (2012).

FURTHER ASSUMPTIONS B

B7. *In addition to Assumption B2, the marginal bandwidths satisfy* $h_j = \phi_j \cdot h$ *for some positive real number* ϕ_j *where* $h = h_n \to 0$ *as* $n \to \infty$.

B8. *The regression operator* $m : \mathbb{R}^\infty \to \mathbb{R}$ *satisfies*

$$\left|m(x) - m(x')\right| \leq \sum_{j=1}^{\infty} c_j \left|x_j - x'_j\right|^\beta \tag{22}$$

for every x' *and* $x = (x_1, x_2, \dots)^\intercal \in \mathbb{R}^\infty$, *and some constant* $\beta \in (0, 1]$, *where* c_j *is some sequence of real constants for which* $\sum_{j=1}^{\infty} c_j \leq 1$ *and* $\sum_{j=1}^{\infty} c_j \phi_j^\beta < \infty$.

REMARK. These additional assumptions help us specify and regulate the bias component. Assumption B7 extends the previous bandwidth condition B2. Obviously, it is consistent with what was previously assumed in B2 since $h \to 0$ implies the coordinate-wise convergence of each marginal bandwidths. With this condition one is able to write the asymptotic bias expression and the order of the bias-variance balancing bandwidth in terms of the common factor h. It is possible to dispense with this condition at the cost of imposing minor modifications in B8; the asymptotic bias will then be written in terms of the infinite sum of a weighted marginal bandwidth h_j, whose convergence needs to be ensured. For the sake of understanding the asymptotic behaviour of the variance component, a further increment condition will be imposed on the sequence of marginal coefficients ϕ_j in Assumption B later. We remark that at this point such an assumption is not necessary as the variance term is not concerned.

Assumption B8 replaces and strengthens Assumption B1, and can be thought of as a variant of Hölder-type continuity; the case of $c_j = 2^{-j}$ and $\beta = 1$ is implied by the Lipschitz condition. Another example of c_j includes $\exp(-j)$. Indeed, under B8 the regression operator becomes a contraction mapping, and the contribution from each marginal dimension decreases in lag or index. This ensures summability of the bias of the estimator and allows one to specify its order of convergence rate, cf. (27) below.

In the context of autoregression where $X_j \equiv Y_{t-j}$ for all j, the model is given by

$$Y_t = m(Y_{t-1}, Y_{t-2}, \ldots) + \varepsilon_t \tag{23}$$

and whether the stationary solution $\{Y_t\}$ indeed exists is an important question, since (23) essentially gives an infinite number of recurrence relations whose solution may not be always well-defined. In the study of a class of general nonlinear AR(d) models, Duflo (1997) and Götze and Hipp (1994) assumed what is called the Lipschitz mixing condition (or the strong contraction condition), which is essentially (22) replaced by finite d-sum on the right hand side. In our context, Assumption B8 plays a similar role; Doukhan and Wintenberger (2008) showed that (22) with $\sum_{j=1}^{\infty} c_j < 1$, is sufficient for the existence of a solution: for some measurable f,

$$Y_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots), \tag{24}$$

where ε_t is an i.i.d. sequence. Wu (2011) arrived at the same conclusion under the assumption of $\sum_{j=1}^{\infty} c_j = 1$; the specific restrictions on c_j are chosen to reflect their findings, despite the fact that we are not restricting the error process $\{\varepsilon_t\}$ to be an independent sequence in our model setup.

Before we proceed, we remark that from now on the rate condition stipulated in (17) is slightly strengthened as follows (and Assumption A2 is modified accordingly):

$$A2' : \; v_2(r)^{1/2}[\varphi_x(\underline{h}\lambda)]^{-1} n^{1/2} \to 0 \qquad as \;\; n \to \infty. \tag{25}$$

3.2.2 The 'variance component'

We now move on to the second chunk of assumptions that concerns with the 'variance part'. As before, vectors Z and z are taken to mean $(\phi_1^{-1}X_1, \phi_2^{-1}X_2, \ldots)^\intercal$ and $(\phi_1^{-1}x_1, \phi_2^{-1}x_2, \ldots)^\intercal$, respectively, where the vector $x = (x_1, x_2, \ldots)^\intercal$ is the point at which estimation is made, and ϕ_j's are the coefficients in Assumption B7.

Assumptions D

D1. *The induced probability measure P_{z-Z} is dominated by the measure P_Z, and its Radon-Nikodym density $dP_{z-Z}/dP_Z =: p^*$ is continuous and is bounded away from zero at 0; i.e., $p^*(0) > 0$.*

D2. *The distribution F of X_s^2, where each X_s is the marginal regressor, is regularly varying near zero with strictly positive index $(-\rho) > 0$.*

D3. *Further to B7, the bandwidth satisfies $h_j = j^p h$ (i.e. $\phi_j = j^p$) with $p > \frac{1}{2}$.*

D4. *The conditional variance* $\mathrm{Var}[Y_t|X_t = u] = \sigma^2(u)$ *is continuous in some neighbourhood of x; i.e.* $\sup_{u \in \mathcal{E}(x,\underline{h}\lambda)}[\sigma^2(u) - \sigma^2(x)] = o(1)$. *Similarly, the cross-conditional moment $E[(Y_t - m(x))(Y_s - m(x)|X_t = u, X_s = v] = \sigma(u, v)$, $t \neq s$ is continuous in some neighbourhood of (x, x).*

D5. $R_{nt} := (EK_1)^{-1}\{K_t(Y_t - m(x)) - EK_t(Y_t - m(x))\}$ *belongs to the domain of attraction of a normal distribution.*

Remark. Assumption D1, which is similar to what is assumed in Mas (2012), concerns with a transition of the shifted small ball probability to the centred small deviation (whose asymptotic behaviour is more accessible). The explicit form of the derivative (and hence of the relationship between the two probabilities) cannot be easily computed in general. Nonetheless, in the special case of the Gaussian process Z with covariance operator Σ it is known by Sytaya (1974) and Zolotarev (1986) that

$$P\Big(\|z - Z\| \le \epsilon\Big) \simeq P\Big(\|Z\| \le \epsilon\Big) \exp\Big\{ -\frac{1}{2}\|\Sigma^{-1/2}z\|^2 \Big\} \quad \text{as} \quad \epsilon \to 0. \tag{26}$$

The reader is directed to Li and Shao (2001) for detailed discussion on this asymptotic equivalence relation. Note for later reference (cf. $\mathcal{V}_1$ and $\mathcal{V}_2$ in pages 17 and 18) that Σ can be expressed in terms of the a_j constants (in Assumption C) which govern the dependence between the marginal regressors and the bandwidth weights ϕ_j:

$$\mathrm{cov}(Z) = \Sigma = (DA)(A^*D),$$

where $A = (a_{ij}) = (a_{i-j})$ and $D = \mathrm{diag}(\phi_1, \phi_2, \ldots)$.

Condition D2 is equivalent to saying

$$\lim_{x \to \infty} \frac{F(1/(\gamma x))}{F(1/x)} = \gamma^\rho$$

where ρ is the index of variation which is strictly negative. Under the condition, Dunker, Lifshits and Linde (1998, cf. Conditions I and L) derived the explicit behaviour of the small ball probability, which forms a vital part of our results. We require the function $F(1/x)$ to be regularly varying in order to ensure that the small ball probability *well-behaves* (near infinity) in asymptotic sense. Since only those having strictly negative ρ satisfy the condition, the distribution F of the squared regressor must be such that $F(1/x)$ decreases (as $x \to \infty$) at a *reasonable speed.* By reasonable we mean that the relative weight of decrease follows a power law, and the variation should be continuous. A large class of common distributions satisfies this condition; for example, Gamma, Beta, Pareto, Exponential, Weibull, and also Chi-squared distribution (in which case each X_s is Gaussian). Indeed, both D1 and D2 hold under Gaussianity (i.e. when condition C2 is assumed).

The specific bandwidth increment condition assumed in D3 is one framework under which the explicit behaviour of the small ball probability can be specified (cf. Dunker et al. (1998)). In the exceptional case of static regression where the regressors form an i.i.d. sequence, the probability can also be derived when the weights are of an exponential type (i.e. $h_j = e^j h$) up to an unknown function, or are non-increasing in a particular manner (cf. Gao et al. (2011)) similar to the polynomial decay. In this paper however, we shall confine our attention to the case of polynomial law for expositional simplicity and consistency of presentation, since the asymptotic behaviour of the small ball is not yet known in the dependent case for choices other than the polynomial decay as in D3. In practice, we would require some ordering for the marginal regressors in the static regressions case A1, since the influence of marginals is set to decrease via the bandwidth adjustments. The standard conditions in D4 are assumed to deal with the asymptotics of the variance and covariance terms. The last condition is imposed to establish the self-normalized clt without assuming higher moment conditions; relevant discussions can be found e.g. in de la Peña et al. (2009). The condition is not affected by the dependence structure of R_{nt} as the property is inherited to the approximated sum in the Bernstein's blocking procedure; see (70) for details.

With reference to (19) we are now able to derive the following results for the bias and variance components using Assumptions B7, B8 and C, and Corollary 1:

$$\mathcal{B}_n(x) := \Big[E\widehat{m}_2(x) - m(x)\Big] \le h^\beta \lambda^\beta \sum_{j=1}^{\infty} c_j j^{p\beta} \tag{27}$$

$$\mathcal{V}_n(x) := \mathrm{Var}\Big[\widehat{m}_2(x) - E\widehat{m}_2(x)\Big] \simeq \frac{\sigma^2(x)\xi_2}{n\varphi_x(\underline{h}\lambda)\xi_1^2}, \tag{28}$$

where λ and $\widehat{m}_2(\cdot)$ are as in (4) and (18), respectively. Formal derivation is done in 4.2 of the appendix. We now construct the asymptotic normality of our estimator.

3.2.3 Limiting distribution under independence of regressors

We first consider the situation in which there is a set of independent exogenous regressors in the static regression context; that is, when marginal regressors X_s are i.i.d. (i.e. satisfies Assumption C1), and the sample observations follow Assumption A1.

In this case, the asymptotic normality can be established for regressors that follow a wide range of different distributions. Recall that under Assumption D2, the distribution function F (of X^2) is regularly varying with the index of variation $\rho < 0$. Then, by the characterization theorem of Karamata (1933) (see for example Feller (1971)), there always exists a slowly varying function $\ell(x)$ that satisfies

$$F(1/x) = x^\rho \ell(x). \tag{29}$$

Now fix some p, the order of increment constant for bandwidth in Assumption D3, and

$X_j^2 \sim F$ i.i.d.	ρ	$\lim_{x\to\infty} \ell(x) = C_\ell^{-2}$	ζ
Uniform(1,b)	-1	1	n/a
Gamma(α,β)	$-\alpha$	$\beta^\alpha\alpha^{-1}\Gamma(\alpha)^{-1}$	$\frac{\alpha\pi\beta^{-1/2p}}{\sin(\pi/2p)}$
$\exp(\eta)$	-1	η	$\frac{\pi\eta^{-1/2p}}{\sin(\pi/2p)}$
Weibull(α,β)	$-\alpha$	β	n/a
Pareto(θ,μ)	-1	μ/θ	n/a
χ_1^2	$-1/2$	$(2/\pi)^{1/2}$	$\frac{\pi 2^{(1-2p)/2p}}{\sin(\pi/2p)}$

Table 1: Examples of key constants for some common distributions

denote by $\mathcal{L}(t)$ the Laplace transform of X^2. We then define the following constants:

$$C_\ell = \lim_{h\to 0}\left[\ell^{-1/2}\left(h^{-\frac{4p}{2p-1}}\right)\right]$$

$$C^* = \frac{(2\pi)^{(1+2p\rho)}(2p-1)}{\Gamma^{-1}(1-\rho)\cdot(2p)^{\frac{2p(\rho+2)-1}{2p-1}}}\cdot\zeta^{\frac{2p(1+\rho)}{2p-1}}, \qquad C^{**} = (2p-1)\cdot\left(\frac{\zeta}{2p}\right)^{2p/(2p-1)}$$

$$\zeta = -\int_0^\infty \frac{u^{-1/2p}\mathcal{L}'(u)}{\mathcal{L}(u)}du, \quad \text{and} \quad \mathcal{V}_1(x) = \frac{C^* C_\ell \xi_2\sigma^2(x)}{p^*(0)\xi_1^2\lambda^{\frac{1+2\rho p}{2p-1}}},$$

where $\Gamma(\cdot)$ is the Gamma function, ξ_1 and ξ_2 are the constants specified in (11) which cancel out in case of uniform kernel, λ is the upper bound of the support of the kernel, $p^*(\cdot)$ is the Radon-Nikodym derivative in D1. The underlying arguments for the formulation of these constants can be found in Dunker, Lifshits and Linde (1998). To aid the exposition, we compute the constants for some common, regularly varying distributions in Table 1 below. The main result of this subsection now follows. The theorem derives the limiting distribution of the infinite-dimensional Nadaraya-Watson type estimator for every choice of i.i.d. regressors satisfying Assumption D2, and a set of sufficient regularity conditions under which the expression is valid.

THEOREM 2. *Suppose B2-B8 and D1-D4 hold. Let the marginal regressors X_s satisfy Assumption C1. Then the estimator (16) w.r.t. the sample observations $\{Y_t, X_t^\intercal\}_{t=1}^n$ satisfying A1 is asymptotically normal with the following limiting distribution:*

$$\sqrt{nh^{\frac{1+2\rho p}{2p-1}}\exp\left\{-C^{**}(\lambda h)^{-\frac{2}{2p-1}}\right\}}\left[\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\right] \Longrightarrow N\left(0, \mathcal{V}_1(x)\right), \qquad (30)$$

where $\mathcal{B}_n(x) = O(h^\beta)$ is is the bias component as in (27).

3.2.4 Limiting distribution under Gaussianity & dependence of regressors

The strict independence condition between the regressors assumed in the previous section can be relaxed to allow some mild dependence specified in Assumption C2. In doing so, we make use of the result derived by Hong, Lifshits and Nazarov (2016, Theorem 1.1), where the asymptotics of the small deviation probability of Gaussian dependent sequences is investigated. This setting not only grants sufficient flexibility in the static regression case, but moreover allows one to compute the distributional result for the dynamic regression context, where the regressor vector consists of time lags of the response or a covariate with dependence structure stipulated in Assumption A2 (and thereby satisfying C2). The cost we pay for this modification is the Gaussianity assumption on the regressors.

With reference to Table 1 above, we can easily compute the constants C^* and C^{**} for the Gaussian case, denoted C^*_G and C^{**}_G respectively, as follows:

$$C^*_G = \frac{(2\pi)^{(1-p)}(2p-1)}{2\cdot(2p)^{\frac{3p-1}{2p-1}}}\cdot\left[\frac{\pi 2^{(1-2p)/2p}}{\sin(\pi/2p)}\right]^{\frac{-p}{2p-1}}, \qquad C^{**}_G = \frac{2p-1}{2}\left(\frac{\pi}{2p\sin\frac{\pi}{2p}}\right)^{\frac{2p}{2p-1}}.$$

For the square summable sequence a_j in (21) define

$$C_{\mathcal{A}} = \left[\frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{j=-\infty}^{\infty} a_j \exp(\mathrm{i}js)\right|^{1/p} ds\right]^p \quad \text{and} \quad \mathcal{V}_2(x) = \frac{C^*_G C_\ell \xi_2 \xi_1^{-2}\sigma^2(x)}{e^{-\frac{1}{2}\|\Sigma^{-1/2}z\|_2^2}(C_{\mathcal{A}}\lambda)^{\frac{1-p}{2p-1}}}.$$

where $\sigma^2(\cdot)$ is the conditional variance defined in Assumption D4 and $z = (z_j) = (j^{-p}x_j)$. Recall that for uniform (Box) kernel $\xi_2 = \xi_1^2$ so they cancel out in $\mathcal{V}_2(x)$.

With other constants defined as before, we now have the following asymptotic normality for the case of dependent regressors. We reiterate that the result covers both the static and dynamic regressions context (A1 and A2), and is invariant to (30) upto a constant as long as the cross-dependence structure satisfies Assumption C2.

Theorem 3. *Suppose B2-B8 and D1-D4 hold. Let the regressor $X = (X_1, X_2, \ldots)^\intercal$ is jointly normally distributed with zero mean and the covariance operator Σ, and satisfies C2. Then, the estimator (16) with respect to sample observations $\{Y_t, X_t^\intercal\}_{t=1}^n$ satisfying either A1 or A2 is asymptotically normal with the following limiting distribution:*

$$\sqrt{nh^{\frac{1-p}{2p-1}}\exp\Big\{-C^{**}_G(C_{\mathcal{A}}\lambda h)^{-\frac{2}{2p-1}}\Big\}}\Big[\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\Big] \Longrightarrow N\left(0, \mathcal{V}_2(x)\right), \quad (31)$$

where $\mathcal{B}_n(x)$ is the bias component in (27).

Remark. The additional constant $C_{\mathcal{A}}$ is a function of the sequence a_j, and represents the dependence structure allowed between the regressors. This suggests an important finding that says allowing for dependence does not incur much penalty; we conjecture that similar conclusion would hold for regressors of different distributions, and leave it for future studies. The exponential term in the denominator of the asymptotic variance arises from the asymptotic equivalence relationship between the shifted and non-shifted small deviation for ℓ_2-valued Gaussian variables, cf. (26).

In both frameworks of independent regressors and dependent Gaussian regressors we are able to construct the self-normalised central limit theorem; define

$$\Delta_n^2(x) = \sum_{t=1}^n\left(\sum_{s=1}^n K_s\right)^{-2}\left[K_t\big(Y_t - \widehat{m}(x)\big) - \frac{1}{n}\sum_{t=1}^n K_t\big(Y_t - \widehat{m}(x)\big)\right]^2. \quad (32)$$

Corollary 2. *Further to the conditions assumed either in Theorem 2 or Theorem 3, suppose Assumption D5 holds. Then the following central limit theorem holds*

$$\Delta_n^{-1}(x)\Big(\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\Big) \Longrightarrow N(0,1), \tag{33}$$

where $\Delta_n(x)$ is the square root of (32).

Remark. This self-normalized limit distribution gives (pointwise) confidence intervals for $\widehat{m}(x)$, which can be used as a basis for conducting standard statistical inference.

3.3 Optimal Bandwidth

We now briefly discuss the issue of bandwidth choice. As in the finite-dimensional framework, the results in the previous section confirm the existence of the trade-off relationship between the order of the bias and variance terms. As the bandwidth goes up, the upper bound of the variance gets smaller while that of the bias increases, vice versa. Therefore we may search for the optimal bandwidth h_{opt} that balances the order of those two quantities by solving their equivalence relation.

For example, as for the case of Gaussian regressor we have

$$h^{\beta} \sim \sqrt{\frac{\exp\left(Ch^{-2/(2p-1)}\right)}{nh^{\frac{1-p}{2p-1}}}} \tag{34}$$

so that

$$\left[2\beta + \frac{1-p}{2p-1}\right] \cdot \log h - Ch^{-\frac{2}{2p-1}} \sim -\log n.$$

Taking $h \sim (\log n)^a$ for some $a < 0$ balances the leading terms on both sides:

$$\left[2\beta + \frac{1-p}{2p-1}\right] \cdot a \cdot \log\log n - C(\log n)^{-\frac{2}{2p-1}\cdot a} \sim -\log n. \tag{35}$$

The explicit order a that solves (35) can be expressed in terms of n, β and p. Writing $\vartheta := [2\beta + (1-p)/(2p-1)]$ and $\chi := 2/(2p-1)$ for notational simplicity, and solving for a we have

$$a_{opt} = \frac{\vartheta \cdot \mathcal{W}\left(\frac{\chi}{\vartheta} \cdot n^{\chi/\vartheta}\right) - \chi \log n}{\vartheta\chi \cdot \log\log n}, \tag{36}$$

where $\mathcal{W}(y)$ is the Lambert W function (see e.g. Olver et al. (2010)), which returns the solution x of $y = x \cdot e^x$. From (36) the optimal bandwidth $h_{opt} \sim (\log n)^{a_{opt}}$ follows.

Remark. We can look for the optimal bandwidth for the cases of non-Gaussian regressors by following exactly the same manner as above; tedious details are omitted here. As regards the solution in (36), since the mapping $x \mapsto x \cdot e^x$ is not an injection, the solution may be multi-valued on the negative domain, i.e. $y < 0$. This does not happen in (36) provided $\beta \geq 1/4$ (however big p is), because $(1-p)/(2p-1)$ is bounded away from $-1/2$; in this case, the coefficient of the double logarithmic term in (35) is strictly smaller or equal to zero.

Since the log terms dominate the double logarithm in (35) as the sample size n increases, it can be readily expected that the optimal value of a in (36) converges to a limit in such a way that the leading orders are balanced. Below we introduce without formal justification a trivial result that gives the lower bound (infimum) of the optimal bandwidth (and hence of the optimal rate that balances the bias and variance, see also Mas (2012, Theorem 3)). The result holds for other choices of the distribution of the regressors, since the exponent of the leading term $-2/(2p-1)$ remains invariant as it was shown in (30) and (31). Lastly, we also note that p is restricted by the degree of contraction c_j in B8. For example, denoting by $f(\beta)$ the upper bound of p satisfying $\sum_{j=1}^{\infty} c_j j^{p\beta} < \infty$ and $\sum_{j=1}^{\infty} c_j \leq 1$, we have $p < f(\beta) \equiv 1/\beta$ when $c_j = (1/2)j^{-2}$.

COROLLARY 3. *For any fixed choice of $p \in (1/2, f(\beta))$ and the distribution F of X^2 satisfying D2, the order of the optimal bandwidth a_{opt} satisfies*

$$a_{opt} \downarrow \left(-\frac{2p-1}{2}\right) \quad \text{as} \quad n \to \infty, \tag{37}$$

which suggests that the lower bound of the optimal bandwidth is given by

$$(\log n)^{-\frac{2p-1}{2}} \preceq h_{opt} \sim (\log n)^{a_{opt}}. \tag{38}$$

3.4 Uniform consistency

Uniform consistency of the Nadaraya-Watson estimator was first studied by Nadaraya (1964, 1970) and subsequently by numerous others. To name a few earlier literature, Devroye (1978) moderated the regularity conditions required in the previous papers, and Robinson (1983) proved consistency with respect to dependent sample data. In functional statistics literature, only uniform consistency with respect to i.i.d. sample has been established so far, see Ferraty et al. (2010). As it was done in their paper we introduce the notion of Kolmogorov's entropy.

DEFINITION 5. *Given some $\eta > 0$, let $L(S, \eta)$ be the smallest number of open balls in E of radius η needed to cover the set $S \subset E$. Then Kolmogorov's η-entropy is defined as $\log L(S, \eta)$.*

From the definition it can be readily expected that Kolmogorov's entropy is dependent heavily on the nature of the space we work on, and is closely related to the rate of convergence of the estimator.

It is well known that the regression function cannot be estimated uniformly over the entire space, see e.g. Bosq (1996). In our infinite dimensional framework, even greater restrictions apply; since we are working on infinite sequence spaces, none of their subsets can be covered by a finite number of balls, so that $L(S, \eta) = \infty$. Therefore, we propose to adopt a truncation argument and consider uniform consistency over a set whose effective dimension is increasing in sample size n. In particular, we define the set

$$S_\tau := \left\{u | (u_i)_{i \in \mathbb{Z}^+}, u_j = 0 \text{ for all } j > \tau, \|u\|_\infty \leq \lambda\right\} \tag{39}$$

where $\tau = \tau_n$ is some increasing sequence and λ is fixed, and consider uniform consistency over this compact set.

Then Kolmogorov's entropy of the set S_τ is given as follows:

LEMMA 2. *Kolmogorov's η-entropy of S_τ defined in (39) with $\tau = \tau_n(\to \infty)$ and $\lambda > 0$ is*

$$\log L\big(S_\tau, \eta\big) = \log \left[\left(\frac{2\lambda\sqrt{\tau}}{\eta} + 1 \right)^\tau \right]. \tag{40}$$

REMARK. We note that (40) is indeed in line with common intuition; as the dimension τ increases, the number of balls with some fixed radius required to cover the set goes off to infinity. The proof of this result can be done by exploiting the splitting technique and then by attempting to cover the polyhedron of increasing dimension. See appendix for details. From this result it follows that for fixed λ and $\eta = \eta_n$, Kolmogorov's entropy $\log L(S_\tau, \eta)$ is of order $O(\tau \log \tau - \tau \log \eta)$.

We now introduce some further assumptions needed for uniform consistency:

ASSUMPTIONS E

E1. *For sufficiently large n, Kolmogorov's η-entropy $\log L(S_\tau, \eta)$ satisfies*

$$\frac{(\log n)^{8+2\epsilon}}{n\varphi_x(\underline{h})} \leq \log L(S_\tau, \eta) \leq \frac{\sqrt{n\varphi_x(\underline{h})}}{(\log n)^{1+\epsilon}} \quad \text{for some} \quad \epsilon \in (0, 1/2). \tag{41}$$

Furthermore, $0 < \varphi_x(\underline{h}) \preceq h < \infty$ *and* $(\log n)^2/(n\varphi_x(\underline{h})) \longrightarrow 0$ *as* $n \to \infty$.

E2. *The kernel function K is Lipschitz continuous on* $[0, \lambda]$.

REMARK. The first part of assumption E1 specifies the rate at which Kolmogorov's entropy should behave in sample size n (hence in dimension $\tau = \tau_n$). From the upper and lower bound it readily follows that $n\varphi(h)$ must be of order bigger than $(\log n)^{6+2\epsilon}$. This assumption allows sufficient generality; for example, the restriction that the bias-variance optimal bandwidth satisfies $h \succeq (\log n)^{-(2p-1)/2}$ gives $n\varphi(h) \preceq (\log n)^{(2p-1)\beta}$. In this case, assumption (41) is valid as long as p is moderately large enough relative to $\beta \leq 1$ in such a way that $6 + 2\epsilon \leq (2p-1)\beta$. The second part of E1 is standard and the last condition straightforwardly follows by (41) and only slightly strengthens the bandwidth condition in Assumption B2.

We now introduce the main result of this section. Note that in the sequel, (with a slight abuse of notation) X is taken to denote the regressor, but with zeros after its $\tau^{th}(= \tau_n \to \infty$ as $n \to \infty)$ entry; that is, $X = (X_1, X_2, \ldots, X_\tau, 0, 0, \ldots)^\intercal$ (so that the original X is recovered as $n \to \infty$). Also, the regression operator and the estimator with respect to this truncated regressor are denoted by $m_\tau(\cdot)$ and $\widehat{m}_\tau(\cdot)$, respectively. The aforementioned assumptions are understood to be modified accordingly.

For uniform consistency we impose a slightly stricter condition on the response:

B4′. The response Y_t is satisfies the following tail condition: There exists some positive constant γ_1 and C such that $P(|Y_t| > u) \leq C \exp(1 - u^{\gamma_1})$ for any $u > 0$.

For example, a Gaussian random variable satisfies B4$'$ with $\gamma_1 < 2$. The condition is also satisfied by many unbounded variables and all those bounded ones as well. We also impose a stronger condition on mixing coefficients; from hereafter, by A1$'$ and A2$'$ we mean Assumptions A1 and A2 but with the arithmetic mixing rate condition strengthened to the following exponential mixing condition (cf. Definition 1):

$$\alpha(r) \leq \exp(-\varsigma r^{\gamma_2}) \tag{42}$$

where $\varsigma > 1$ and γ_2 is a positive constant such that $\gamma := 1/(\gamma_1^{-1} + \gamma_2^{-1}) \geq 1$. In case of the bounded response (i.e. $|Y_t| \leq C$), γ_1 is taken to be ∞ so that $\gamma_2 = \gamma \geq 1$.

THEOREM 4. *Suppose that Assumptions B2, B3, B4$'$, B5-B8, D1-D3 and E1-E2 hold. Let the marginal regressors X_s satisfy C1, and take $\tau = \tau_n \sim (\log n)$. Then the estimator $\widehat{m}_\tau(\cdot)$ with respect to sample observations $\{Y_t, X_t\}_{t=1}^n$ satisfying A1$'$ is uniformly consistent for $m(x) = m(x_1, \ldots)$ over S_τ:*

$$\sup_{x \in S_\tau} \left| \widehat{m}_\tau(x) - m_\tau(x) \right| = O_P\left(h^\beta + \sqrt{\frac{(\log n)^2 \exp\left(h^{-2/(2p-1)}\right)}{n h^{\frac{1-p}{2p-1}}}} \right). \tag{43}$$

If alternatively X_s is Gaussian and satisfies C2, then the same conclusion holds with respect to sample observations satisfying either A1$'$ or A2$'$.

REMARK. We may choose the optimal bandwidth as before; following the same arguments in the pointwise case, choosing $h \sim (\log n)^a$ and solving for n gives

$$a_{opt} = \frac{\vartheta \cdot \mathcal{W}\left[\frac{\chi}{\vartheta} \exp(-\frac{\chi}{\vartheta} 2 \log\log n - \chi \log n)\right] + 2\chi \log\log n - \chi \log n}{\vartheta \chi \log\log n}.$$

And because the order of the leading terms is $(\log n)^{-(2p-1)/2}$ as in the pointwise case, it is straightforward to see that the lower bound of the optimal bandwidth in Corollary 2 still continues to hold; that is, $h_{opt} \succeq (\log n)^{-(2p-1)/2}$. This is again invariant to the choice of distribution F of the squared regressor.

COROLLARY 4. *Suppose conditions assumed in Theorem 4 hold. Then, upon choosing $h \sim (\log n)^{a_{opt}}$, we have*

$$\sup_{x \in S_\tau} \left| \widehat{m}_\tau(x) - m_\tau(x) \right| = O_P\left([\log n]^{\beta \cdot a_{opt}} \right). \tag{44}$$

Without going into details we briefly remark that in case of bounded response, a special case of B4$'$ as aforementioned, the results in this section can be strengthened to almost sure sense due to availability of suitable exponential inequalities.

4 Concluding remarks

In this paper we studied the nonparametric estimation problem of the regression function when the number of regressor is potentially infinite. Both pointwise and uniform consistency results are obtained with explicit logarithmic rates of convergence, and the asymptotic normality is established. Our theory is flexible in the sense that it applies to a number of different static and dynamic regressive frameworks.

The theories presented in this paper can be readily extended to estimate the generalized conditional expectation $E(f(Y)|X=\cdot)$, where f is some known Borel function. A straightforward example is the conditional variance, which is particularly important in financial applications. Other quantities of interest in prediction such as the conditional median or mode can also be studied. This could be done via nonparametrically estimating the conditional distribution $P(Y \leq y|X=\cdot) = E(1\{-\infty, y](Y)|X=\cdot)$, but would necessarily require a slightly different set of assumptions.

Whether an algebraic convergence rate can be obtained by assuming an additive structure remains unanswered. If the X_j's are i.i.d., then one can estimate an additive model sequentially. In this case one could also estimate average partial effect in a general model by just estimating the one dimensional average derivative.

5 Appendix: Proofs of the main results

5.1 Proof of Theorem 1

From the decomposition (19):

$$\widehat{m}(x) - m(x) = \frac{E\widehat{m}_2(x) - m(x)}{\widehat{m}_1(x)} + \frac{\widehat{m}_2(x) - E\widehat{m}_2(x)}{\widehat{m}_1(x)} - \frac{m(x)[\widehat{m}_1(x) - 1]}{\widehat{m}_1(x)},$$

we see that it suffices to show $E\widehat{m}_2(x) - m(x) \to 0$ and $\widehat{m}_2(x) - E\widehat{m}_2(x) \to^P 0$, since $\widehat{m}_1(x) \to^P 1$ would then follow from the latter and complete the proof.

As for the former 'bias component', denoting by $\mathcal{E}(x, \lambda\underline{h})$ the infinite dimensional hyperellipsoid centred at $x = (x_j)_j \in \mathbb{R}^\infty$ with semi-axes h_j in each direction we have

$$\begin{aligned} E\widehat{m}_2(x) - m(x) &= E\left(\frac{1}{nEK_1}\sum_{t=1}^n K_tY_t - m(x)\right) \\ &= \frac{1}{EK_1}EK_1Y_1 - \frac{EK_1}{EK_1}m(x) = \frac{1}{EK_1}E\Big[E\Big[\big(Y_1 - m(x)\big)K_1\Big|X\Big]\Big] \\ &= \frac{1}{EK_1}E\Big[\Big[m(X) - m(x)\Big]K_1\Big] \leq \sup_{u\in\mathcal{E}(x,\lambda\underline{h})}\big|m(u) - m(x)\big| \longrightarrow 0 \quad (45) \end{aligned}$$

as $n \to \infty$, where K_t is the shorthand notation for $K(\|H^{-1}(x - X_t)\|)$ as introduced in the main text before. The second equality is justified by stationarity that is preserved under measurable transformation, and the last inequality is due to compact support of the kernel and continuity of the regression operator at x (Assumption B1).

The next step concerns with the latter 'variance component' $\widehat{m}_2 - E\widehat{m}_2$; its mean-squared convergence to zero will be shown. Writing

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n}\sum_{t=1}^n \frac{1}{EK_1}\Big\{K_tY_t - E(K_tY_t)\Big\} =: \frac{1}{n}\sum_{t=1}^n Q_{nt}, \quad (46)$$

we remark that the arguments to follow depend upon the temporal dependence structure of Q_{nt}. In the static regression case, Q_{nt} is a measurable function of $Y_t, X_{1t}, X_{2t}, \ldots$, and hence inherits their joint dependence structure. That is, Q_{nt} is arithmetically α-mixing with the rate specified in A1. In the dynamic regressions case (which covers the autoregression framework), the dependence of Q_{nt} is defined via K_t which is near epoch dependent on (Y_t, V_t) as specified in Assumption A2; this bypasses the issue of Q_{nt} being dependent upon infinite past of Y_t and/or V_t. We proceed with these two cases separately.

CASE 1: STATIC REGRESSION. Clearly, it is sufficient to prove $\mathrm{Var}(\widehat{m}_2 - E\widehat{m}_2) \to 0$ for mean squared convergence. Since Q_{nt} is stationary over time we have

$$\mathrm{Var}(\widehat{m}_2 - E\widehat{m}_2) = \frac{1}{n^2}\sum_{t=1}^{n}\mathrm{Var}(Q_{nt}) + \frac{2}{n^2}\sum\sum_{1\le i<j\le n}\mathrm{Cov}\big(Q_{ni}, Q_{nj}\big) \tag{47}$$

$$= \frac{1}{n}\mathrm{Var}(Q_{n1}) + \frac{2}{n^2}\sum\sum_{1\le j-i<n}\mathrm{Cov}\big(Q_{ni}, Q_{nj}\big)$$

$$= \frac{1}{n}\mathrm{Var}(Q_{n1}) + \frac{2}{n^2}\sum_{s=1}^{n-1}(n-s)\cdot\mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big) =: A_1 + A_2. \tag{48}$$

Now, by (8), (10) and Assumption A it follows that

$$A_1 = \frac{1}{nE^2K_1}\mathrm{Var}\left(K_1Y_1 - EY_1K_1\right) = \frac{\mathrm{Var}\,(K_1Y_1)}{nE^2K_1}$$

$$\le \frac{EK_1^2Y_1^2}{nE^2K_1} = \frac{E(E(Y_1^2|X_1)K_1^2)}{nE^2K_1} \le \frac{C}{n\varphi_x(\lambda\underline{h})} \to 0 \tag{49}$$

as $n\to\infty$.

We now move on to the second term A_2 and investigate the covariance term. Since measurable transformations of mixing variables preserve the mixing property, using Davydov's inequality, see Davydov (1968, Lemma 2.1) or Bosq (1996, Corollary 1.1) and stationarity we have

$$\big|\mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| = \left|\mathrm{Cov}\left(Y_1\frac{K_1}{EK_1}, Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right| \le \frac{C\{E|Y_1K_1|^{2+\delta}\}^{\frac{2}{2+\delta}}}{\varphi_x(\underline{h}\lambda)^2\cdot s^{k\delta/(2+\delta)}}. \tag{50}$$

In the meantime,

$$\big|\mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| = \left|\mathrm{Cov}\left(Y_1\frac{K_1}{EK_1}, Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right|$$

$$\le \left|E\left(Y_1\frac{K_1}{EK_1}Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right| + \left|E\left(Y_1\frac{K_1}{EK_1}\right)E\left(Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right|$$

$$\le \frac{C}{\varphi_x(\underline{h}\lambda)^2}\left|E\left(K_1K_{s+1}\right)\right| + \frac{C'}{E^2K_1}\left|E\left(K_1\right)E\left(K_{s+1}\right)\right|$$

$$\le \frac{C}{\varphi_x(\underline{h}\lambda)^2}\cdot\psi_x(\lambda\underline{h}; 1, s+1) + C' \le C'' \tag{51}$$

by stationarity, law of iterated expectation, boundedness of regression function, and Assumption B6, B5 (along with the upper bound $\psi(\lambda\underline{h}; 1, s+1)$ of EK_1K_{s+1} obtained as a direct consequence of B5 following similar arguments used for Lemma 1).

With reference to (50) and (51), we take some increasing sequence $u_n \to \infty$ such that $u_n = o(n)$, and write

$$\sum_{s=1}^{n-1} \big|\mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| = \sum_{s=1}^{u_n-1} \big|\mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| + \sum_{s=u_n}^{n-1} \big|\mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big)\big|$$
$$\le C''\big(u_n - 1\big) + \sum_{s=u_n}^{n-1} \frac{C s^{-k\delta/(2+\delta)}}{\varphi_x(\underline{h}\lambda)^2} = O\left(u_n + \frac{u_n^{-k\delta/(2+\delta)+1}}{\varphi_x(\underline{h}\lambda)^2}\right), \qquad (52)$$

which is $O\big(\varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)}\big)$ upon choosing $u_n \sim \varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)}$.

Consequently, since $k \ge 2(2+\delta)/\delta$ it follows that

$$A_2 := \frac{2}{n^2}\sum_{s=1}^{n-1}(n-s)\cdot \mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big) = \frac{2}{n}\sum_{s=1}^{n-1}\left(1 - \frac{s}{n}\right)\cdot \mathrm{Cov}\big(Q_{n1}, Q_{n,s+1}\big)$$
$$= O\big(n^{-1}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)} + n^{-2}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)}\big)$$
$$= O\big(n^{-1}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)}\big) = o(1) \qquad (53)$$

by Assumption B2, and the desired result is obtained.

Case 2: Dynamic Regression.[5] We return back to (46):

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big\{K_tY_t - E(K_tY_t)\Big\} =: \frac{1}{n}\sum_{t=1}^{n} Q_{nt}. \tag{54}$$

In this framework $K_t = K(\|H^{-1}(x - X_t)\|)$ is a (measurable) function of $(Y_{t-1}, Y_{t-2}, \ldots)$. Despite loosing the mixing property, K_t inherits stationarity of the mixing process $\{Y_t\}$. We write $K_{t,(r)} = \Psi(Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-r+1}) = E(K_t|Y_t, \ldots, Y_{t-r+1})$, where Ψ denotes a measurable map and r is as in Assumption A2. Clearly, $K_{t,(r)}$ preserves the mixing dependence structure of Y_t with mixing coefficient $\alpha(\ell - (r-1))$ since $\sigma(K_{s,(r)};\ s \geq t + \ell) \subset \sigma((Y_s, \ldots, Y_{s-r+1});\ s \geq t + \ell) = \sigma(Y_s;\ s \geq t + \ell - (r-1))$.

Now write

$$\begin{aligned}\widehat{m}_2 - E\widehat{m}_2 &= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[K_{t,(r)}Y_t - E\big(K_{t,(r)}Y_t\big)\Big] + \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[K_tY_t - K_{t,(r)}Y_t\Big] \\ &\quad + \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[E\big(K_{t,(r)}Y_t\big) - E(K_tY_t)\Big] = R_1 + R_2 + R_3,\end{aligned} \tag{55}$$

and first consider the last term R_3.

Fix some increasing sequence $q = q_n \to \infty$, and write $Y_{t,L} := Y_t 1\{|Y_t| \leq q\}$ and $Y_{t,U} = Y_t 1\{|Y_t| > q\}$. Then

$$\begin{aligned}EY_tK_{t,(r)} &= EY_tK\big(\|H^{-1}(x - X_t)\|\big) - EY_{t,U}K\big(\|H^{-1}(x - X_t)\|\big) \\ &\quad + EY_{t,L}K_{t,(r)} - EY_{t,L}K\big(\|H^{-1}(x - X_t)\|\big) \\ &\quad + EY_{t,U}K_{t,(r)} = D_1 + D_2 + D_3.\end{aligned} \tag{56}$$

The second part of D_1 is given by

$$\begin{aligned}&EY_{t,U}K\big(\|H^{-1}(x - X_t)\|\big) \leq E|Y_t|1_{\{|Y_t|>q\}}K\big(\|H^{-1}(x - X_t)\|\big) \\ &\leq q^{-(\delta+1)}E|Y_t|^{2+\delta}1_{\{|Y_t|>q\}}K_t \leq Cq^{-(\delta+1)}E|Y_t|^{2+\delta}1_{\{|Y_t|>q\}} = o(q^{-(\delta+1)})\end{aligned} \tag{57}$$

because $1_{\{|Y_t|>q\}} = o(1)$ as $n \to \infty$. Following similar arguments on D_3 we have $D_1 + D_3 = EY_tK_t + o(q^{-(\delta+1)})$. So we are now left with the middle term D_2:

$$D_2 \leq E\,|Y_{t,L}|\,\big|K_t - K_{t,(r)}\big| = O\left(q\sqrt{v_2(r_n)}\right) \tag{58}$$

by Hölder's inequality. Therefore, from (56), (57) and (58) we see that

$$R_3 = \frac{1}{nEK_1}\sum_{t=1}^{n}\Big[EK_{t,(r)}Y_t - E(K_tY_t)\Big] = o\left(\frac{q^{-(\delta+1)}}{\varphi_x(\lambda\underline{h})}\right) + O\left(\frac{q\sqrt{v_2(r_n)}}{\varphi_x(\lambda\underline{h})}\right), \tag{59}$$

and upon choosing $q = (\varphi_x(\underline{h}\lambda)/n)^{-1/(2(\delta+1))}$ we have $o(\varphi^{-1}q^{-(\delta+1)}) = o(\varphi^{-1}(\varphi/n)^{1/2}) = o(n^{-1/2}\varphi^{-1/2}) = o(1)$. Furthermore,

$$\begin{aligned}O\left(\frac{1}{\varphi_x(\underline{h}\lambda)}q\sqrt{v_2(r_n)}\right) &= O\left(\frac{1}{\varphi_x(\underline{h}\lambda)}\cdot\left(\frac{\varphi_x(\underline{h}\lambda)}{n}\right)^{-1/(2(\delta+1))}\sqrt{v_2(r_n)}\right) \\ &= O\left(\frac{\sqrt{v_2(r_n)}}{[\varphi_x(\underline{h}\lambda)]^{(2\delta+3)/(2\delta+2)}n^{-1/(2(\delta+1))}}\right) = o(1)\end{aligned} \tag{60}$$

[5] For the sake of notational simplicity, we will write the proofs for the dynamic regression framework in terms of its autoregressive special case throughout the appendix. That is, some lags of the response variable Y_t here possibly represent the lags of the covariate V_t.

by Assumption A2, yielding $R_3 = o(1)$, and hence $R_2 = o_p(1)$.

As for the first term that remains,

$$
\begin{aligned}
R_1 &= \frac{1}{n}\sum_{t=1}^{n}\left[\frac{K_{t,(r)}Y_t - E(K_tY_t)}{EK_1}\right] + \frac{1}{n}\sum_{t=1}^{n}\left[\frac{E(K_tY_t) - E(K_{t,(r)}Y_t)}{EK_1}\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} E\big(Q_{nt}|Y_t, Y_{t-1}, \ldots, Y_{t-r+1}\big) - R_3 \\
&= \frac{1}{n}\sum_{t=1}^{n} Q_{nt,(r)} + o\left(\frac{q^{-(\delta+1)}}{\varphi_x(\underline{h}\lambda)}\right) + O\left(\frac{\sqrt{v_2(r_n)}}{[\varphi_x(\underline{h}\lambda)]^{(2\delta+3)/(2\delta+2)}n^{-1/(2(\delta+1))}}\right). \qquad (61)
\end{aligned}
$$

Since $Q_{nt,(r)}$ is α-mixing, we can work with the first term by following similar arguments in the regression case. Specifically, due to boundedness of the kernel and the mixing properties, the bound in (50) can be constructed. As for the constant bound constructed in (51), we rewrite

$$
\begin{aligned}
\frac{\mathrm{Cov}\big(Y_1K_{1,(r)}, Y_{s+1}K_{s+1,(r)}\big)}{\varphi_x(\lambda\underline{h})^2} &= \frac{\mathrm{Cov}\big(Y_1[K_{1,(r)} - K_1],\ Y_{s+1}[K_{s+1,(r)} - K_{s+1}]\big)}{\varphi_x(\lambda\underline{h})^2} \\
&+ \frac{\mathrm{Cov}\big(Y_1[K_{1,(r)} - K_1], Y_{s+1}K_{s+1,(r)}\big)}{\varphi_x(\lambda\underline{h})^2} \\
&+ \frac{\mathrm{Cov}\big(Y_1, Y_{s+1}[K_{s+1,(r)} - K_{s+1}]\big)}{\varphi_x(\lambda\underline{h})^2} + \frac{\mathrm{Cov}\big(Y_1K_1, Y_{s+1}K_{s+1}\big)}{\varphi_x(\lambda\underline{h})^2} \\
&= \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 + \mathcal{G}_4.
\end{aligned}
$$

By (50), $\mathcal{G}_4 \le C$ by (51). Further,

$$
\begin{aligned}
\mathcal{G}_1 \le &\left|\frac{E(Y_1Y_{s+1}[K_{1,(r)} - K_1][K_{s+1,(r)} - K_{s+1}])}{\varphi_x(\lambda\underline{h})^2}\right| \\
&+ \left|\frac{E(Y_1[K_{1,(r)} - K_1]) \cdot E(Y_{s+1}[K_{s+1,(r)} - K_{s+1}])}{\varphi_x(\lambda\underline{h})^2}\right| \le C'\frac{v_2(r)}{\varphi_x(\lambda\underline{h})^2} \to 0
\end{aligned}
$$

by Assumption B6 and by the fact that

$$
\left(\frac{\sqrt{v_2(r_n)}}{\varphi_x(\underline{h}\lambda)}\right) \le \left(\frac{\sqrt{v_2(r_n)}}{\varphi_x(\underline{h}\lambda)}\right) \cdot (n/\varphi)^{1/(2\delta+2)} \longrightarrow 0
$$

by (17) in Assumption A2. Similarly, $\mathcal{G}_2$ and $\mathcal{G}_3$ can be easily shown to converge to zero in large sample.

Now choosing an increasing sequence $u_n \sim [\varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)} + r_n] \to \infty$ such that $r_n/u_n = o(1)$, we see that (ignoring the array notation in $Q_{nt,(r)}$ for simplicity)

$$
\begin{aligned}
\sum_{s=1}^{n-1}\big|\mathrm{Cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\big| &= \sum_{s=1}^{u_n-1}\big|\mathrm{Cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\big| + \sum_{s=u_n}^{n-1}\big|\mathrm{Cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\big| \\
&\le C\big(\varphi_x(\underline{h}\lambda)^{-\frac{2(2+\delta)}{(k\delta)}} + r_n\big) + \sum_{s=u_n}^{n-1}\frac{C(s - r_n + 1)^{-k\delta/(2+\delta)}}{\varphi_x(\underline{h}\lambda)^2} = O\Big(\varphi_x(\underline{h}\lambda)^{-\frac{2(2+\delta)}{(k\delta)}}\Big),
\end{aligned}
$$

since the mixing coefficient for $Q_{nt,(r)}$ denoted $\alpha'(n)$ is given by $\alpha(n-(r-1))$ for $n \ge r$. It now follows by the same arguments in (53) that the first term in (61) converges to zero, and hence $R_1 = o_p(1)$, which is the result we desired. ■

5.2 Proof of Theorem 2 and 3

We start by recalling the bias component discussed in (45). Additional assumptions B7, B8 and D3 allow us to proceed further as follows:

$$\begin{aligned}
\mathcal{B}_n(x) &= E\widehat{m}_2(x) - m(x) = E\bigg(\frac{1}{nEK_1}\sum_{t=1}^{n} K_tY_t - m(x)\bigg) \\
&= \frac{1}{EK_1}EK_1Y_1 - \frac{EK_1}{EK_1}m(x) = \frac{1}{EK_1}E\Big[E\Big[\big(Y_1 - m(x)\big)K_1\Big|X\Big]\Big] \\
&= \frac{1}{EK_1}E\Big[\Big[m(X) - m(x)\Big]K_1\Big] \le \sup_{u\in\mathcal{E}(x,\lambda\underline{h})}\big|m(u) - m(x)\big| \\
&\le \sup_{u\in\mathcal{E}(x,\lambda\underline{h})}\sum_{j=1}^{\infty} c_j\big|u_j - x_j\big|^{\beta} = \sum_{j=1}^{\infty} c_j(\lambda h\phi_j)^{\beta} = h^{\beta}\bigg(\lambda^{\beta}\sum_{j=1}^{\infty} c_j j^{p\beta}\bigg) < \infty. \qquad (62)
\end{aligned}$$

Now rewriting the decomposition (19) as

$$\begin{aligned}
\widehat{m}(x) - m(x) - \mathcal{B}_n(x) \\
&= \frac{\mathcal{B}_n(x)\cdot\big[1 - \widehat{m}_1(x)\big]}{\widehat{m}_1(x)} + \frac{\widehat{m}_2(x) - E\widehat{m}_2(x) - m(x)\big[\widehat{m}_1(x) - 1\big]}{\widehat{m}_1(x)},
\end{aligned}$$

and noting that $\widehat{m}_1(x) \to^p 1$ (an immediate consequence of Theorem 1), we see that it suffices to derive the limiting distribution of

$$\begin{aligned}
\widehat{m}_2(x) &- E\widehat{m}_2(x) - m(x)[\widehat{m}_1(x) - 1] \\
&= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[K_tY_t - m(x)K_t - E(K_tY_t) + m(x)EK_t\Big] =: \frac{1}{n}\sum_{t=1}^{n} R_{nt}. \qquad (63)
\end{aligned}$$

By Assumption B6, D3, D4, and the law of iterated expectations, the asymptotic variance of the triangular array R_{nt} is given by

$$\begin{aligned}
\mathrm{Var}(R_{nt}) &= \frac{\mathrm{Var}[K_t(Y_t - m(x))]}{E^2K_1} \\
&\simeq \frac{1}{E^2K_1}\Big\{E\Big[K_t(Y_t - m(x))\Big]^2 - E^2\big[K_t(Y_t - m(x))\big]\Big\} \\
&= \frac{1}{E^2K_1}\bigg\{E\big[\sigma^2(X)K_1^2\big] + E\bigg(\Big[m(X) - m(x)\Big]^2K_1^2\bigg)\bigg\} \\
&= \frac{1}{E^2K_1}\bigg\{\sigma^2(x)EK_1^2 + E\bigg(\Big[\sigma^2(X) - \sigma^2(x)\Big]K_1^2\bigg) + o(1)EK_1^2\bigg\} \\
&= \frac{EK_1^2}{E^2K_1}(\sigma^2(x) + o(1)) \simeq \frac{\sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2}. \qquad (64)
\end{aligned}$$

Following similar arguments and using the latter assumption of D4, it can be readily shown that the covariance term is of smaller order than (64), which together shows (28). Under Assumption D1 the small ball probability can be written in terms of the centered small deviation and $p^*(\cdot)$, the Radon-Nikodym derivative of the induced probability measure P_{z-Z} with respect to P_Z:

$$\begin{aligned}\varphi_x(\lambda\underline{h}) &= P\big(X \in \mathcal{E}(x, \lambda\underline{h})\big)\\ &= P\bigg(\sum_{j=1}^{\infty} j^{-2p}\big(x_j - X_j\big)^2 \le h^2\lambda^2\bigg) = P\Big(\|z - Z\| \le h\lambda\Big)\\ &= \int_{B(0,h\lambda)} dP_{z-Z}(u) = \int_{B(0,h\lambda)} p^*(u)\, dP_Z(u)\\ &\simeq p^*(0)\cdot P(\|Z\| \le h\lambda) = p^*(0) \times P\left(\sum_{j=1}^{n} j^{-2p} X_j^2 \le h^2\lambda^2\right), \end{aligned} \tag{65}$$

where the latter probability can be explicitly specified by substituting $r = h^2\lambda^2$, $A = 2p$, and $a = 2p/(2p-1)$ in Proposition 4.1 of Dunker et al. (1998) for the i.i.d. case. As for the case where the regressors are dependent, i.e. when the X_j's satisfy Assumption C2, the small ball probability can be specified in view of Theorem 1.1 of Hong, Lifshits and Nazarov (2016). Finally we have,

$$\frac{\sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2} = \frac{1}{\phi(h)}\cdot\frac{\sigma^2(x)\xi_2}{p^*(0)\xi_1^2}\cdot\frac{C^*C_\ell}{\lambda^{\frac{1+2\rho p}{2p-1}}}, \tag{66}$$

where $\phi(h) = h^{(1+2\rho p)/(2p-1)}\exp\{-C^{**}(\lambda h)^{-2/(2p-1)}\}$, and (as defined before)

$$C_\ell = \lim_{h\to 0}\left[\ell^{-1/2}\Big(h^{-\frac{4p}{2p-1}}\Big)\right] \qquad C^* = \frac{(2\pi)^{(1+2p\rho)}(2p-1)}{\Gamma^{-1}(1-\rho)\cdot(2p)^{\frac{2p(\rho+2)-1}{2p-1}}}\cdot\zeta^{\frac{2p(1+\rho)}{2p-1}}$$

and $\Gamma(\cdot)$ is the Gamma function, ξ_1 and ξ_2 are the constants specified in (11), and λ is the upper bound of the support of the kernel.

In constructing the central limit theorem we consider the normalized statistic $R_{nt}^* := \sqrt{\phi(h)}\cdot R_{nt}$ and derive the self normalized limiting distribution of $(1/\sqrt{n})\cdot R_{nt}^*$. We shall only prove the autoregression case, where an additional step of mixing approximation is added to the standard regression case; the asymptotic normality for the regression case in a functional context was established in Masry (2005). In many places of the remainder of this proof we shall closely follow their proof of Theorem 4.

We make use of Bernstein's blocking method and partition $\{1,\ldots,n\}$ by $2k(= 2k_n \to \infty)$ number of blocks of two different sizes that alternate (hereafter referred to as the "big" and "small" blocks) and lastly a single block (the "last block") that covers the remainder. The size of the alternating blocks is given by a_n and b_n respectively, where the one for the "big-blocks" a_n is set to dominate that for the "small-blocks" b_n in large sample, i.e. $b_n = o(a_n)$. More specifically, we take

$$k_n = \lfloor n/(a_n + b_n)\rfloor \quad \text{and} \quad a_n = \lfloor\sqrt{n\phi(h)}/q_n\rfloor$$

where $q_n \to \infty$ is a sequence of integer; it then clearly follows that $a_n/n \to 0$ and $a_n/\sqrt{n\phi(h)} \to 0$. We also assume $(n/a_n)\cdot\alpha^*(b_n) = (n/a_n)\cdot\alpha(b_n - r + 1) \to 0$, where α^* is the mixing coefficient of $R_{nt,(r)}^* = E(R_{nt}^*|\mathcal{F}_{t-r+1}^{t-1})$.

By construction above we can write $\sqrt{n}^{-1}\sum_{t=1}^{n} R_{nt}^{*}$ as the sum of the groups of big-blocks $\mathcal{B}$, small-blocks $\mathcal{S}$ and the remainder block $\mathcal{R}$ defined as

$$\mathcal{B} := \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\Xi_{1,j} = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt}^{*}\right)$$

$$\mathcal{S} := \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\Xi_{2,j} = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R_{nt}^{*}\right)$$

$$\mathcal{R} := \frac{1}{\sqrt{n}}\Xi_{3,j} = \frac{1}{\sqrt{n}}\left(\sum_{t=k(a+b)+1}^{n} R_{nt}^{*}\right).$$

The aim is to show that the contributions from the small and the last remaining block are negligible, and that the big-blocks are asymptotically independent.

We first consider the big blocks $\mathcal{B}$. Given r as in Assumption 2, and $R_{nt,(r)}^{*} = E(R_{nt}^{*}|Y_t,\ldots,Y_{t-r+1})$ we have

$$\mathcal{B} = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt,(r)}^{*}\right) + \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a}\left[R_{nt,(r)}^{*} - R_{nt}^{*}\right]\right) = \mathcal{Q}_1 + \mathcal{Q}_2.$$

As for the second term, consider

$$\begin{aligned}
\frac{1}{\sqrt{n}}E\mathcal{Q}_2 &\leq \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\sum_{t=j(a+b)+1}^{j(a+b)+a} E\left|R_{nt,(r)}^{*} - R_{nt}^{*}\right| \\
&= \frac{1}{EK_1}\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\sum_{t=j(a+b)+1}^{j(a+b)+a} E\left|K_tY_t - Y_tE(K_t|Y_t,Y_{t-1},\ldots,Y_{t-r+1})\right| \\
&\leq \frac{1}{\sqrt{n}}\frac{1}{\varphi_x(\underline{h}\lambda)}\sum_{j=0}^{k-1}\sum_{t=j(a+b)+1}^{j(a+b)+a} E|Y_t||K_t - K_{t,(r)}| \\
&\leq \frac{1}{\sqrt{n}}\frac{1}{\varphi_x(\underline{h}\lambda)}\sum_{j=0}^{k-1}\sum_{t=j(a+b)+1}^{j(a+b)+a}\left(E|Y_t|^2\right)^{1/2}\left(E|K_t - K_{t,(r)}|^2\right)^{1/2} \\
&\leq C\cdot\frac{1}{\sqrt{n}}k_na_n\frac{\sqrt{v_2(r_n)}}{\varphi_x(\lambda\underline{h})} = O\left(\frac{\sqrt{n\cdot v_2(r_n)}}{\varphi_x(\lambda\underline{h})}\right) = o(1),
\end{aligned}$$

which implies that $\sqrt{n}^{-1}\mathcal{Q}_2 = o_p(1)$.

We now show asymptotic independence of terms in $\mathcal{Q}_1$, on noting that $\Xi_{1,j}'s$ are independent if for all real t_j

$$\left|E\left[\sum_{j=0}^{k-1}\exp\left(\mathrm{i}t_j\Xi_{1,j}\right)\right] - \prod_{j=0}^{k-1}E\left[\exp(\mathrm{i}t_j\Xi_{1,j})\right]\right| \tag{67}$$

is zero, see for instance Applebaum (2009, page 18). Applying the Volkonskii-Rozanov inequality (see Fan and Yao (2003, page 72)), it can be shown that (67) is bounded above by $C(n/a_n)\cdot\alpha(b_n - r + 1) \to 0$, implying asymptotic independence.

Moving on to the small blocks, due to stationarity we have

$$\begin{aligned}
\operatorname{Var}(\mathcal{S}) &= \frac{1}{n}\operatorname{Var}\left(\sum_{j=0}^{k-1}\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R_{nt}^{*}\right) \\
&= \frac{1}{n}\sum_{j=0}^{k-1}\operatorname{Var}\left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R_{nt}^{*}\right) + \frac{1}{n}\sum^{k-1}\sum_{j\neq l}\operatorname{Cov}\left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R_{nt}^{*}, \sum_{s=l(a+b)+a+1}^{(l+1)(a+b)} R_{ns}^{*}\right) \\
&= \frac{1}{n}\sum_{j=0}^{k-1}\left(b_n \operatorname{Var}(R_{nt}^{*}) + \sum_{t\neq l}^{b_n}\operatorname{Cov}(R_{nt}^{*}, R_{nl}^{*})\right) + \frac{1}{n}\sum^{k-1}\sum_{j\neq l}\sum_{i,j=1}^{b_n}\operatorname{Cov}\big(R_{n,i+w_j}^{*}, R_{n,r+w_l}^{*}\big) \\
&= Q_1 + Q_2 + Q_3.
\end{aligned}$$

where $w_j = j(a+b)+a$.

Regarding the first term, similar arguments used in deriving (64) yield

$$Q_1 = \frac{1}{n}k_n b_n \frac{\left[\varphi_x(\underline{h}\lambda)^{1/2}\right]^2 \sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2} = \frac{k_n b_n \sigma^2(x)\xi_2}{n\xi_1^2} \longrightarrow 0 \tag{68}$$

because $k_n b_n/n \sim b_n/(a_n+b_n) \to 0$. Now moving on to Q_2 and Q_3, the sum of covariances can be dealt with in the same manner as we did for the variance using (64), so $Q_2 \to 0$. Similarly for Q_3, implying $\operatorname{Var}(\mathcal{S}) \to 0$ as desired. Convergence result for the remainder $\mathcal{R}$ can be established similarly, and is bounded by $C(a_n+b_n)/n \to 0$.

The results above suggest that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} R_{nt}^{*} = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt}^{*}\right) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j + o_p(1), \tag{69}$$

and the desired result holds in view of (62) and the CLT for triangular array upon checking the Lindeberg condition (which is omitted here due to its similarity with Masry (2005, page 174-175)). Corollary 2 now follows because

$$\begin{aligned}
\sqrt{n\phi(h)}\left(\frac{\widehat{m} - m - \mathcal{B}_n}{\sqrt{n\phi(h)}\Delta_n}\right) &= \frac{\sqrt{n}\frac{1}{n}\sum_{t=1}^{n} R_{nt}^{*}}{\sqrt{\frac{1}{n}\sum_t \widehat{R}_{nt}^{*,2}}} = \frac{\frac{1}{\sqrt{n}}\sum_{t=1}^{n} R_{nt}^{*}}{\sqrt{\frac{1}{n}\sum_t R_{nt}^{*,2} + o_p(1)}} \\
= \frac{\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt}^{*} + o_p(1)}{\sqrt{\frac{1}{n}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R_{nt}^{*}\right)^2 + o_p(1)}} &= \frac{\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j + o_p(1)}{\sqrt{\frac{1}{n}\sum_{j=0}^{k-1}\eta_j^2 + o_p(1)}} \Longrightarrow N(0,1)
\end{aligned} \tag{70}$$

by Theorem 4.1 of de la Peña et al. (2009), since the denominator converges in probability to a strictly positive quantity $(\sigma^2(x)\xi_2/\xi_1^{-2})$, and that η_j belongs to the domain of attraction of a normal distribution by definition and (69). ■

5.3 Proof of Lemma 1 and 2

Lemma 1 is a straightforward extension of Lemma 4.3 and 4.4 of Ferraty and Vieu (2006), and hence is omitted. Lemma 2 can be shown by noting that for each n the τ_n-dimensional polyhedron $D := \{w = (w_i)_{i\leq\tau} \in \mathbb{R}^{\tau}, |w_i| \leq \lambda\}$ can be covered by $([2\lambda\sqrt{\tau}/\varepsilon+1])^{\tau}$ number of balls of radius ε, see Chaté and Courbage (1997), and then following the proof steps of Theorem 2 in Jia et al. (2003). ■

5.4 Proof of Theorem 4

As before, we start from the decomposition (19):

$$\widehat{m}(x) - m(x) = \frac{1}{\widehat{m}_1(x)} \left(\left[\widehat{m}_2(x) - E\widehat{m}_2(x)\right] + \left[E\widehat{m}_2(x) - m(x)\right] - m(x)\left[\widehat{m}_1(x) - 1\right] \right).$$

We recall from (65) that $\varphi_x(\lambda \underline{h}) \sim \varphi(\lambda \underline{h})$ and that the small deviation for the truncated regressor $X = (X_1, \ldots, X_\tau, 0, 0, \ldots)$ denoted $\varphi^{\mathcal{T}}(\lambda \underline{h})$ satisfies

$$\varphi(\lambda \underline{h}) = P\left(\sum_{j=1}^{\infty} j^{-2p} X_j^2 \leq h^2 \right) \leq P\left(\sum_{j=1}^{\tau} j^{-2p} X_j^2 \leq h^2 \right) = \varphi^{\mathcal{T}}(\lambda \underline{h}). \tag{71}$$

In the first step of the proof we show

$$\sup_{x \in \mathcal{S}_\tau} \left| \widehat{m}_2(x) - E\widehat{m}_2(x) \right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda \underline{h})}} \right). \tag{72}$$

We cover the set $\mathcal{S}_\tau$ defined in (39) with $L = L(\mathcal{S}_\tau, \eta)$ number of balls of radius η denoted by I_k, each of which is centred at x_k, $k = 1, \ldots, L$. i.e. $S_\tau \subset \bigcup_{k=1}^{L_n} B(x_{k,n}, \eta)$. Then it follows that

$$\begin{aligned}
\sup_{x \in \mathcal{S}_\tau} \left| \widehat{m}_2(x) - E\widehat{m}_2(x) \right| &= \max_{1 \leq k \leq L_n} \sup_{x \in I_k \cap \mathcal{S}_\tau} \left| \widehat{m}_2(x) - E\widehat{m}_2(x) \right| \\
&= \max_{1 \leq k \leq L_n} \sup_{x \in I_k \cap \mathcal{S}_\tau} \left| \widehat{m}_2(x) - \widehat{m}_2(x_k) + \widehat{m}_2(x_k) - E\widehat{m}_2(x_k) + E\widehat{m}_2(x_k) - E\widehat{m}_2(x) \right| \\
&\leq \max_{1 \leq k \leq L_n} \sup_{x \in I_k \cap \mathcal{S}_\tau} \left| \widehat{m}_2(x) - \widehat{m}_2(x_k) \right| + \max_{1 \leq k \leq L_n} \sup_{x \in I_k \cap \mathcal{S}_\tau} \left| E\widehat{m}_2(x_k) - E\widehat{m}_2(x) \right| \\
&\quad + \max_{1 \leq k \leq L_n} \left| \widehat{m}_2(x_k) - E\widehat{m}_2(x_k) \right| \quad =: R_1 + R_2 + R_3,
\end{aligned} \tag{73}$$

where $\widehat{m}_2(x_k) = (EK_1)^{-1} \sum_{t=1}^{n} Y_t K_{t,k}$ and $K_{t,k} = K(\|H^{-1}(x_k - X_t)\|)$.

We first consider R_1:

$$\begin{aligned}
R_1 &= \max_{1 \leq k \leq L_n} \sup_{x \in I_k \cap S_\tau} \left| \widehat{m}_2(x) - \widehat{m}_2(x_k) \right| \\
&= \max_{1 \leq k \leq L_n} \sup_{x \in I_k \cap S_\tau} \left| \frac{1}{nEK_1} \sum_{t=1}^{n} Y_t K\left(\|H^{-1}(x - X_t)\|\right) - Y_t K\left(\|H^{-1}(x_k - X_t)\|\right) \right| \\
&\leq \max_{1 \leq k \leq L_n} \sup_{x \in I_k \cap S_\tau} \frac{C}{n\varphi(\lambda \underline{h})} \sum_{t=1}^{n} \left| Y_t K_t - Y_t K_{t,k} \right|.
\end{aligned}$$

Now, because the kernel function is assumed to be Lipschitz continuous by Assumption E2, it follows that

$$R_1 \leq \frac{1}{n} \sum_{t=1}^{n} \frac{C'|Y_t|}{\varphi(\underline{h}\lambda)} \eta h^{-1} =: \frac{1}{n\varphi(\underline{h}\lambda)} \sum_{t=1}^{n} J_t,$$

where J_t is α-mixing under both assumptions A1′ and A2′. Then for some $\delta > 0$, on choosing $\eta = \log n/n$ and by Assumption B4′ we see that the tail condition (which is required for the exponential inequality to be applied below) continues to hold for J_t.

Also, using Assumption B6 we see that

$$E|J_t| \leq \frac{E(E(|Y_t||X))\eta}{h} \leq \frac{C\eta}{h}. \tag{74}$$

By Lemma 2 we can specify the Kolmogorov's entropy for S_τ with $\eta = \log n/n^2$:

$$\log L\Big(S, \frac{\log n}{n^2}\Big) = C \log \left[\left(\frac{2\lambda n^2}{\sqrt{\log n}} + 1 \right)^{\log n} \right] \sim \log n \times \log \left[\frac{2\lambda n^2}{\sqrt{\log n}} \right]$$

for sufficiently large n and λ, implying that the order of Kolmogorov's $\frac{\log n}{n^2}$ entropy is

$$O\Big(\log L\Big(S_\tau, \frac{\log n}{n^2}\Big)\Big) = O\Big((\log n)^2 - \log n[\log\log n] \Big) = O\Big(\big(\log n\big)^2 \Big). \tag{75}$$

Now, we apply the Fuk-Nagaev inequality for exponentially mixing variables of Merlevède, Peligrad and Rio (2009, 1.7) with $\varepsilon = \varepsilon_0[\log L\big(S, \frac{\log n}{n^2}\big)/(n\varphi(\lambda\underline{h}))]^{1/2}$ and $r = (\log n)^2$ for some positive constant ε_0. Since

$$s_n^2 := \sum_{t=1}^{n}\sum_{s=1}^{n} \mathrm{Cov}\,(J_t, J_s) \leq C\left(\frac{(\log n)^2}{n^2h^2}\right) = O\big(n\varphi(\lambda\underline{h})\log n\big),$$

and due to exponential mixing assumption it follows that

$$\begin{aligned}
&P\left(\max_{1\leq k\leq L_n} \sup_{x\in I_k\cap S_\tau} \Big|\widehat{m}_2(x) - \widehat{m}_2(x_k)\Big| > \varepsilon_0 \sqrt{\frac{\log L\big(S, \frac{\log n}{n^2}\big)}{n\varphi(\lambda\underline{h})}} \right) \\
&\leq\ 4\left(1 + \frac{n^2\varphi(\lambda\underline{h})^2\varepsilon_0^2 \log L\big(S, \frac{\log n}{n^2}\big)}{16(\log n)^2 s_n^2 n\varphi(\lambda\underline{h})}\right)^{-\frac{(\log n)^2}{2}} + \frac{16Cn}{\sqrt{n\varphi(\lambda\underline{h})}\log n} e^{-\varsigma\left\{\frac{\sqrt{n\varphi(\lambda\underline{h})}}{(\log n)}\right\}^{\gamma}} \\
&=\ 4\left(1 + \frac{n\varphi(\lambda\underline{h})\varepsilon_0^2}{16s_n^2}\right)^{-\frac{(\log n)^2}{2}} + \frac{16C\sqrt{n}}{\sqrt{\varphi(\lambda\underline{h})}\log n} \exp\big(-\varsigma\log L_n\big) \\
&\leq\ 4\exp\left(-\frac{\varepsilon_0^2(\log n)^2 n\varphi(\lambda\underline{h})}{32s_n^2}\right) + \left(\frac{Cn^2}{\sqrt{\log n}}\right)L_n^{-\varsigma} \\
&\leq\ 4\exp\left(-\frac{\varepsilon_0^2\log n}{32}\right) + \left(\frac{Cn^2}{\sqrt{\log n}}\right)\left(\frac{\sqrt{\log n}}{n^2}\right)^{\varsigma\log n} \longrightarrow 0
\end{aligned} \tag{76}$$

by the Taylor expansion of $\log(1+\epsilon)$ for sufficiently small $\epsilon > 0$.

Hence by (74) and Assumption E1 it follows that

$$\begin{aligned}
R_1 &= \max_{1\leq k\leq L_n} \sup_{x\in I_k\cap\mathcal{S}_\tau} \Big|\widehat{m}_2(x) - \widehat{m}_2(x_k)\Big| \leq O\Big(\frac{\eta}{h}\Big) + O_P\left(\sqrt{\frac{\log L\big(S, \frac{\log n}{n^2}\big)}{n\varphi(\lambda\underline{h})}}\right) \\
&= O\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right) + O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right) = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right).
\end{aligned} \tag{77}$$

As for the second term R_2, we have

$$R_2 \leq \max_{1\leq k\leq L_n} \sup_{x\in I_k\cap\mathcal{S}_\tau} E\big|\widehat{m}_2(x) - \widehat{m}_2(x_k)\big| = O\Big(\frac{\eta}{h}\Big) = O\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \tag{78}$$

Next we move on to the last component:

$$R_3 = \max_{1\leq k\leq L_n} \big|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)\big| =: \max_{1\leq k\leq L_n} \big|W_n(x_k)\big| \tag{79}$$

where

$$\begin{aligned} W_n(x) = \widehat{m}_2(x) - E\widehat{m}_2(x) &= \frac{1}{nEK_1}\sum_{t=1}^{n}\Big[Y_tK_t - EY_tK_t\Big] \\ &\leq \frac{C}{n\varphi_x^{\mathcal{T}}(\underline{h}\lambda)}\sum_{t=1}^{n}\Big[Y_tK_t - EY_tK_t\Big] = \frac{C}{n\varphi_x(\underline{h}\lambda)}\sum_{t=1}^{n}U_{nt}. \end{aligned}$$

where $U_{nt} = Y_tK_t - EY_tK_t$, and by elementary arguments

$$P\left(\max_{1\leq k\leq L_n} |\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)| > \varepsilon\right) \leq L_n \cdot \sup_{x\in\mathcal{S}} P\big(|W_n(x)| > \varepsilon\big). \tag{80}$$

Due to the dependence of Q_{nt} on X_t we consider the cases of static and dynamic regressions separately, because the asymptotic arguments to follow depends upon the temporal dependence structure of Q_{nt}.

In the static case, we first examine the situation where the response is unbounded and satisfies the exponential tail condition in B4$'$. Since

$$s_n^2 = \sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{Cov}\,(U_{nt}, U_{ns}) = O\left(n\varphi_x^{\mathcal{T}}(\underline{h}\lambda)\right),$$

we apply the Fuk-Nagaev inequality for exponentially mixing variables once again. Writing $L_n := L\left(S, \frac{\log n}{n^2}\right)$ and taking $\varepsilon = \varepsilon_0[\log L\big(S, \frac{\log n}{n^2}\big)/(n\varphi(\lambda\underline{h}))]^{1/2}$ and $r = (\log n)^{2+\epsilon}$, $\epsilon \in (0, 1/2)$ for some $\varepsilon_0 > 0$, we have

$$\begin{aligned} &P\left(\Big|\widehat{m}_2(x) - E\widehat{m}_2(x)\Big| > \varepsilon_0\sqrt{\frac{\log L_n}{n\varphi(\lambda\underline{h})}}\right) \leq P\left(\left|\sum_{t=1}^{n}U_{nt}\right| > n\varphi_x^{\mathcal{T}}(\underline{h}\lambda)\varepsilon_0\sqrt{\frac{\log L_n}{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}\right) \\ &\leq 4\left(1 + \frac{n\varphi^{\mathcal{T}}(\lambda\underline{h})\varepsilon_0^2\log L_n}{16(\log n)^{2+\epsilon}s_n^2}\right)^{-\frac{(\log n)^{2+\epsilon}}{2}} + \frac{16Cn}{\sqrt{n\varphi^{\mathcal{T}}(\lambda\underline{h})}\log n}\exp\left(-\varsigma\left\{\frac{\sqrt{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}{(\log n)^{1+\epsilon}}\right\}^{\gamma}\right) \\ &\leq 4\left(1 + \frac{\varepsilon_0^2\log L_n}{16(\log n)^{2+\epsilon}}\right)^{-\frac{(\log n)^{2+\epsilon}}{2}} + \frac{16C\sqrt{n}}{\sqrt{\varphi^{\mathcal{T}}(\lambda\underline{h})}\log n}\exp\big(-\varsigma\log L_n\big) \\ &\leq 4\exp\left(-\frac{\varepsilon_0^2\log L_n}{32}\right) + \left(\frac{Cn^2}{\sqrt{\log n}}\right)L_n^{-\varsigma} \leq 4L_n^{-\frac{\varepsilon_0^2}{32}} + \left(\frac{Cn^2}{\sqrt{\log n}}\right)\left(\frac{\sqrt{\log n}}{n^2}\right)^{\varsigma\log n} \to 0 \end{aligned}$$

because $\gamma \geq 1$ and $L_n = O((n^2/\sqrt{\log n})^{\log n})$.

Now since $\varsigma > 1$, by choosing ε_0 large enough it follows by (80) that

$$R_3 = \max_{1\le k\le L_n} \left|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)\right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \tag{81}$$

In the special case when the response is bounded, the same result continues to hold with $\gamma_1 = \infty$ (so that $\gamma_2 = \gamma(\ge 1)$).

An alternative proof for the case of bounded response could be done by applying the exponential inequality of Bosq (1996, Theorem 1.3.2) for α-mixing sequences as follows: Noting that $|Q_t| \le C/\varphi_x(\underline{h}\lambda) =: b$, $\forall t$, and that $\sigma^2(r) := p \cdot \mathrm{Var}(Q_t) = O(p/\varphi(\underline{h}\lambda))$ (where $p = n/(2q)$ and $q = \log n\sqrt{n}/\sqrt{\varphi}$) by the Cauchy-Schwarz inequality and Assumption B4 we have

$$v^2(r) = \frac{2}{p^2}\sigma^2(r) + \frac{b\varepsilon}{2} \le \frac{Cq}{n\varphi_x(\underline{h}\lambda)} + \frac{C\varepsilon}{\varphi_x(\underline{h}\lambda)} \le \frac{C'\varepsilon}{\varphi_x(\underline{h}\lambda)},$$

where $\varepsilon = \varepsilon_0\sqrt{\log L_n/(n\varphi)}$ and $L_n := L\left(S, \frac{\log n}{n^2}\right)$, and by Assumption A1 that

$$\begin{aligned}
P\left(\left|\widehat{m}_2(x) - E\widehat{m}_2(x)\right| > \varepsilon_0\sqrt{\frac{\log L_n}{n\varphi(\lambda\underline{h})}}\right) &\le 4e^{-\varepsilon^2 q/(8v^2(r))} + 22\sqrt{1 + \frac{4b}{\varepsilon}}q\alpha\left(\left[\frac{n}{2q}\right]\right) \\
&\le 4\exp\left\{-\frac{\varepsilon_0 q\varphi\sqrt{\log L_n}}{8\sqrt{n\varphi}}\right\} + 22\left(1 + \frac{4\sqrt{n\varphi}}{\varphi\log n}\right)^{1/2}\frac{\log n\sqrt{n}}{\sqrt{\varphi}}\alpha\left(\left[\frac{\sqrt{n\varphi}}{2\log n}\right]\right) \\
&\le 4\exp\left\{-\frac{\varepsilon_0\log L_n}{8}\right\} + \exp\left(-\varsigma\left\{\frac{\sqrt{n\varphi(\lambda\underline{h})}}{\log n}\right\}^{\gamma_2}\right) \\
&\le 4L_n^{-\varepsilon_0/8} + \frac{C(\log n)^{1/2}n^{3/4}}{\varphi(\lambda\underline{h})L_n^{\varsigma}} \to 0.
\end{aligned}$$

In the dynamic regression case (i.e. under C2), the same conclusion can be derived by starting from (79) and exploiting the mixing approximation argument:

$$\begin{aligned}
\max_{1\le k\le L_n}\left|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)\right| &= \max_{1\le k\le L_n}\left|W_n(x_k)\right| = \max_{1\le k\le L_n}\left|n^{-1}\sum_{t=1}^n Q_{nt,k}\right| \\
&\le \max_{1\le k\le L_n}\left|\frac{1}{n}\sum_{t=1}^n Q_{nt,k,(r)}\right| + \sup_{x\in\mathcal{S}}\frac{1}{n}\sum_{t=1}^n\left|Q_{nt,(r)} - Q_{nt}\right| \\
&= O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right) + O_P\left(\frac{\sqrt{v_2(r)}}{\varphi(\lambda\underline{h})}\right) = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right),
\end{aligned}$$

since $\sqrt{n}\sqrt{v_2(r)}(\log n)^{-1}/\sqrt{\varphi} \le \sqrt{n}\sqrt{v_2(r)}/\varphi \to 0$ by (25).

Now returning back to where we started, viewing $\widehat{m}_1(x)$ as a special case of $\widehat{m}_2(x)$ with $Y_t = 1$ $\forall t$, we can repeat the above procedure, yielding (since $E\widehat{m}_1(x) = 1$)

$$\sup_{x\in\mathcal{S}_\tau}\left|\widehat{m}_1(x) - 1\right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \tag{82}$$

The proof is now complete in view of (71), (72), (77), (78), (81), (82), contributions from the bias component, Proposition 4.1 of Dunker, Lifshits and Linde (1998), and Theorem 1.1 of Hong, Lifshits and Nazarov (2016). ■

References

[1] de Acosta, A., 1983. *Small deviations in the functional central limit theorem with applications to functional laws of the iterated logarithm.* Annals of Probability, 11(1), 78-101.

[2] Andrews, D.W.K. (1984). Non-Strong Mixing Autoregressive Processes. Journal of Applied Probability 21(4) · December 1984

[3] Andrews, D. W.K., 1995. *Nonparametric kernel estimation for semiparametric models.* Econometric Theory, 11(3), 560-586.

[4] Applebaum, D., 2009. *Lévy Processes and Stochastic Calculus.* Cambridge University Press.

[5] Azais, J. M., Fort, J. C. (2013). *Remark on the finite-dimensional character of certain results of functional statistics.* Comptes Rendus Mathematique, 351(3), 139-141.

[6] Bierens, H. J., 1983. *Uniform consistency of kernel estimators of a regression function under generalized conditions.* Journal of the American Statistical Association, 78(383), 699-707.

[7] Bierens, H. J., 1987. *Kernel estimators of regression functions.* In Advances in econometrics: Fifth world congress, 1, 99-144.

[8] Bingham, N. H., Goldie, C. M., Teugels, J. L., 1987. *Regular Variation.* Cambridge University Press.

[9] Billingsley, P., 1968. *Convergence of Probability Measures.* Wiley.

[10] Borovkov, A. A., Ruzankin, P. S., 2008. *On small deviations of series of weighted random variables.* Journal of Theoretical Probability, 21(3), 628-649.

[11] Bosq, B., 1996. *Nonparametric statistics for stochastic processes: estimation and prediction.* Springer-Verlag, New York.

[12] Bradley, R. C., 2005. *Basic properties of strong mixing conditions. A survey and some open questions.* Probability Surveys, 2(2), 107-144.

[13] Bulinskii, A. V., Shiryaev, A. N., 2003. *Theory of Random Processes* (in Russian). Fizmatlit.

[14] Chaté, H., Courbage, M. (ed.), 1997. *Special issue on lattice dynamics.* Physica D, 103, 1-612.

[15] Chen, R., Tsay, R. S., 1993. *Nonlinear additive ARX models.* Journal of the American Statistical Association, 88(423), 955-967.

[16] Davidson, J., 1994. *Stochastic Limit Theory: An Introduction for Econometricians.* Oxford University Press.

[17] Davydov, Y. A., 1968. *Convergence of distributions generated by stationary stochastic processes.* Theory of Probability and Applications, 13(4), 691-696.

[18] Delsol, L., 2009. *Advances on asymptotic normality in non-parametric functional time series analysis.* Statistics, 43(1), 13-33.

[19] Devroye, L. P., 1978. *The uniform convergence of the Nadaraya-Watson regression function estimate.* Canadian Journal of Statistics, 6(2), 179-191.

[20] Devroye, L., 1981. *On the almost everywhere convergence of nonparametric regression function estimates.* Annals of Statistics, 9(6), 1310-1319.

[21] Doukhan, P., 1994. *Mixing.* Springer, New York.

[22] Doukhan, P., Wintenberger, O., 2008. *Weakly dependent chains with infinite memory.* Stochastic Processes and their Applications, 118(11), 1997-2013.

[23] Duflo, M., 1997. *Random iterative models.* Springer-Verlag, Berlin.

[24] Dunker, T., Lifshits, M. A., Linde, W., 1998. *Small deviation probabilities of sums of independent random variables.* In: Eberlein, E. (ed.), High dimensional probability, volume 43 of Progress in Probability., Birkhauser, Basel, 59-74.

[25] Fan, J., Masry, E., 1992. *Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes.* Journal of Multivariate Analysis, 43(2), 237-271.

[26] Fan, J., 1990. *A remedy to regression estimators and nonparametric minimax efficiency.* Department of Statistics, University of North Carolina at Chapel Hill.

[27] Fan, J., Yao, Q., 2005. *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

[28] Feller, W., 1971. *Introduction to Probability Theory and Its Applications*, Vol. 2. Wiley, New York.

[29] Ferraty, F., Laksaci, A., Tadj, A., Vieu, P., 2010. *Rate of uniform consistency for nonparametric estimates with functional variables.* Journal of Statistical Planning and Inference, 140, 335-352.

[30] Ferraty, F., Romain, Y., 2010. *The Oxford Handbook of Functional Data Analysis.* Oxford University Press.

[31] Ferraty, F., Vieu, P., 2002. *The functional nonparametric model and application to spectrometric data.* Computational Statistics, 17(4), 545-564.

[32] Ferraty, F., Vieu, P., 2006. *Nonparametric functional data analysis: Theory and Practice.* Springer, New York.

[33] Fuk, D. K., Nagaev, S. V., 1971. *Probability inequalities for sums of independent random variables.* Theory of Probability and its applications, 16(4), 643-660.

[34] Gao, F., Hannig, J., Torcaso, F., 2003. *Comparison Theorems for Small Deviations of Random Series.* Electronic Journal of Probability, 8(21), 1-17.

[35] Geenens, G., 2011. *Curse of dimensionality and related issues in nonparametric functional regression.* Statistics Surveys, 5, 30-43.

[36] Götze, F., Hipp, C., 1994. *Asymptotic distribution of statistics in time series.* Annals of Statistics, 22(4), 2062-2088.

[37] Greblicki, W., Krzyzak, A., 1980. *Asymptotic properties of kernel estimates of a regression function.* Journal of Statistical Planning and Inference, 4(1), 81-90.

[38] Hädle, W. K., 1990. *Applied Nonparametric Regression.* (Vol. 27). Cambridge University Press.

[39] Hong, S. Y., Lifshits, M., Nazarov. A., 2016. *Small deviations in L_2-norm for Gaussian dependent sequences.* Electronic Communications in Probability, 21, 1-9.

[40] Ibragimov, I. A., 1962. *Some limit theorems for stationary processes.* Theory of Probability and its Applications, 23, 291-300.

[41] Ibragimov, I. A., Linnik., Y. B., 1971. *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen.

[42] Jia, Q., Zhou, S., Yin. F., 2003. *Kolmogorov entropy of global attractor for dissipative lattice dynamical systems.* Journal of Mathematical Physics, 44, 5804-5801.

[43] Karamata, J., 1933. *Sur un mode de croissance regulire. Theoremes fondamentaux* (in French). Bulletin de la Societe Mathematique de France, 61, 55-62.

[44] Lewandowski, M., Ryznar, M., 1995. *Anderson inequality is strict for Gaussian and stable measures.* Proceedings of the American Mathematical Society, 123(12), 3875-3880.

[45] Li, W. V., 2012. *Small value probabilities in analysis and mathematical physics.* Presented at the Arizona School of Analysis and Mathematical Physics, Tucson, Arizona, United States in March 15, 2012. Retrieved from the link `http://math.arizona.edu/~mathphys/school_2012/WenboLi.pdf`.

[46] Li, W. V., Shao, Q. M., 2001. *Gaussian processes: inequalities, small ball probabilities and applications.* Handbook of Statistics, 19, 533-598.

[47] Linton, O. B., Sancetta, A., 2009. *Consistent estimation of a general nonparametric regression function in time series.* Journal of Econometrics, 152(1), 70-78.

[48] Lu, Z., 2001. *Asymptotic normality of kernel density estimators under dependence.* Annals of the Institute of Statistical Mathematics, 53(3), 447-468.

[49] Mas, A., 2012. *Lower bound in regression for functional data by small ball probability representation in Hilbert space*, Electronic Journal of Statistics, 6, 1745-1778.

[50] Masry, E., 2005. *Nonparametric regression estimation for dependent functional data: asymptotic normality.* Stochastic Processes and their Applications, 115(1), 155-177.

[51] Masry, E., Fan, J., 1997. *Local polynomial estimation of regression function for mixing processes.* Scandinavian Journal of Statistics, 24, 1965-1979.

[52] Merlevède, F., Peligrad, M., Rio, E., 2009. *A Bernstein type inequality and moderate deviations for weakly dependent sequences.* Preprint available at `http://arxiv.org/abs/0902.0582`.

[53] Nadaraya, E. A., 1964. *On estimating regression.* Theory of Probability & Its Applications, 9(1), 141-142.

[54] Nadaraya, E. A., 1970. *Remarks on non-parametric estimates for density functions and regression curves.* Theory of Probability & Its Applications, 15(1), 134-137.

[55] Olver, F. W. J., Lozier, D. W., Boisvert, R. F., Clark, C. W., 2010. *NIST Handbook of Mathematical Functions.* Cambridge University Press.

[56] Pagan, A., Ullah, A., 1988. *The econometric analysis of models with risk terms.* Journal of Applied Econometrics, 3(2), 87-105.

[57] Parzen, E., 1962. *On estimation of a probability density function and mode.* Annals of Mathematical Statistics, 33(3), 1065-1076.

[58] de la Peña, V. H., Lai, T. L., Shao, Q. M. 2008. *Self-normalized processes: Limit theory and Statistical Applications.* Springer, New York.

[59] Parzen, E. (1962). *On estimation of a probability density function and mode.* Annals of Mathematical Statistics, 33(3), 1065-1076.

[60] Phillips, P. C., Park, J. Y., 1998. *Nonstationary density estimation and kernel autoregression.* Cowles Foundation for Research in Economics.

[61] Ramsay, J. O., Silverman, B. W., 2002. *Applied functional data analysis: methods and case studies.* Springer, New York.

[62] Rio, E., 2000. *Theorie asymptotique des processus alé atoires faiblement dependants* (in French). Springer, Maththematiques and Applications, 31.

[63] Robinson, P. M., 1983. *Nonparametric estimators for time series.* Journal of Time Series Analysis, 4(3), 185-207.

[64] Rosenblatt, M., 1956. *A central limit theorem and a strong mixing condition.* Proceedings of the National Academy of Sciences, 42(1), 43-47.

[65] Roussas, G. G., 1989. *Consistent regression estimation with fixed design points under dependence conditions.* Statistics & Probability Letters, 8(1), 41-50.

[66] Roussas, G. G., 1990. *Nonparametric regression estimation under mixing conditions.* Stochastic Processes and Their Applications, 36(1), 107-116.

[67] Schuster, E. F., 1972. *Joint asymptotic distribution of the estimated regression function at a finite number of distinct points.* Annals of Mathematical Statistics, 43(1), 84-88.

[68] Skorohod, A. B., 1967. *On the densities of probability measures in functional spaces.* Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66) Vol. II: Contributions to Probability Theory, 1, 163-182. University of California Press, Berkely, CA.

[69] Stone, C., 1980. *Optimal rates of convergence for nonparametric estimators.* Annals of Statistics, 8(6), 1348-1360.

[70] Stone, C., 1982. *Optimal global rates of convergence for nonparametric regression.* Annals of Statistics, 10(4), 1040-1053.

[71] Sytaya, G. N., 1974. *Certain asymptotic representations for a Gaussian measure in Hilbert space* (in Russian). Theory of Random Process, 2, 93-104.

[72] Watson, G. S., 1964. *Smooth regression analysis*, Sankhya Ser. A, 26(4), 359-372.

[73] Wu, W. B., 2011. *Asymptotic theory for stationary processes.* Statistics and Its Interface, 4, 207-226.

[74] Zolotarev, V. M., 1986. *Asymptotic behavior of the Gaussian measure in* ℓ_2. Journal of Soviet Mathematics, 35(2), 2330-2334.