Electronic Journal of Statistics
Vol. 6 (2012) 1373–1391
ISSN: 1935-7524
DOI: 10.1214/12-EJS716

Convergence of nonparametric functional regression estimates with functional responses

Heng Lian

Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
Singapore 637371
e-mail: henglian@ntu.edu.sg

Abstract: We consider nonparametric functional regression when both predictors and responses are functions. More specifically, we let $(X_1, Y_1), \dots, (X_n, Y_n)$ be random elements in $\mathcal{F} \times \mathcal{H}$ where $\mathcal{F}$ is a semi-metric space and $\mathcal{H}$ is a separable Hilbert space. Based on a recently introduced notion of weak dependence for functional data, we showed the almost sure convergence rates of both the Nadaraya-Watson estimator and the nearest neighbor estimator, in a unified manner. Several factors, including functional nature of the responses, the assumptions on the functional variables using the Orlicz norm and the desired generality on weakly dependent data, make the theoretical investigations more challenging and interesting.

AMS 2000 subject classifications: Primary 62G08, 60G10.
Keywords and phrases: Bernstein's inequality for martingale differences, Nadaraya-Watson estimate, nearest neighbor estimate, nonparametric functional regression, Orlicz norm.

Received March 2012.

1. Introduction

The problem of regression with functional predictors has been receiving increasing interests nowadays, boosted by more and more datasets with observations that can be naturally perceived as curves. This trend starts with the popular monograph Ramsay and Silverman (2002) that gives a detailed exposition of functional linear models. The existing literature contains numerous theoretical and empirical studies on functional linear models (Cardot, Ferraty and Sarda, 1999; Cuevas, Febrero and Fraiman, 2002; James, 2002; Müller and Stadtmüller, 2005; Yao, Müller and Wang, 2005; Cai and Hall, 2006; Hall and Horowitz, 2007; Crambes, Kneip and Sarda, 2009; Febrero-Bande, Galeano and Gonzalez-Manteiga, 2010). Nonparametric methods with functional predictors and scalar responses appear later (Ferraty and Vieu, 2002, 2004, 2006; Preda, 2007; Biau, Cerou and Guyader, 2010), which by now have been widely accepted by the statistical community as a more flexible approach to functional regression with fewer structural assumptions imposed. As this area naturally develops and matures, the situation where the responses are also curves begins to receive more attention (Aguilera, Ocana and Valderrama, 2008; Crambes and Mas, 2009; Horváth, Kokoszka and Reimherr, 2009). For example, one might predict annual precipitation using temperature measurements as in Ramsay and Silverman (2005), or predict future hourly electricity consumption based on past history as in Antoch et al. (2008). Although these two studies follow the parametric approach to functional regression, it is clear that nonparametric approach is a viable alternative (Lian, 2007).

On the other hand, the assumption of independence in most theoretical investigations carried out so far is often too restrictive in many applications. The necessity to respond properly to data dependence is clearly demonstrated by the example given in Ferraty, Goia and Vieu (2002) where a functional observation denotes the monthly electricity consumption over a year and thus it is unrealistic to assume that electricity consumption in one year is independent of that of the previous year. In previous studies regarding nonparametric functional regression, dependence is incorporated based on some mixing conditions (Ferraty and Vieu, 2004). Here we instead use the notion of $L^4 - m-$approximability advocated in Hörmann and Kokoszka (2010); Gabrys, Horváth and Kokoszka (2010) (with some appropriate minor extensions). The advantage compared to using mixing conditions is that the $L^4 - m-$approximability condition is easily verified in many examples as shown in Hörmann and Kokoszka (2010).

In the more classical setting, the observation pairs reside in the Euclidean spaces. In this paper, we carry out a theoretical investigation of nonparametric functional regression with functional responses on dependent data. Two related classes of nonparametric estimates have been proposed, the k-nearest neighbor estimate (k-NN) and the Nadaraya-Watson kernel estimate. Because of their similarity in many aspects, we will try to unify the proofs for these two as much as possible. We will show almost sure convergence of these nonparametric estimators based on assumptions on Orlicz norms of the functional variables. The condition involving the Orlicz norm is more general than the usual moment condition and thus it is of theoretical interest to generalize existing results in term of the Orlicz norm. Also, as seen in the results (for example Corollary 1) in Section 2, the choice of the neighborhood size in the k-NN estimator has an interesting interaction with the error assumption. With a stronger Orlicz norm (more restrictive assumption) on error, the required condition on the neighborhood size can be relaxed. Due to the functional nature of the responses and the assumption of weak dependence, the theoretical investigation poses serious challenges and some novel construction of martingale difference sequence will be introduced.

This work can be regarded as an extension of Lian (2011); Ferraty et al. (2011) which treated the k-NN and the Nadaraya-Watson estimators respectively on independent data with more restrictive error distribution assumptions. We provide a *unified* analysis of the two estimators. In Section 2.1, some background material on Orlicz norm and weak dependence is introduced. The main theoretical results are presented in Section 2.2 which apply to both the k-NN estimator and the Nadaraya-Watson estimator. Section 2.3 discusses more specifically how the general theorem can be applied to the two estimators. The technical proofs are contained in Section 3.

Finally, we note that throughout the paper we use C to denote a generic constant that assumes different values at different places.

2. Almost sure convergence of nonparametric estimates

2.1. On the notion of Orlicz norm and weak dependence

In this subsection we review the concept of Orlicz norm and collect some of its simple properties as a lemma here for easy reference later. Although all of the properties are simple and most are well-known, some others seem to be new (such as Lemma 1 (vi)(vii)) which we cannot find in the existing literature. We also review and extend the notion of L^4-m-approximability of a data sequence using the more general Orlicz norm instead of L^p norm.

Following van der Vaart and Wellner (1996), let ψ be a convex, increasing function on $[0,\infty)$ with $\psi(0)=0$ and let X be a real-valued random variable. The Orlicz norm (or ψ-Orlicz norm to emphasize its dependence on ψ) is defined as

$$\|X\|_\psi = \inf\{C>0 : E[\psi(|X|/C)] \le 1\},$$

which can be shown to be indeed a norm. For random elements X taking values in a normed space, the Orlicz norm of $\|X\|$ (which is a real-valued random variable) is also denoted by $\|X\|_\psi$ for simplicity.

There are two commonly used ψ functions: $\psi(x)=x^p$ and $\psi(x)=\exp\{x^p\}-1$, $p\ge 1$, and throughout the paper we use ψ_p to denote the latter. With $\psi(x)=x^p$, the Orlicz norm is simply the L^p norm $(E[|X|^p])^{1/p}$. With $\psi(x)=\psi_p(x)=\exp\{x^p\}-1$, the finiteness of Orlicz norm of X is closely related to the exponential decay of its tail probability, the exact statement of which is contained in the following Lemma together with other simple properties concerning the Orlicz norm.

Lemma 1. *Below we assume ψ is a valid function that defines an Orlicz norm, that is, ψ is convex, increasing on $[0,\infty)$ with $\psi(0)=0$. X is a random variable.*

(i) $P(|X|>x) \le 1/\psi(x/\|X\|_\psi), \forall x \ge 0$.
(ii) *If* $P(|X|>x) \le K\exp\{-Cx^p\}$ *for all* $x\ge 0$ *and some constants* K *and* C, *then* $\|X\|_{\psi_p} \le ((1+K)/C)^{1/p}$.
(iii) *If* $\tilde\psi(x)=\psi(ax)$ *for some* $a>0$, *then* $\|X\|_{\tilde\psi} = a\|X\|_\psi$.
(iv) *If* $\tilde\psi(x)\le a\psi(x)$ *for some* $a\ge 1$, *then* $\|X\|_{\tilde\psi} \le a\|X\|_\psi$.
(v) *If* $\tilde\psi(x)=\phi(\psi(ax))$ *for some* $a>0$ *and some concave increasing function* ϕ *with* $\phi(0)=0$ *and* $\phi(1)=1$, *then* $\|X\|_{\tilde\psi}\le a\|X\|_\psi$.
(vi) *If* $\tilde\psi(x):=\psi(x^{1/p}), p\ge 1$ *is convex, then* $\||X|^p\|_{\tilde\psi}\le \|X\|_\psi^p$.
(vii) $\|E[X|\mathcal{G}]\|_\psi \le \|X\|_\psi$, *for any* σ*-algebra* $\mathcal{G}$.

Proof. Results (i) and (ii) can be found in Section 2.2 of van der Vaart and Wellner (1996). (iii) is obvious by the definition of Orlicz norm. To prove (iv), we note that $E\tilde\psi(|X|/a\|X\|_\psi) \le aE\psi(|X|/a\|X\|_\psi) \le E\psi(|X|/\|X\|_\psi) \le 1$, where we used that $\psi(x/a)\le\psi(x)/a$ due to the convexity of ψ. For (v), since

$E\tilde{\psi}(|X|/a\|X\|_\psi) = E\phi(\psi(|X|/\|X\|_\psi)) \le \phi(E\psi(|X|/\|X\|_\psi)) \le \phi(1) = 1$ (using Jensen's inequality), we get $\|X\|_{\tilde{\psi}} \le a\|X\|_\psi$ by definition. For (vi), the result follows from $E\tilde{\psi}(|X|^p/\|X\|_\psi^p) = E\psi(|X|/\|X\|_\psi) \le 1$. Finally, (vii) follows from

$$\begin{aligned} E\psi(E[X|\mathcal{G}]/\|X\|_\psi) &= E\psi(E[X/\|X\|_\psi|\mathcal{G}]) \\ &\le E(E(\psi(X/\|X\|_\psi)|\mathcal{G})) \\ &= E\psi(X/\|X\|_\psi) \le 1, \end{aligned}$$

where we used $\psi(E[X/\|X\|_\psi|\mathcal{G}]) \le E[\psi(X/\|X\|_\psi)|\mathcal{G}]$ due to convexity of ψ. □

We already noted that L^p norm is a special case of Orlicz norm when $\psi(x) = x^p$. On the other hand, based on Lemma 1 (v), one can show that $\|X\|_p \le C\|X\|_{\psi_q}$ for any $p, q \ge 1$ and $\|X\|_{\psi_{q_1}} \le C'\|X\|_{\psi_{q_2}}$ if $q_1 \le q_2$, (where C, C' are universal constants that only depends on p, q, q_1, q_2). In this sense the norm $\|.\|_{\psi_q}$ is stronger than L^p, and the more so with larger q.

As explained in the introduction, for data collected sequentially over time, the assumption of independence is not realistic. In Hörmann and Kokoszka (2010), the authors formalize the notion of dependence for functional data using $L^4 - m-$approximability. Instead of using the L^4 norm which is sufficient for the purpose of those studies, we instead use the Orlicz norm here.

Definition 1. Given a function ψ that defines an Orlicz norm, a sequence $\{X_i\}_{i=1}^\infty$ (taking values in a normed space) with finite Orlicz norm is said to be $\psi - m-$approximable if we have the representation

$$X_i = h(\alpha_i, \alpha_{i-1}, \ldots),$$

where the α_k are independent and identically distributed random elements of a measurable space and h is a measurable function. In addition, we assume that if

$$X_i^{(m)} = h(\alpha_i, \alpha_{i-1}, \ldots, \alpha_{i-m+1}, \alpha'_{i-m}, \alpha'_{i-m-1} \ldots),$$

with α'_k independent copies of α_k, then

$$\sum_{m=1}^\infty \|X_m - X_m^{(m)}\|_\psi < \infty.$$

For a $\psi - m-$approximable sequence $\{X_i\}$, we say it is $\psi - m-$approximable with decay rate γ_k if $\sum_{m=k}^\infty \|X_m - X_m^{(m)}\|_\psi = O(\gamma_k)$.

In Hörmann and Kokoszka (2010), several examples of $L^p - m-$approximable sequence are given, minor modifications of these can produce more general $\psi - m-$approximable sequences. For example, a functional autoregressive process (Example 2.1 in Hörmann and Kokoszka (2010)) is $\psi - m-$approximable as long as the innovation noise has finite ψ-Orlicz norm, by the same arguments. Although not explicitly stated there, a functional autoregressive process is $\psi - m-$approximable with exponential decay rate: $\gamma_m = O(\exp\{-Cm\})$ for some constant C. This example can also be extended to the more general linear process as in Proposition 2.1 of Hörmann and Kokoszka (2010). Other examples there, such as functional bilinear process, and functional ARCH, could be adapted to obtain $\psi - m-$approximable processes.

2.2. Nonparametric estimates and convergence rate

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a stationary (in a strong sense) sequence of $\mathcal{F} \times \mathcal{H}$-valued random elements with $E\|Y\| < \infty$, where $\mathcal{F}$ is a semi-metric space with semi-metric $d(.,.)$ and $\mathcal{H}$ is a Hilbert space with norm $\|.\|$. Note that we say $d(.,.)$ is semi-metric if (i) $d(x,x) = 0$, $\forall x \in \mathcal{F}$; (ii) $d(x,y) = d(y,x)$, $\forall x, y \in \mathcal{F}$; (iii) $d(x,y) \leq d(x,z) + d(z,y)$, $\forall x, y, z \in \mathcal{F}$. This definition of semi-metric might be different from that used in some mathematics literature. However, since the works on nonparametric functional data analysis (Ferraty and Vieu, 2004, 2006) are so popular by now, we deem it appropriate to follow their definition of semi-metric.

The regression function is $r(t) = E(Y|X = t)$ and we can write $Y_i = r(X_i) + \epsilon_i$ where $\epsilon_i = Y_i - E(Y_i|X_i) \in \mathcal{H}$ are mean zero noises (in the sense of Bochner integral, see Ledoux and Talagrand (1991)). *In this subsection, we always consider probabilities and expectations conditional on* $\{X_i\}$*, in effect treating it as fixed.* The asymptotic results stated are thus conditional on predictors even though we do not state this explicitly in the following. The implications of random predictors are treated in the next subsection after we present the general convergence results in this subsection.

The regression function can be estimated by local weighting of responses

$$\hat{r}(t) = \sum_{i=1}^{n} W_i(t) Y_i, \tag{1}$$

where $(W_1(t), \ldots, W_n(t))$ is a probability vector of weights. Here the weights actually depend on n, but we make this implicit in our notations for simplicity (similarly for other quantities below such as b, k, H, v_i, etc.). Note that $W_i(t)$ can be a function of all $X_k, k = 1, \ldots, n$, instead of X_i only, as is the case for k-NN estimates (see the examples below). Since in this paper we only investigate pointwise convergence at a fixed point t, we will use the notation $(W_1, \ldots, W_n)$ in the following for simplicity.

We rank $(X_i, Y_i), i = 1, \ldots, n$, based on increasing value of $d(X_i, t)$ (ties are broken by indices) and obtain a vector $(R_1, \ldots, R_n)$ such that X_{R_i} is the ith nearest neighbor of t. Let $v_i = W_{R_i}$, we can write (1) equivalently as

$$\hat{r}(t) = \sum_{i=1}^{n} v_i Y_{R_i}. \tag{2}$$

Our consideration of weak dependence leads to extra complications in the proofs. If the observations are independent, then obviously Y_{R_i} are also independent. However, if $(Y_1, Y_2, \ldots)$ is merely stationary, then $(Y_{R_1}, Y_{R_2}, \ldots)$ is no longer stationary in general since the order of observations is broken. We will thus use representation (1) in most parts of our proofs, although representation (2) is easier to manipulate in the study of k-NN estimates for independent data.

Example 1 (Simple nearest neighbor estimate)**.** Take $v_i = 1/k$ for $i \leq k$ and $v_i = 0$ for $i > k$, so that the regression function estimate is just the average of responses corresponding to the k nearest neighbors of t. Even in this simplest case, although v_i is only a deterministic sequence, W_i still depends on all $X_j, 1 \leq j \leq n$ since all predictors jointly determine t's neighbors. More generally, we can take v_i to be a deterministic sequence with $v_1 \geq v_2 \geq \cdots \geq v_n$ thus putting more weights on data closer to t.

Example 2 (Nearest neighbor estimate based on kernel)**.** Take

$$W_i = K(d(X_i, t)/H)/\sum_j K(d(X_j, t)/H),$$

where K is a kernel function and H is the distance of the kth nearest neighbor to t. Mathematically,

$$H = \inf\{h \in R : \sum_{i=1}^{n} I\{X_i \in B(t, h)\} \geq k\}, \tag{3}$$

where $B(t, h) = \{t' \in \mathcal{F} : d(t', t) \leq h\}$ and $I\{.\}$ denotes the indicator function. In this subsection, since we condition on predictors $\{X_i\}$, H is a known fixed value.

Example 3 (Nadaraya-Watson estimate)**.** Take

$$W_i = K(d(X_i, t)/H)/\sum_j K(d(X_j, t)/H),$$

which has exactly the same form as in the previous example. However, here H is a predetermined value usually called the bandwidth parameter, not derived from distance of t's kth nearest neighbor. Typically, one applies the same value of H for all values of t. Thus compared to nearest neighbor estimate, the Nadaraya-Watson estimate is not adaptive to the local sparseness of data. In this subsection when conditioning on predictors and for a given t, of course Nadaraya-Watson estimator is same as that in Example 2 since H is fixed in both cases. The differences will appear in the next subsection.

Naturally we need the following assumption on the regression function to obtain nontrivial rates of convergence.

Assumption 1. *r is bounded and Lipschitz continuous. That is $\|r(t)\| \leq B, \forall x \in \mathcal{F}$ and $\|r(t) - r(t')\| \leq Md(t, t')^{\alpha}$.*

In fact, since we only consider pointwise convergence, it suffices that r is Lipschitz continuous on an open neighborhood of t. We nevertheless use the above assumption for simplicity in statements.

Assumption 2. *We assume $v_1 \geq v_2 \geq \cdots \geq v_n$. Moreover, integer k is chosen to satisfy $k/n \to 0$ and $k/\log n \to \infty$.*

Although Assumption 2 as stated is more amenable for use for k-NN estimates, it can also be used for Nadaraya-Watson estimate, which will be clear in the next subsection. We also impose the following assumptions on the noise.

Assumption 3. *Given a convex increasing function ψ with $\psi(0) = 0$, and suppose for some constants $C > 0$, some concave increasing function ϕ with $\phi(0) = 0, \phi(1) = 1$, we have that $x^r \leq \phi(\psi(Cx))$ for some $r \geq 2$. Moreover, $M := \|\epsilon_i\|_\psi < \infty$ and the stationary sequence $(\epsilon_1, \epsilon_2, \ldots)$ is $\psi - m -$approximable with decay rate $\{\gamma_k\}$.*

In the above assumption, the Orlicz norm is used for bounding the tail probability of noises (Lemma 1 (i)) as well as controlling the dependence. It is possible of course to use different ψ for these two different purposes, but using the same ψ seems to be most natural since they concern the same noises. The assumption $x^r \leq \phi(\psi(Cx))$ deserves some explanation. By Lemma 1 (v), this implies that the r-th moment of the noise variable is finite, for some $r \geq 2$ and it is in particular satisfied by $\psi(x) = x^p$ for $p \geq r$ and $\psi(x) = \psi_q(x)$ for $q \geq 1$. When a stronger ψ-Orlicz norm is used, Assumption 3 imposes a stronger constraint, but the summability conditions in Theorem 1 below are easier to satisfy.

Our main result for functional nonparametric estimates with functional responses is the following. Due to that we aim for generality of the result here, the statement of the conditions is complicated involving summability of many sequences. We try to clarify the theorem with several remarks that follow. The theorem can accomodate both k-NN estimators and Nadaraya-Watson estimators, and reduces to known results when the data are independent with moment conditions on the error process.

Theorem 1. *Suppose assumptions 1, 2 and 3 hold, and $\sum_{i=k+1}^n v_i = O(b)$, $(\sum_{i=1}^n v_i^2)^{1/2} = O(c_2)$ with $b, c_2 \to 0$. Also, we denote by H the distance to t from its kth nearest neighbor, and we assume $H \to 0$. If one can find sequences $a_n \to 0, L_n \to 0, x_n \to 0, m_n$ with m_n an integer between 1 and n, such that (in the rest of the paper these sequences are simply denoted by a, L, x, m)*

- *(*) The four sequences, $\exp\{-Ca^2/(aL + m^2 c_2^2 + x)\}$ for some constant C big enough, $1/\psi(\sqrt{x/2}/(\gamma_1 c_2))$, $(m/a)/\psi^{1-1/r}(L/(2Mmv_1))$, and $1/\psi \times (a/(2nv_1\gamma_m))$, are all summable over n.*

Then $\|\hat{r}(t) - r(t)\| = O(b + H^\alpha + a + (\gamma_1 v_1)^{1/2})$ almost surely.

Remark 1. Here we present a unified result for both nearest-neighbor estimate and the Nadaraya-Watson estimate. The sequence m is related to the dependence of the data sequence, and roughly speaking we should choose larger m for data with stronger dependency. For the nearest-neighbor estimate, k is a pre-specified constant and typically b and c_2 are explicit functions of k and thus deterministic. On the other hand, H depends on k through (3) and thus depends on predictors. The situation for the Nadaraya-Watson estimate is exactly the opposite. H will be prespecified (typically as a function of sample size) and k is the number of predictors falling into the ball with radius H and thus depends on data. Similarly, v_i as order statistics of W_i depend on predictor values.

Remark 2. Because of the requirement $\sum_{n=1}^{\infty} \exp\{-Ca^2/(aL + m^2c_2^2 + x)\} < \infty$, we see that the sequence a cannot converge faster than mc_2 and thus we will focus on cases where this rate is achievable up to some logarithmic terms in the following.

Remark 3. In the convergence rate, b and H^α represent the bias while a comes from the variance of the estimator. For independent data, $\gamma_1 = 0$ and the term $(\gamma_1 v_1)^{1/2}$ does not appear. More generally, this term can be ignored as long as $v_1 = O(c_2^2)$, by Remark 2 above. As an example, we obviously have $v_1 = c_2^2 = 1/k$ for the simplest k-NN estimate with $v_i = 1/k, i \le k$. In the next subsection, one will see that for the Nadaraya-Watson estimate in Example 3 above, we also have that v_1 and c_2^2 are of the same order under mild assumptions.

As presented above, which aims for generality rather than clarity, it is hard to see what the convergence rate is in typical situations, and thus we discuss the rates in some special cases in the rest of this subsection.

Independent case When the data are independent, $1/\psi(\sqrt{x/2}/(\gamma_1 c_2))$ and $1/\psi(a/(2nv_1\gamma_m))$ are zero (Informally, $\gamma_m = 0$ when data are independent and we take $\psi(\infty) = \infty$. More rigorously, it can be seen from the proofs that these two terms are zero), and we can take $m = 1$, $x = 0$. Taking $L = c_2$ and $a = (\log n)c_2$, the first sequence in (*) is then obviously summable. So as long as $1/\left(a\psi^{1-1/r}\left(c_2/(2Mv_1)\right)\right)$ is summable, we have convergence rate $(\log n)c_2$. For the simplest nearest neighbor estimate with $v_i = 1/k, i \le k$, we have $c_2 = 1/\sqrt{k}$. The expression $1/a/\psi^{1-1/r}\left(c_2/(2Mv_1)\right)$ is simplified to $\sqrt{k}/((\log n)\psi^{1-1/r}(\sqrt{k}/2M))$. For $\psi(x) = x^p$ or $\psi(x) = \exp\{x^p\} - 1$, this obviously is a restriction on k, in particular that k should diverge fast enough at a certain rate. We note that by existing results on k-NN estimate for independent data with scalar responses, the variance term is expected to be $c_2 = 1/\sqrt{k}$, which agrees with the rate here up to a logarithmic term. In summary, we have

Corollary 1. *In the independent case, for the simplest k-NN estimate with* $v_i = 1/k, i \le k$, *if* $\sum_{n=1}^{n} \sqrt{k}/\psi^{1-1/r}(\sqrt{k}/2M) < \infty$ *where* $M = \|\epsilon_i\|_\psi$, *then* $\|\hat{r}(t) - r(t)\| = O(H^\alpha + (\log n)/\sqrt{k})$ *almost surely.*

This result is almost the same as Corollary 1 in Lian (2011), except that here we used the more general Orlicz norm for the errors (and thus an extra summability assumption is required). We also note that for the Nadaraya-Watson estimate in Example 3, discussions in the next subsection suggest that the convergence behavior is very much the same under reasonable assumptions.

It is worth pointing out here that H and k are related through the small ball probability, which plays a prominent role in nonparametric functional regression (Ferraty and Vieu, 2004, 2006). The details are contained in the next subsection.

Weakly dependent case Here the convergence rate is determined by the interplay of ψ and $\{\gamma_m\}$ in a more complicated way. For example, qualitatively, the summability of $1/\psi(a/(2nv_1\gamma_m))$ is easier to be satisfied the smaller is γ_m (weaker dependence). Moreover, the choice of x must take into account the trade off between the summability of $\exp\{-Ca^2/(aL+m^2c_2^2+x)\}$ and the summability of $1/\psi(\sqrt{x/2}/(\gamma_1 c_2))$ (the former is an increasing function of x while the latter is a decreasing function of x). Similarly, the choice of m must take into account the trade off between summability of $(m/a)/\psi^{1-1/r}(L/(2Mmv_1))$ and $1/\psi(a/(2nv_1\gamma_m))$ (the former is an increasing function of m while the latter is typically a decreasing function of m). Ignoring the complication of choosing m, when $\psi(x)=\psi_p(x)=\exp\{x^p\}-1$, the following corollary gives one possible situation where it is possible to set $a=mc_2$ up to an extra logarithmic term.

Corollary 2. *When $\psi=\psi_p, p\geq 1$, we have convergence rate $\|\hat{r}(t)-r(t)\| = O(b+H^\alpha+(\log n)^2 mc_2+(\gamma_1 v_1)^{1/2})$ as long as $1/\psi(C(\log n)^2 m/(n\gamma_m))$ is summable for C large enough.*

Proof. Take $x=C(\log n)^2c_2^2$ (C large enough) and $L=C(\log n)mc_2$, the first expression in (*) is then satisfied if $a=C(\log n)^2mc_2$. Moreover, $1/\psi(\sqrt{x/2}/(\gamma_1c_2)) \leq 1/\psi(C\log n)$ is summable. Using the trivial inequalities $c_2\geq v_1$ and $v_1\geq 1/n$, we get $m/a\leq n$ and thus $(m/a)/\psi^{1-1/r}(L/(2Mmv_1)) \leq n/\psi^{1-1/r}(C\log n)$ is summable. Finally, for the last sequence in (*), we have

$$\sum_n 1/\psi(a/(2nv_1\gamma_m)) \leq \sum_n 1/\psi(C(\log n)^2 m/(n\gamma_m)) < \infty,$$

by assumption in the statement of this corollary.

Finally, we note that in the above corollary, if $\gamma_m=e^{-Cm}$ for some $C>0$, then we can take $m\sim\log n$ so that all sequences in (*) are summable, and the rate of convergence is $(\log n)^3c_2$. □

2.3. On the properties of H and k with random covariates

In the previous subsection, we treat the predictor as fixed and the convergence rate depends on the sequence $\{X_i\}$. Here we study the behavior of some of the quantities that appeared in the rates when X_i is a random stationary sequence in typical situations. Results obtained in this subsection can be combined with Theorem 1 to obtain more explicit convergence rates. The necessity of studying H (for NN estimator) or k (for Nadaraya-Watson estimator) is seen from Remark 1 in the previous subsection.

When X_i are random, we will make use of the important quantity $\varphi(h) := P(\{t': t'\in B(t,h)\})$ which is called the small ball probability. Its importance has been demonstrated in Ferraty and Vieu (2006) for functional kernel regression with scalar responses. In particular, the use of $\varphi(h)$ in a functional setting replaces the common assumption on the existence of a density for X when X belongs to some Euclidean space. It is easy to see that in the classical setting with mild assumptions on the density of $X\in R^d$, we have $\varphi(h)\sim h^d$.

Nearest neighbor estimate We only consider the simplest k-NN estimate as in Example 1 with $v_i = 1/k, i = 1, \ldots, k$. Then in the convergence rates, $b = 0$, $c_2^2 = \sum_i W_i^2 = 1/k$ and $\max_i W_i = 1/k$. Thus only the quantity H depends on $\{X_i\}$. If the sequence $\{X_i\}$ contains independent elements, one can show $H = O(\varphi(2k/n))$ almost surely as in the following proposition, which was contained in Lian (2011).

Proposition 1. *In the independent case, suppose $k/n \to 0$ and $k/\log n \to \infty$. Let H be the distance from t to its k-th nearest neighbor as defined in (3), then $P(H > \varphi^{-1}(2k/n), i.o.) \to 0$, where i.o. means "infinitely often" and $\varphi^{-1}(x) := \inf\{h : \varphi(h) \geq x\}$.*

In Ferraty and Vieu (2006), the authors distinguished two types of processes: the fractal type processes and the exponential type processes. The former is characterized by $\phi(h) \sim h^\tau$, for some $\tau > 0$ and the latter characterized by $\phi(h) \sim \exp\{-(1/h^{\tau_1})\log(1/h^{\tau_2})\}, \tau_1 > 0, \tau_2 \geq 0$. The fractal type processes are similar to finite dimensional covariates in many aspects, while for infinite dimensional case such as when the covariate curves belong to some smoothness class, exponential type processes are more typical. For example, the Brownian motion is of exponential type. The paper van der Vaart and van Zanten (2008) provides other more complicated Gaussian processes all of which are of exponential type. Combining Proposition 1 above with Corollary 1, we obtain the rates $O([\varphi^{-1}(2k/n)]^\alpha + (\log n)/\sqrt{k})$ for independent data. When the optimal k is chosen, it is easy to see that for exponential type processes the convergence rates are logarithmic in the sample size, much slower than the classical finite-dimensional cases. Also note that this slow rate is largely determined by the term $[\varphi^{-1}(2k/n)]^\alpha$ which converges to zero logarithmically whether k increases logarithmically or polynomially in n.

For weakly dependent sequence $\{X_i\}$, in particular assuming $\{X_i\}$ is $\psi-m-$approximable with $\|d(X_1, X_1^{(m)})\|_\psi = \beta_m, \sum_{m=1}^\infty \beta_m < \infty$ (a minor extension to Definition 1 is needed here since $X_i \in \mathcal{F}$ which is not a normed space, thus we need to use $d(.,.)$ instead of $X_1 - X_1^{(m)}$), we can show the following proposition whose proof is deferred to the next section. Note that although we used the same notation as before, ψ here is different from that in Assumption 3 since here we are considering the predictor sequence instead of the noise sequence.

Proposition 2. *Suppose for some $h > \varphi^{-1}(2k/n)$, there exists some sequence $1 \leq m \leq n$ such that $k/n \to 0$, $k/(m\log n) \to \infty$ and $\sum_{n=1}^\infty n/\psi((h - \varphi^{-1}(2k/n))/\beta_m) < \infty$. Then we have $H \leq h$ for n large enough, almost surely.*

Nadaraya-Watson estimate Here $W_i = K(d(X_i,t)/H)/\sum_i K(d(X_i,t)/H)$ and we only consider the simple case where kernel function K satisfies $cI_{[-1,1]} \leq K \leq CI_{[-1,1]}$ for some $C > c > 0$. Unlike k-NN estimate, here H is predetermined. In Assumption 2, we let k be the number of covariates inside the ball $B(t,H)$ and thus if X_i is not one of the k nearest neighbors of t, we have $W_i = 0$ and thus $b = \sum_{k+1}^n v_i = 0$ in the convergence rate in Theorem 1. Since H is predetermined in Nadaraya-Watson estimates, the only quantity in the convergence rates that depends on X_i is $v_1 = \max_i W_i$ and $c_2 = (\sum_i W_i^2)^{1/2}$. Since $v_1 \le C/\sum_i K(d(X_i,t)/H) \le C/ck$, $v_1 \ge c/Ck$ as long as $k \ge 1$, and $c_2 \sim 1/\sqrt{k}$ which can be easily shown, we only need to study the asymptotic behavior of k, the number of predictors inside the ball $B(t,H)$.

With $\{X_i\}$ an independent sequence, we have

Proposition 3. *Suppose* $H \to 0$, $n\varphi(H)/\log n \to \infty$, *then* $n\varphi(H)/2 \le k \le 2n\varphi(H)$ *for* n *large enough, almost surely.*

On the other hand, for a $\psi - m-$approximable sequence $\{X_i\}$ with $\|d(X_1, X_1^{(m)})\|_\psi = \beta_m$, we have

Proposition 4. *Suppose* H'' *and* H' *are two sequences with* $H' < H < H''$ *and there exists a sequence* $1 \le m \le n$ *such that* $n\varphi(H')/(m\log n) \to \infty$, $\sum_{n=1}^\infty n/\psi((H''-H)/\beta_m) < \infty$ *and* $\sum_{n=1}^\infty n/\psi((H-H')/\beta_m) < \infty$. *Then we have* $n\varphi(H')/2 \le k \le 2n\varphi(H'')$ *for* n *large enough, almost surely.*

The proofs for these two propositions are very similar to those for Propositions 1 and 2, and thus omitted.

In the above we focused exclusively on the cases with $b = 0$ for simplicity. The extra bias $b \ne 0$ will appear when, for example, positive weights are put on all observations for k-NN estimates, or a noncompact kernel such as the Gaussian kernel is used in the Nadaraya-Watson estimates. The magnitude of b depends on the specific choices of the weights or kernel bandwidth.

3. Proofs

Based on two different representations of the nonparametric estimate in (1) and (2), we decompose $\|\hat{r}(t) - r(t)\|$ into the bias term and the variance term,

$$\|\hat{r}(t) - r(t)\| \le \|\sum_i v_i(r(X_{R_i}) - r(t))\| + \|\sum_i W_i\epsilon_i\|. \tag{4}$$

The bias term is easier to deal with. In fact,

$$\begin{aligned}\|\sum_i v_i(r(X_{R_i}) - r(t))\| &\le 2B\sum_{i=k+1}^n v_i + \|\sum_{i=1}^k v_i(r(X_{R_i}) - r(t))\| \\ &= O(b + H^\alpha),\end{aligned}$$

by Assumptions 1 and 2.

Now we deal with the variance term. Let $\eta_i = W_i\epsilon_i$, $S_n = \sum_{i=1}^n \eta_i$ and the following arguments are conditional on $\{X_1,\dots,X_n\}$ (in effect treating W_i as nonrandom weights). Following the idea of Section 6.3 in Ledoux and Talagrand (1991), we write $\|S_n\| - E\|S_n\| = \|\sum_{i=1}^n \eta_i\| - E\|\sum_{i=1}^n \eta_i\| = \sum_{i=1}^n e_i$, with $e_i = E[\|S_n\|\,|\mathcal{G}_i] - E[\|S_n\|\,|\mathcal{G}_{i-1}]$ where $\mathcal{G}_i$ is the $\sigma-$algebra generated by $\epsilon_1,\dots,\epsilon_i$ ($\mathcal{G}_0$ is the trivial $\sigma-$algebra). It is easy to see that $\{e_i\}$ is a *real-valued* martingale difference sequence which potentially enables us to use relevant exponential type inequalities. However, in general it seems at least not easy to obtain directly appropriate moment bounds for e_i in order to apply, for example, Lemma 8.9 in van der Geer (2000) (Bernstein's inequality for martingale differences), and thus we instead work with the quantity

$$\begin{aligned} d_i &= E[\|S_n\|\,|\mathcal{G}_i] - E[\|S_n\|\,|\mathcal{G}_{i-1}] - E[\|S_n - \eta_i - \cdots - \eta_{i+m-1}\|\,|\mathcal{G}_i] \\ &\quad + E[\|S_n - \eta_i - \cdots - \eta_{i+m-1}\|\,|\mathcal{G}_{i-1}], \end{aligned}$$

where m is the same as that in the statement of the theorem and, as discussed in Remarks following the theorem, need to be chosen appropriately (as a side note, $m = 1$ suffices for independent data in which case we actually have $d_i = e_i$). If $i + m - 1 > n$, the expression $S_n - \eta_i - \cdots - \eta_{i+m-1}$ is taken to mean $S_n - \eta_i - \cdots - \eta_n$. Obviously d_i is still a martingale difference sequence. We denote $f_i = E[\|S_n - \eta_i - \cdots - \eta_{i+m-1}\|\,|\mathcal{G}_i] - E[\|S_n - \eta_i - \cdots - \eta_{i+m-1}\|\,|\mathcal{G}_{i-1}]$ and thus $e_i = d_i + f_i$.

Lemma 2 shows that

$$|d_i| \le \sum_{j=i}^{i+m-1} W_j E(\|\epsilon_j\||\mathcal{G}_i) + \sum_{j=i}^{i+m-1} W_j E(\|\epsilon_j\||\mathcal{G}_{i-1}), \tag{5}$$

and

$$E(d_i^2|\mathcal{G}_{i-1}) \le m \sum_{j=i}^{i+m-1} W_j^2 E(\|\epsilon_j\|^2|\mathcal{G}_{i-1}). \tag{6}$$

Lemma 3 shows that

$$P(\sum_i f_i > a) \le 2/\psi\left(a/(2nv_1\gamma_m)\right). \tag{7}$$

Lemma 4 shows that

$$E\|S_n\| = O(c_2 + \sqrt{\gamma_1 v_1}). \tag{8}$$

Aided by these results, we can bound the variance term $\|S_n\|$ in three steps.

Step 1: Let $d_i' = d_i I\{|d_i| \le L\}$ for some $L > 0$. We have $P(\sum_{i=1}^n (d_i' - E(d_i'|\mathcal{G}_{i-1})) > a) \le \exp\{-Ca^2/(aL + m^2c_2^2 + x)\} + 1/\psi\left(\sqrt{x}/(\sqrt{2}\gamma_1 c_2)\right), \forall a \ge 0, x \ge 0$.

Let $\tilde{\psi}(x) := \psi(\sqrt{x})$. By Assumption 3, $\tilde{\psi}$ is convex and increasing and thus defines an Orlicz norm. Using (6), we have

$$\begin{aligned} &\sum_{i=1}^n E(d_i^2|\mathcal{G}_{i-1}) \\ \le\ & m \sum_{i=1}^n \sum_{j=i}^{i+m-1} W_j^2 E(\|\epsilon_j\|^2|\mathcal{G}_{i-1}) \\ =\ & m \sum_{i=1}^n \sum_{j=i}^{i+m-1} W_j^2 E(\|\epsilon_j^{(j-i+1)} + \epsilon_j - \epsilon_j^{(j-i+1)}\|^2|\mathcal{G}_{i-1}) \end{aligned}$$

$$\leq \quad 2m^2 E(\|\epsilon_1\|^2) \sum_{i=1}^{n} W_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i}^{i+m-1} W_j^2 E(\|\epsilon_j - \epsilon_j^{(j-i+1)}\|^2 | \mathcal{G}_{i-1}),$$

where in the last line above we use that $\epsilon_j^{(j-i+1)}$ is independent of $\mathcal{G}_{i-1}$, and also use the inequality $\|\epsilon_j^{(j-i+1)} + \epsilon_j - \epsilon_j^{(j-i+1)}\|^2 \leq 2\|\epsilon_j^{(j-i+1)}\|^2 + 2\|\epsilon_j - \epsilon_j^{(j-i+1)}\|^2$ which follows from the parallelogram identity. Furthermore,

$$\begin{aligned}
& \|2 \sum_{i=1}^{n} \sum_{j=i}^{i+m-1} W_j^2 E(\|\epsilon_j - \epsilon_j^{(j-i+1)}\|^2 | \mathcal{G}_{i-1})\|_{\tilde{\psi}} \\
\leq \quad & 2 \sum_{i=1}^{n} \sum_{j=i}^{i+m-1} W_j^2 \left\| E(\|\epsilon_j - \epsilon_j^{(j-i+1)}\|^2 | \mathcal{G}_{i-1}) \right\|_{\tilde{\psi}} \\
\leq \quad & 2 \sum_{i=1}^{n} \sum_{j=i}^{i+m-1} W_j^2 \left\| \|\epsilon_j - \epsilon_j^{(j-i+1)}\|^2 \right\|_{\tilde{\psi}} \\
\leq \quad & 2 \sum_{i=1}^{n} \sum_{j=i}^{i+m-1} W_j^2 \left(\|\epsilon_j - \epsilon_j^{(j-i+1)}\|_{\psi} \right)^2 \\
\leq \quad & 2\gamma_1^2 \sum_{i} W_i^2,
\end{aligned}$$

where we used Lemma 1 (vii) for the second inequality above and Lemma 1 (vi) for the third inequality above. Then, using Lemma 1 (i), we have for any $x \geq 0$

$$P(\sum_{i=1}^{n} E(d_i^2 | \mathcal{G}_{i-1}) > 2m^2 E(\|\epsilon_1\|^2) c_2^2 + x) \leq 1/\psi(\sqrt{x}/(\sqrt{2}\gamma_1 c_2)). \tag{9}$$

Using $|d_i' - E(d_i'|\mathcal{G}_{i-1})| \leq 2L$ and $E[(d_i' - E(d_i'|\mathcal{G}_{i-1}))^2 | \mathcal{G}_{i-1}] \leq E(d_i'^2 | \mathcal{G}_{i-1}) \leq E(d_i^2|\mathcal{G}_{i-1})$, we get $E(|d_i' - E(d_i'|\mathcal{G}_{i-1})|^k | \mathcal{G}_{i-1}) \leq (2L)^{k-2} E(d_i^2|\mathcal{G}_{i-1}), \forall k \geq 2$. Since $d_i' - E(d_i'|\mathcal{G}_{i-1}), i \leq n$ is a martingale difference sequence, using Lemma 8.9 in van der Geer (2000) (Bernstein's inequality for martingales) together with (9), we obtain the desired bound as follows:

$$\begin{aligned}
& P(\sum_{i=1}^{n} (d_i' - E(d_i'|\mathcal{G}_{i-1})) > a) \\
\leq \quad & P(\sum_{i=1}^{n} (d_i' - E(d_i'|\mathcal{G}_{i-1})) > a \text{ and } \sum_{i=1}^{n} E(d_i^2|\mathcal{G}_{i-1}) \leq 2m^2 E(\|\epsilon_1\|^2) c_2^2 + x) \\
& \quad + P(\sum_{i=1}^{n} E(d_i^2|\mathcal{G}_{i-1}) > 2m^2 E(\|\epsilon_1\|^2) c_2^2 + x) \\
\leq \quad & \exp\{-Ca^2/(aL + m^2 c_2^2 + x)\} + 1/\psi\left(\sqrt{x}/(\sqrt{2}\gamma_1 c_2)\right).
\end{aligned}$$

Step 2: Let $d_i'' = d_i - d_i' = d_i I\{|d_i| > L\}$. We have

$$P(\sum_{i} |d_i'' - E(d_i''|\mathcal{G}_{i-1})| > a) \leq Cm/\left(a\psi^{1-1/r}\left(\frac{L}{2Mmv_1}\right)\right),$$

where $M = \|\epsilon_1\|_{\psi}$.

From (5), we have that $\|d_i\|_\psi \le 2\sum_{j=i}^{i+m-1} W_j\|\epsilon_j\|_\psi = 2M\sum_{j=i}^{i+m-1} W_j$, and thus using Lemma 1 (i) and (v), $P(d_i > L) \le 1/\psi(\frac{L}{2M\sum_{j=i}^{i+m-1} W_j})$, and $\|d_i\|_r \le C\|d_i\|_\psi \le C\sum_{j=i}^{i+m-1} W_j$.

Using Hölder's inequality, we have

$$\begin{aligned}
& E(|d_i'' - E(d_i''|\mathcal{G}_{i-1})|) \\
\le\ & 2E(|d_i''|) \\
=\ & 2E(|d_i|I\{|d_i| > L\}) \\
\le\ & 2\{E(|d_i|^r)\}^{1/r} P(|d_i| > L)^{1-1/r} \\
\le\ & C(\sum_{j=i}^{i+m-1} W_j)/\psi^{1-1/r}\left(\frac{L}{2M\sum_{j=i}^{i+m-1} W_j}\right) \\
\le\ & C(\sum_{j=i}^{i+m-1} W_j)/\psi^{1-1/r}\left(\frac{L}{2Mmv_1}\right),
\end{aligned}$$

and thus, using Markov's inequality, we have $P(\sum_i |d_i'' - E(d_i''|\mathcal{G}_{i-1})| > a) \le E[\sum_i |d_i'' - E(d_i''|\mathcal{G}_{i-1})|]/a \le Cm/\left(a\psi^{1-1/r}\left(\frac{L}{2Mmv_1}\right)\right)$.

Step 3: Finally, we demonstrate the bound for the variance term in (4).

Using $E(d_i|\mathcal{G}_{i-1}) = E(d_i'|\mathcal{G}_{i-1}) + E(d_i''|\mathcal{G}_{i-1}) = 0$, we have that $d_i = d_i' - E(d_i'|\mathcal{G}_{i-1}) + (d_i'' - E(d_i''|\mathcal{G}_{i-1}))$ and then

$$\begin{aligned}
& P(\|S_n\| - E\|S_n\| > 3a) \\
=\ & P(\sum_i d_i + f_i > 3a) \\
\le\ & P(\sum_i d_i > 2a) + P(\sum_i f_i > a) \\
\le\ & P(\sum_i (d_i' - E(d_i'|\mathcal{G}_{i-1})) > a) + P(\sum_i (d_i'' - E(d_i''|\mathcal{G}_{i-1})) > a) \\
& + P(\sum_i f_i > a) \\
\le\ & \exp\{-Ca^2/(aL + m^2c_2^2 + x)\} + 1/\psi\left(\sqrt{x}/(\sqrt{2}\gamma_1 c_2)\right) \\
& + C(m/a)/\psi^{1-1/r}\left(\frac{L}{2Mmv_1}\right) + 2/\psi\left(a/(2nv_1\gamma_m)\right),
\end{aligned}$$

by the previous two steps and (7). The above expression is summable by assumption of the theorem, and an application of the Borel-Cantelli Lemma leads to $\|S_n\| - E\|S_n\| = O(a)$. Combining this with (8), the variance term is thus $\|S_n\| = O(a + c_2 + (\gamma_1 v_1)^{1/2})$. As noted in Remark 1 following the theorem, the term c_2 can be omitted since we always have $c_2 = O(a)$.

Lemma 2. *Using the notation in the proof of Theorem 1, we have*

$$|d_i| \leq \sum_{j=i}^{i+m-1} E(\|\eta_j\||\mathcal{G}_i) + \sum_{j=i}^{i+m-1} E(\|\eta_j\||\mathcal{G}_{i-1}),$$
$$E(d_i^2|\mathcal{G}_{i-1}) \leq m \sum_{j=i}^{i+m-1} E(\|\eta_j\|^2|\mathcal{G}_{i-1}).$$

Proof. Since $d_i = E(\|S_n\| - \|S_n - \sum_{j=i}^{i+m-1} \eta_j\||\mathcal{G}_i) - E(\|S_n\| - \|S_n - \sum_{j=i}^{i+m-1} \eta_j\||\mathcal{G}_{i-1})$, the first equation is obvious.

Denote $\xi_i = E(\|S_n\| - \|S_n - \sum_{j=i}^{i+m-1} \eta_j\||\mathcal{G}_i)$, then $d_i = \xi_i - E(\xi_i|\mathcal{G}_{i-1})$. Using the interpretation of $E(\xi_i|\mathcal{G}_{i-1})$ as the projection of ξ_i, we have $E(d_i^2|\mathcal{G}_{i-1}) \leq E(\xi_i^2|\mathcal{G}_{i-1}) \leq m \sum_{j=i}^{i+m-1} E(\|\eta_j\|^2|\mathcal{G}_{i-1})$, proving the second equation. □

Lemma 3. *For* $f_i = E[\|S_n - \eta_i - \cdots - \eta_{i+m-1}\| \,|\mathcal{G}_i] - E[\|S_n - \eta_i - \cdots - \eta_{i+m-1}\| \,|\mathcal{G}_{i-1}]$ *as in the proof of Theorem 1, we have*

$$P(\sum_i f_i > a) \leq 2/\psi\left(a/(2n(\max_{1\leq j\leq n} W_j)\gamma_m)\right).$$

Proof. By the definition of $\psi-m-$approximability for sequence ϵ_i, we have that

$$\begin{aligned} & E[\|W_1\epsilon_1 + \cdots + W_{i-1}\epsilon_{i-1} + W_{i+m}\epsilon_{i+m}^{(m)} + \cdots + W_n\epsilon_n^{(n-i)}\| \,|\mathcal{G}_i] \\ = \; & E[\|W_1\epsilon_1 + \cdots + W_{i-1}\epsilon_{i-1} + W_{i+m}\epsilon_{i+m}^{(m)} + \cdots + W_n\epsilon_n^{(n-i)}\| \,|\mathcal{G}_{i-1}]. \end{aligned}$$

Thus $f_i = f_i' - f_i''$ where

$$\begin{aligned} f_i' \; = \; & E[\|W_1\epsilon_1 + \cdots + W_{i-1}\epsilon_{i-1} + W_{i+m}\epsilon_{i+m} + \cdots + W_n\epsilon_n\| \,|\mathcal{G}_i] \\ & - E[\|W_1\epsilon_1 + \cdots + W_{i-1}\epsilon_{i-1} + W_{i+m}\epsilon_{i+m}^{(m)} + \cdots + W_n\epsilon_n^{(n-i)}\| \,|\mathcal{G}_i], \\ f_i'' \; = \; & E[\|W_1\epsilon_1 + \cdots + W_{i-1}\epsilon_{i-1} + W_{i+m}\epsilon_{i+m} + \cdots + W_n\epsilon_n\| \,|\mathcal{G}_{i-1}] \\ & - E[\|W_1\epsilon_1 + \cdots + W_{i-1}\epsilon_{i-1} + W_{i+m}\epsilon_{i+m}^{(m)} + \cdots + W_n\epsilon_n^{(n-i)}\| \,|\mathcal{G}_{i-1}]. \end{aligned}$$

Since $|f_i'| \leq E(\|W_{i+m}\epsilon_{i+m} - W_{i+m}\epsilon_{i+m}^{(m)}\| + \cdots + \|W_n\epsilon_n - W_n\epsilon_n^{(n-i)}\| \,|\mathcal{G}_i)$, using Lemma 1 (vii), we have $\|f_i'\|_\psi \leq (\max_{1\leq j\leq n} W_j)\gamma_m$ and thus $\|\sum_{i=1}^n f_i'\|_\psi \leq n(\max_{1\leq j\leq n} W_j)\gamma_m$. Using Lemma 1 (i) we get

$$P(\sum_i f_i' > a/2) \leq 1/\psi\left(a/(2n(\max_{1\leq j\leq n} W_j)\gamma_m)\right).$$

By exactly the same arguments $P(\sum_i f_i'' > a/2) \leq 1/\psi(a/(2n(\max_{1\leq j\leq n} W_j)\gamma_m))$, and the Lemma is proved. □

Lemma 4. *Let* $S_n = \sum_{i=1}^n \eta_i = \sum_{i=1}^n W_i\epsilon_i$ *as in Theorem 1, we have* $E\|S_n\| = O\left(c_2 + \sqrt{\gamma_1 \max_i W_i}\right)$.

Proof. We have

$$
\begin{aligned}
& (E\|S_n\|)^2 \\
= \;& (E\|\sum_i W_i\epsilon_i\|)^2 \le E(\|\sum_i W_i\epsilon_i\|^2) \\
= \;& E\sum_i\sum_j W_iW_j\langle\epsilon_i,\epsilon_j\rangle \\
= \;& \sum_i W_i^2 E\|\epsilon_i\|^2 + 2\sum_{i=1}^n\sum_{j=i+1}^n W_iW_jE\langle\epsilon_i,\epsilon_j\rangle \\
= \;& O(c_2^2) + 2\sum_{i=1}^n\sum_{j=i+1}^n W_iW_jE\langle\epsilon_i,\epsilon_j-\epsilon_j^{(j-i)}\rangle \\
\le \;& O(c_2^2) + 2\sum_{i=1}^n\sum_{j=i+1}^n W_iW_jE(\|\epsilon_i\|\,\|\epsilon_j-\epsilon_j^{(j-i)}\|) \\
\le \;& O(c_2^2) + 2\sum_{i=1}^n\sum_{j=i+1}^n W_iW_j(E(\|\epsilon_i\|^2)^{1/2}(E\|\epsilon_j-\epsilon_j^{(j-i)}\|^2)^{1/2} \\
\le \;& O(c_2^2) + C\sum_{i=1}^n\sum_{j=i+1}^n W_iW_j\|\epsilon_i\|_\psi\|\epsilon_j-\epsilon_j^{(j-i)}\|_\psi,
\end{aligned}
$$

where we used that ϵ_j and $\epsilon_j^{(j-i)}$ are independent, and Assumption 3 on ψ. Finally, we see that $\sum_{i=1}^n\sum_{j=i+1}^n W_iW_j\|\epsilon_i\|_\psi\|\epsilon_j-\epsilon_j^{(j-i)}\|_\psi \le (\max_j W_j)(\sum_i W_i)\times \sum_{m=1}^\infty \|\epsilon_1-\epsilon_1^{(j-i)}\|_\psi = O(\gamma_1 \max_i W_i)$. □

Proof of Proposition 2. Consider the approximating sequence $(X_1^{(m)}, X_2^{(m)}, \ldots)$. Define the zero mean random variables $Y_i^{(m)} = I\{X_i^{(m)} \in B(x,c)\} - \tilde{p}$ where $c = \varphi^{-1}(2k/n)$ and $\tilde{p} = \varphi(c) = 2k/n$. Divide the sequence $(Y_1^{(m)}, Y_2^{(m)}, \ldots)$ into m groups (we assume n/m is an integer for simplicity in presentation without loss of generality) as follows:

$$
\begin{aligned}
\text{group 1:}\quad & Y_1^{(m)}, Y_{1+m}^{(m)}, Y_{1+2m}^{(m)}, \ldots, Y_{1+(n/m-1)m}^{(m)}, \\
\text{group 2:}\quad & Y_2^{(m)}, Y_{2+m}^{(m)}, Y_{2+2m}^{(m)}, \ldots, Y_{2+(n/m-1)m}^{(m)}, \\
\vdots\quad & \vdots \\
\text{group } m\text{:}\quad & Y_m^{(m)}, Y_{2m}^{(m)}, Y_{3m}^{(m)}, \ldots, Y_n^{(m)}.
\end{aligned}
$$

Because of the construction, the random variables within one group are independent of each other. Let $Z_i, i = 1, \ldots m$ be the sum of random variables within each group. Using Bernstein's inequality, we have $P(|Z_i| > x) \le$

$2\exp\{-\frac{1}{2}x^2/(n\tilde{p}/m + x/3)\}$, and thus

$$
\begin{aligned}
& P(\sum_{i=1}^{n} I\{X_i^{(m)} \in B(x,c)\} \le k) \\
\le \quad & P(|\sum_{i=1}^{n}(I\{X_i^{(m)} \in B(x,c)\} - \tilde{p})| \ge n\tilde{p} - k) \\
= \quad & P(|\sum_{i=1}^{m} Z_i| \ge n\tilde{p} - k) \\
\le \quad & mP(|Z_1| > (n\tilde{p} - k)/m) \\
\le \quad & 2m\exp\{-\frac{1}{2}(\frac{n\tilde{p} - k}{m})^2/(n\tilde{p}/m + (n\tilde{p} - k)/(3m)) \\
= \quad & 2m\,\exp\{-(3/14)k/m\}. \qquad (10)
\end{aligned}
$$

We also have that

$$
\begin{aligned}
& P(H > h) \le P(\sum_{i=1}^{n} I\{X_i \in B(x,h)\} \le k) \\
\le \quad & P(\sum_{i=1}^{n} I\{X_i^{(m)} \in B(x,c)\} \le k) \\
& +P(\exists i, s.t.\ X_i^{(m)} \in B(x,c) \text{ and } X_i \notin B(x,h)) \\
\le \quad & P(\sum_{i=1}^{n} I\{X_i^{(m)} \in B(x,c)\} \le k) + P(\exists i, s.t.\ d(X_i^{(m)}, X_i) > h - c) \\
\le \quad & P(\sum_{i=1}^{n} I\{X_i^{(m)} \in B(x,c)\} \le k) + n/\psi((h-c)/\beta_m) \qquad (11) \\
\le \quad & 2m\,\exp\{-(3/14)k/m\} + n/\psi((h-c)/\beta_m), \qquad (12)
\end{aligned}
$$

where we used Lemma 1 (i) in (11) and used (10) in (12). The lemma follows from the Borel-Cantelli Lemma. □

Acknowledgements

The author thanks the Editor, the Associate Editor and two referees for their helpful comments that lead to significant improvement of the manuscript. This research is supported by a Singapore MOE Tier 1 Grant 36/09.

References

Aguilera, A., Ocana, F. and Valderrama, M. (2008). Estimation of functional regression models for functional responses by wavelet approximation. In *Functional and Operatorial Statistics*, (S. DaboNiang and F. Ferraty, eds.). *Contributions to Statistics* 15-21. MR2478482

Antoch, J., Prchal, L., De Rosa, M. R. and Sarda, P. (2008). Functional linear regression with functional response: Application to prediction of electricity consumption. In *Functional and Operatorial Statistics*, (S. DaboNiang and F. Ferraty, eds.). *Contributions to Statistics* 23-29. MR2478483

Biau, G., Cerou, F. and Guyader, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory* **56** 2034-2040. MR2654492

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34** 2159-2179. MR2291496

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters* **45** 11-22. MR1718346

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics* **37** 35-72. MR2488344

Crambes, C. and Mas, A. (2009). Asymptotics of prediction in the functional linear regression with functional outputs. *Arxiv preprint arXiv:0910.3070*.

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics-Revue Canadienne De Statistique* **30** 285-300. MR1926066

Febrero-Bande, M., Galeano, P. and Gonzalez-Manteiga, W. (2010). Measures of influence for the functional linear model with scalar response. *Journal of Multivariate Analysis* **101** 327-339. MR2564343

Ferraty, F., Goia, A. and Vieu, P. (2002). Functional nonparametric model for time series: a fractal approach for dimension reduction. *Test* **11** 317-344. MR1947601

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics* **17** 545-564. MR1952697

Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *Journal of nonparametric statistics* **16** 111-125. MR2053065

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. *Springer series in statistics*. Springer, New York, NY. MR2229687

Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2011). Kernel regression with functional response. *Electronic Journal of Statistics* **5** 159–171. MR2786486

Gabrys, R., Horváth, L. and Kokoszka, P. (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association* **105** 1113-1125. MR2752607

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35** 70-91. MR2332269

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics* **38** 1845-1884. MR2662361

Horváth, L., Kokoszka, P. and Reimherr, M. (2009). Two sample inference in functional linear models. *Canadian Journal of Statistics* **37** 571-591. MR2588950

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64** 411-432. MR1924298

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces: isoperimetry and processes*. Springer-Verlag, Berlin ; New York. MR1102015

Lian, H. (2007). Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics-Revue Canadienne De Statistique* **35** 597-606. MR2381399

Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics* **5** 31–40. MR2773606

Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics* **33** 774-805. MR2163159

Preda, C. (2007). Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference* **137** 829-840. MR2301719

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. Springer, New York. MR1910407

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*, 2nd ed. *Springer series in statistics*. Springer, New York. MR2168993

van der Geer, S. A. (2000). *Applications of empirical process theory*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36** 1435-1463. MR2418663

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Verlag. MR1385671

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics* **33** 2873-2903. MR2253106