OPTIMAL UNIFORM CONVERGENCE RATES AND ASYMPTOTIC NORMALITY FOR SERIES ESTIMATORS UNDER WEAK DEPENDENCE AND WEAK CONDITIONS

By

Xiaohong Chen and Timothy M. Christensen

December 2014

COWLES FOUNDATION DISCUSSION PAPER NO. 1976

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions[*]

Xiaohong Chen[†] and Timothy M. Christensen[‡]

First version January 2012; Revised August 2013, July 2014

Abstract

We show that spline and wavelet series regression estimators for weakly dependent regressors attain the optimal uniform (i.e. sup-norm) convergence rate $(n/\log n)^{-p/(2p+d)}$ of Stone (1982), where d is the number of regressors and p is the smoothness of the regression function. The optimal rate is achieved even for heavy-tailed martingale difference errors with finite $(2+(d/p))$th absolute moment for $d/p<2$. We also establish the asymptotic normality of t statistics for possibly nonlinear, irregular functionals of the conditional mean function under weak conditions. The results are proved by deriving a new exponential inequality for sums of weakly dependent random matrices, which is of independent interest.

JEL Classification: C12, C14, C32

Keywords: Nonparametric series regression; Optimal uniform convergence rates; Weak dependence; Random matrices; Splines; Wavelets; (Nonlinear) Irregular Functionals; Sieve t statistics

[*] We thank the guest coeditor, two anonymous referees, Bruce Hansen, Jianhua Huang, Stefan Schneeberger, and conference participants of SETA2013 in Seoul and AMES2013 in Singapore for useful comments. This paper is an expanded version of Sections 4 and 5 of our Cemmap Working Paper CWP56/13 (Chen and Christensen, 2013), the remaining parts of which are currently undergoing a major revision. Support from the Cowles Foundation is gratefully acknowledged. Any errors are the responsibility of the authors.

[†] Corresponding Author. Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, CT 06520, USA. Tel: +1 203 432 5852; fax: +1 203 432 6167. E-mail address: `xiaohong.chen@yale.edu`

[‡] Department of Economics, New York University, 19 West 4th Street, New York, NY 10012, USA. E-mail address: `timothy.christensen@nyu.edu`

1 Introduction

We study the nonparametric regression model

$$\begin{aligned} Y_i &= h_0(X_i) + \epsilon_i \\ E[\epsilon_i|X_i] &= 0 \end{aligned} \tag{1}$$

where $Y_i \in \mathbb{R}$ is a scalar response variable, $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a d-dimensional regressor (predictor variable), and the conditional mean function $h_0(x) = E[Y_i|X_i = x]$ belongs to a Hölder space of smoothness $p > 0$. We are interested in series least squares (LS) estimation[1] of h_0 under sup-norm loss and inference on possibly nonlinear functionals of h_0 allowing for weakly dependent regressors and heavy-tailed errors ϵ_i.

For i.i.d. data, Stone (1982) shows that $(n/\log n)^{-p/(2p+d)}$ is the minimax lower bound in sup-norm risk for estimation of h_0 over a Hölder ball of smoothness $p > 0$. For strictly stationary beta-mixing regressors, we show that spline and wavelet series LS estimators $\widehat{h}$ of h_0 attain the optimal uniform rate of Stone (1982) under a mild unconditional moment condition $E[|\epsilon_i|^{2+(d/p)}] < \infty$ imposed on the martingale difference errors.

More generally, we assume the error process $\{\epsilon_i\}_{i=-\infty}^{\infty}$ is a martingale difference sequence but impose no explicit weak dependence condition on the regressor process $\{X_i\}_{i=-\infty}^{\infty}$. Rather, weak dependence of the regressor process is formulated in terms of convergence of a certain random matrix. We verify this condition for absolutely regular (beta-mixing) sequences by deriving a new exponential inequality for sums of weakly dependent random matrices. This new inequality then leads to a sharp upper bound on the sup-norm variance term of series LS estimators with an arbitrary basis. When combined with a general upper bound on the sup-norm bias term of series LS estimators, the sharp sup-norm variance bound immediately leads to a general upper bound on the sup-norm convergence rate of series LS estimators with an arbitrary basis and weakly dependent data.

In our sup-norm bias and variance decomposition of series LS estimators, the bound on the sup-norm bias term depends on the sup norm of the empirical L^2 projection onto the linear sieve space. The sup norm of the empirical L^2 projection varies with the choice of the (linear sieve) basis. For spline regression with i.i.d. data, Huang (2003b) shows that the sup norm of the empirical L^2 projection onto

[1] Other terms for series LS appearing in the literature include series regression and linear sieve regression, but we use series LS hereafter. The series LS estimator falls into the general class of nonparametric sieve M-estimators.

splines is bounded with probability approaching one (wpa1). Using our new exponential inequality for sums of weakly dependent random matrices, his bound is easily extended to spline regression with weakly dependent regressors. In addition, we show that, for either i.i.d. or weakly dependent regressors, the sup norm of the empirical L^2 projection onto compactly supported wavelet bases is also bounded wpa1. These tight bounds lead to sharp sup-norm bias control for spline and wavelet series LS estimators. They in turn imply that spline and wavelet series LS estimators achieve the optimal sup-norm convergence rate even for weakly dependent data and heavy-tailed errors.[2]

Sup-norm (uniform) convergence rates of series LS estimators have previously been studied by Newey (1997), de Jong (2002), Song (2008) and Chen and Liao (2014) for i.i.d. data, and Lee and Robinson (2013) for spatially dependent data. But the uniform convergence rates obtained in these papers are slower than the optimal rate of Stone (1982).[3] In a rough note, Chen and Huang (2003) derived the optimal sup-norm rate for spline series LS estimators with i.i.d. data under the condition $E[|\epsilon_i|^{2+\delta}] < \infty$ for some $\delta > d/p$.[4] In an independent work, Belloni, Chernozhukov, Chetverikov, and Kato (2014) show that spline and local polynomial partition series[5] LS estimators attain the optimal sup-norm rate with i.i.d. data under the conditional moment condition $\sup_x E[|\epsilon_i|^{2+\delta} |X_i = x] < \infty$ for some $\delta > d/p$. Our result contributes to the literature by establishing that spline and wavelet series LS estimators attain the optimal sup-norm rate with either i.i.d. data or strictly stationary beta-mixing regressors under the weaker unconditional moment requirement $E[|\epsilon_i|^{2+(d/p)}] < \infty$.

As another application of our new exponential inequality, under very weak conditions we obtain sharp L^2 convergence rates for series LS estimators with weakly dependent regressors. For example, under the minimal bounded conditional second moment restriction ($\sup_x E[|\epsilon_i|^2 |X_i = x] < \infty$), our L^2-norm rates for trigonometric polynomial, spline or wavelet series LS estimators attain Stone (1982)'s optimal L^2-norm rate of $n^{-p/(2p+d)}$ with strictly stationary, exponentially beta-mixing (respectively algebraically beta-mixing at rate γ) regressors with $p > 0$ (resp. $p > d/(2\gamma)$), while the power series LS estimator attains the same optimal rate with exponentially (resp. algebraically) beta-mixing regressors

[2] The error ϵ_i is heavy-tailed in the sense that $E[|\epsilon_i|^{2+\delta}] = \infty$ for $\delta > d/p$ is allowed; say $E[|\epsilon_i|^4] = \infty$ is allowed.

[3] See Hansen (2008), Kristensen (2009), Masry (1996), Cattaneo and Farrell (2013) and the references therein for attainability of the optimal uniform convergence rates with kernel, local linear regression and partitioning estimators.

[4] The authors did not pay attention to the fact that their proof for the optimal sup-norm rate of spline LS estimator actually allows for $\delta > d/p$. They set $\delta = 2$ for the optimal sup-norm rate, but did not like the strong condition $E[|\epsilon_i|^4] < \infty$ and hence did not finish the paper. The authors did circulate the note among some colleagues and their former students who work in this area.

[5] Belloni et al. (2014) recast the local polynomial partitioning estimator of Cattaneo and Farrell (2013) as a series LS estimator.

for $p > d/2$ (resp. $p > d(2+\gamma)/(2\gamma)$). It is interesting to note that for a smooth conditional mean function, we obtain the optimal L^2 convergence rates for these commonly used series LS estimators with weakly dependent regressors without requiring the existence of higher-than-second unconditional moments of the error terms.

We also show that feasible asymptotic inference can be performed on a possibly nonlinear functional $f(h_0)$ using the plug-in series LS estimator $f(\widehat{h})$. We establish the asymptotic normality of $f(\widehat{h})$ and of the corresponding Student t statistic for weakly dependent data under mild low-level conditions. When specializing to general irregular (i.e., slower than $\sqrt{n}$-estimable) but sup-norm bounded *linear* functionals of spline or wavelet series LS estimators with i.i.d. data, we obtain the asymptotic normality of $f(\widehat{h})$

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{V_K^{1/2}} \to_d N(0,1)$$

under remarkably mild conditions of (1) uniform integrability ($\sup_{x \in \mathcal{X}} E[\epsilon_i^2 \{|\epsilon_i| > \ell(n)\}|X_i = x] \to 0$ for any $\ell(n) \to \infty$ as $n \to \infty$), and (2) $K^{-p/d}\sqrt{n/V_K} = o(1)$, $(K \log K)/n = o(1)$, where K is the sieve number of terms, and V_K is the sieve variance that grows with K for irregular functionals. These conditions coincide with the weakest known conditions in Huang (2003b) for the pointwise asymptotic normality of spline LS estimators, except we also allow for other irregular linear functionals of spline or wavelet LS estimators. When specializing to general irregular but sup-norm bounded *nonlinear* functionals of spline or wavelet series LS estimators with i.i.d. data, we obtain asymptotic normality of $f(\widehat{h})$ (and of its t statistic) under conditions (1) and (3) $K^{-p/d}\sqrt{n/V_K} = o(1)$, $K^{(2+\delta)/\delta}(\log n)/n \lesssim 1$ (and $K^{(2+\delta)/\delta}(\log n)/n = o(1)$ for the t statistic) for $\delta \in (0,2)$ such that $E[|\epsilon_i|^{2+\delta}] < \infty$. These conditions are much weaker than the well-known conditions in Newey (1997) for the asymptotic normality of a nonlinear functional and its t statistic of spline LS estimator, namely $K^{-p/d}\sqrt{n} = o(1)$, $K^4/n = o(1)$ and $\sup_x E[|\epsilon_i|^4 |X_i = x] < \infty$. Moreover, under a slightly more restrictive growth condition on K but without the need to increase δ, we show that our mild sufficient conditions for the i.i.d. case extend naturally to the weakly dependent case.

Since economic and financial time series data often have infinite forth moments, the new improved rates and inference results in our paper should be very useful to the literatures on nonparametric estimation and testing of nonlinear time series models (see, e.g., Robinson (1989), Li, Hsiao, and Zinn (2003), Fan and Yao (2003), Chen (2013)). Moreover, our new exponential inequality for sums of weakly dependent random matrices should be useful in series LS estimation of spatially dependent models and in other contexts as well.[6]

The rest of the paper is organized as follows. Section 2 first derives general upper bounds on the sup-norm convergence rates of series LS estimators with an arbitrary basis. It then shows that spline and wavelet series LS estimators attain the optimal sup-norm rates, allowing for weakly dependent data and heavy tailed error terms. It also presents general sharp L^2-norm convergence rates of series LS estimators with an arbitrary basis under very mild conditions. Section 3 provides the asymptotic normality of sieve t statistics for possibly nonlinear functionals of h_0. Section 4 provides new exponential inequalities for sums of weakly dependent random matrices, and a reinterpretation of equivalence of the theoretical and empirical L^2 norms as a criterion regarding convergence of a certain random matrix. Section 5 shows the sup-norm stability of the empirical L^2 projections onto compactly supported wavelet bases, which provides a tight upper bound on the sup-norm bias term for the wavelet series LS estimator. The results in Sections 4 and 5 are of independent interest. Section 6 contains a brief review of spline and wavelet sieve bases. Proofs and ancillary results are presented in Section 7.

Notation: Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues, respectively, of a matrix. The exponent $^-$ denotes the Moore-Penrose generalized inverse. $\|\cdot\|$ denotes the Euclidean norm when applied to vectors and the matrix spectral norm (i.e., largest singular value) when applied to matrices, and $\|\cdot\|_{\ell^p}$ denotes the ℓ^p norm when applied to vectors and its induced operator norm when applied to matrices (thus $\|\cdot\| = \|\cdot\|_{\ell^2}$). If $\{a_n : n \geq 1\}$ and $\{b_n : n \geq 1\}$ are two sequences of non-negative numbers, $a_n \lesssim b_n$ means there exists a finite positive C such that $a_n \leq C b_n$ for all n sufficiently large, and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. $\#\mathcal{S}$ denotes the cardinality of a set $\mathcal{S}$ of finitely many elements. Given a strictly stationary process $\{X_i\}$ and $1 \leq p < \infty$, we let $L^p(X)$ denote the function space consisting of all (equivalence classes) of measurable functions f for which the $L^p(X)$ norm $\|f\|_{L^p(X)} \equiv E[|f(X_i)|^p]^{1/p}$ is finite, and we let $L^\infty(X)$ denote the space of bounded functions under the sup norm $\|\cdot\|_\infty$, i.e., if $f : \mathcal{X} \to \mathbb{R}$ then $\|f\|_\infty \equiv \sup_{x \in \mathcal{X}} |f(x)|$.

2 Uniform Convergence Rates

In this section we present some general results on uniform convergence properties of nonparametric series LS estimators with weakly dependent data.

[6] In our ongoing work on sieve estimation of semi/nonparametric conditional moment restriction models with time series data, this new exponential inequality also enables us to establish asymptotic properties under weaker conditions.

2.1 Estimator and basic assumptions

In nonparametric series LS estimation, the conditional mean function h_0 is estimated by least squares regression of $Y_1, \dots, Y_n$ on a vector of sieve basis functions evaluated at $X_1, \dots, X_n$. The standard series LS estimator of the conditional mean function h_0 is

$$\widehat{h}(x) = b^K(x)'(B'B)^- B'Y \tag{2}$$

where $b_{K1}, \dots, b_{KK}$ are a collection of K sieve basis functions and

$$\begin{aligned} b^K(x) &= (b_{K1}(x), \dots, b_{KK}(x))' && (3) \\ B &= (b^K(X_1), \dots, b^K(X_n))' && (4) \\ Y &= (Y_1, \dots, Y_n)' \,. && (5) \end{aligned}$$

Choosing a particular class of sieve basis function and the dimension K are analogous to choosing the type of kernel and bandwidth, respectively, in kernel regression techniques. The basis functions are chosen such that their closed linear span $B_K = clsp\{b_{K1}, \dots, b_{KK}\}$ can well approximate the space of functions in which h_0 is assumed to belong.

When the data $\{(X_i, Y_i)\}_{i=1}^n$ are a random sample it is often reasonable to assume that X is supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$. However, in a time-series setting it may be necessary to allow the support $\mathcal{X}$ of X to be infinite, as in the example of nonparametric autoregressive regression with a student t distributed error term. See, e.g., Fan and Yao (2003) and Chen (2013) for additional examples and references.

To allow for possibly unbounded support $\mathcal{X}$ of X_i we modify the usual series LS estimator and notion of convergence. First, we weight the basis functions by a sequence of non-negative weighting functions $w_n : \mathcal{X} \to \{0, 1\}$ given by

$$w_n(x) = \begin{cases} 1 & \text{if } x \in \mathcal{D}_n \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\mathcal{D}_n \subseteq \mathcal{X}$ is compact, convex, and has nonempty interior, and $\mathcal{D}_n \subseteq \mathcal{D}_{n+1}$ for all n. The resulting series LS estimator is then

$$\widehat{h}(x) = b_w^K(x)'(B_w'B_w)^- B_w'Y \tag{7}$$

where $b_{K1}, \ldots, b_{KK}$ are a collection of K sieve basis functions and

$$\begin{aligned} b_w^K(x) &= (b_{K1}(x)w_n(x), \ldots, b_{KK}(x)w_n(x))' \qquad (8) \\ B_w &= (b_w^K(X_1), \ldots, b_w^K(X_n))' \,. \qquad (9) \end{aligned}$$

Second, we consider convergence in the (sequence of) weighted sup norm(s) $\|\cdot\|_{\infty,w}$ given by

$$\|f\|_{\infty,w} = \sup_x |f(x)w_n(x)| = \sup_{x\in\mathcal{D}_n} |f(x)| \tag{10}$$

This modification is made because simple functions, such as polynomials of x, have infinite sup norm when X_i has unbounded support, but will have finite weighted sup norm.

Remark 2.1 *When $\mathcal{X}$ is compact we may simply set $w_n(x) = 1$ for all $x \in \mathcal{X}$ and all n. With such a choice of weighting, the series LS estimator with weighted basis trivially coincides with the series LS estimator in (2) with unweighted basis, and $\|\cdot\|_{\infty,w} = \|\cdot\|_\infty$.*

When $\mathcal{X}$ is unbounded there are several possible choices for $\mathcal{D}_n$. For instance, we may take $\mathcal{D}_n = \mathcal{D}$ for all n, where $\mathcal{D} \subseteq \mathcal{X}$ is a fixed compact convex set with nonempty interior. This approach is not without precedent in the nonparametric analysis of nonlinear time series models. For example, Huang and Shen (2004) use a similar approach to trim extreme observations in nonparametric functional coefficient regression models, following Tjøstheim and Auestad (1994). More generally, we can consider an expanding sequence of compact nonempty sets $\mathcal{D}_n \subset \mathcal{X}$ with $\mathcal{D}_n \subseteq \mathcal{D}_{n+1}$ for all n and set $w_n(x) = \{x \in \mathcal{D}_n\}$ for all n. For example, if $\mathcal{X} = \mathbb{R}^d$ we could take $\mathcal{D}_n = \{x \in \mathbb{R}^d : \|x\|_p \leq r_n\}$ where $0 < r_n \leq r_{n+1} < \infty$ for all n. This approach is similar to excluding functions far from the support of the data when performing series LS estimation with a compactly-supported wavelet basis for $L^2(\mathbb{R})$ or $L^2(\mathbb{R}^d)$. We defer estimation with smooth weighting functions of the form $w_n(x) = (1 + \|x\|^2)^{-\omega}$ or $w_n(x) = \exp(-\|x\|^\omega)$ to future research.

We first introduce some mild regularity conditions that are satisfied by typical regression models and most linear sieve bases.

Assumption 1 *(i) $\{X_i\}_{i=-\infty}^{\infty}$ is strictly stationary, (ii) $\mathcal{X} \subseteq \mathbb{R}^d$ is convex and has nonempty interior.*

Assumption 2 *(i) $\{\epsilon_i, \mathcal{F}_{i-1}\}_{i=1}^n$ with $\mathcal{F}_{i-1} = \sigma(X_i, \epsilon_{i-1}, X_{i-1}, \ldots)$ is a strictly stationary martingale difference sequence, (ii) $E[\epsilon_i^2|\mathcal{F}_{i-1}]$ is uniformly bounded for all $i \geq 1$, almost surely, (iii) $E[|\epsilon_i|^{2+\delta}] < \infty$ for some $\delta > 0$.*

Let $N(\mathcal{D}_n, \epsilon)$ denote the internal ϵ-covering number of $\mathcal{D}_n$ with respect to the Euclidean norm (i.e. the minimum number of points $x_1, \ldots, x_m \in \mathcal{D}_n$ such that the collection of ϵ-balls centered at each of $x_1, \ldots, x_m$ cover $\mathcal{D}_n$).

Assumption 3 *(i) $\mathcal{D}_n$ is compact, convex, has nonempty interior, and $\mathcal{D}_n \subseteq \mathcal{D}_{n+1}$ for all n, (ii) there exists $\nu_1, \nu_2 > 0$ such that $N(\mathcal{D}_n, \epsilon) \lesssim n^{\nu_1}\epsilon^{-\nu_2}$.*

Define $\zeta_{K,n} \equiv \sup_x \|b_w^K(x)\|$ and $\lambda_{K,n} \equiv \left[\lambda_{\min}(E[b_w^K(X_i)b_w^K(X_i)'])\right]^{-1/2}$.

Assumption 4 *(i) there exist $\omega_1, \omega_2 \geq 0$ s.t. $\sup_{x \in \mathcal{D}_n} \|\nabla b_w^K(x)\| \lesssim n^{\omega_1} K^{\omega_2}$, (ii) there exist $\varpi_1 \geq 0, \varpi_2 > 0$ s.t. $\zeta_{K,n} \lesssim n^{\varpi_1} K^{\varpi_2}$, (iii) $\lambda_{\min}(E[b_w^K(X_i)b_w^K(X_i)']) > 0$ for each K and n.*

Assumptions 1 and 2 trivially nest i.i.d. sequences, but also allow the regressors to exhibit quite general weak dependence. Note that Assumption 2(ii) reduces to $\sup_x E[\epsilon_i^2|X_i = x] < \infty$ in the i.i.d. case. Suitable choice of δ in Assumption 2(iii) for attainability of the optimal uniform rate will be explained subsequently. Strict stationarity of $\{\epsilon_i\}$ in Assumption 2 may be dropped provided the sequence $\{|\epsilon_i|^{2+\delta}\}$ is uniformly integrable. However, strict stationarity is used to present simple sufficient conditions for the asymptotic normality of functionals of $\widehat{h}$ in Section 3.

Assumption 3 is trivially satisfied when $\mathcal{X}$ is compact and $\mathcal{D}_n = \mathcal{X}$ for all n. More generally, when $\mathcal{X}$ is noncompact and $\mathcal{D}_n$ is an expanding sequence of compact subsets of $\mathcal{X}$ as described above, Assumption 3(ii) is satisfied provided each $\mathcal{D}_n$ is contained in an Euclidean ball of radius $r_n \lesssim n^\nu$ for some $\nu > 0$.[7]

Assumption 4 is a mild regularity condition on the sieve basis functions. When $\mathcal{X}$ is compact and rectangular this assumption is satisfied by all the widely used series (or linear sieve bases) with $\lambda_{K,n} \lesssim 1$, and $\zeta_{K,n} \lesssim \sqrt{K}$ for tensor-products of univariate polynomial spline, trigonometric polynomial or

[7] By translational invariance we may assume that $\mathcal{D}_n$ is centered at the origin. Then $\mathcal{D}_n \subseteq \mathcal{R}_n = [-r_n, r_n]^d$. We can cover $\mathcal{R}_n$ with $(r_n/\epsilon)^d$ ℓ^∞-balls of radius ϵ, each of which is contained in an Euclidean ball of radius $\epsilon\sqrt{d}$. Therefore, $N(\mathcal{D}_n, \epsilon) \leq (\sqrt{d}r_n)^d \epsilon^{-d} \lesssim n^{\nu d}\epsilon^{-d}$.

wavelet bases, and $\zeta_{K,n} \lesssim K$ for tensor-products of power series or orthogonal polynomial bases (see, e.g., Newey (1997), Huang (1998), and Chen (2007)). See DeVore and Lorentz (1993) for additional bases with either $\zeta_{K,n} \lesssim \sqrt{K}$ or $\zeta_{K,n} \lesssim \sqrt{K}$ properties.

Let $\widetilde{b}_w^K(x)$ denote the orthonormalized vector of basis functions, namely

$$\widetilde{b}_w^K(x) = E[b_w^K(X_i)b_w^K(X_i)']^{-1/2}b_w^K(x)\,, \tag{11}$$

and let $\widetilde{B}_w = (\widetilde{b}_w^K(X_1), \ldots, \widetilde{b}_w^K(X_n))'$.

Assumption 5 *Either: (a) $\{X_i\}_{i=1}^n$ is i.i.d. and $\zeta_{K,n}\lambda_{K,n}\sqrt{(\log K)/n} = o(1)$, or*
(b) $\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| = o_p(1)$.

Assumption 5 is a mild but powerful condition that ensures the empirical and theoretical L^2 norms are equivalent over the linear sieve space wpa1 (see Section 4 for details). In fact, to establish many of our results below with weakly dependent data, nothing further about the weak dependence properties of the regressor process $\{X_i\}_{i=1}^n$ needs to be assumed beyond convergence of $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\|$ to zero. In the i.i.d. case, the following Lemma shows that part (a) of Assumption 5 automatically implies $\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| = o_p(1)$.

Lemma 2.1 *Under Assumption 4(iii), if $\{X_i\}_{i=1}^n$ is i.i.d. then*

$$\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| = O_p\left(\zeta_{K,n}\lambda_{K,n}\sqrt{(\log K)/n}\right) = o_p(1)$$

provided $\zeta_{K,n}\lambda_{K,n}\sqrt{(\log K)/n} = o(1)$.

Remark 2.2 *Consider the compact support case in which $\mathcal{X} = [0,1]^d$ and $w_n(x) = 1$ for all $x \in \mathcal{X}$ and all n (so that $b_{Kk}(x)w_n(x) = b_{Kk}(x)$ for all n and K) and suppose the density of X_i is uniformly bounded away from zero and infinity over $\mathcal{X}$. In this setting, we have $\lambda_{K,n} \lesssim 1$. If $\{X_i\}_{i=1}^n$ is i.i.d., then Assumption 5 is satisfied with $\sqrt{(K \log K)/n} = o(1)$ for spline, trigonometric polynomial or wavelet bases, and with $K\sqrt{(\log K)/n} = o(1)$ for (tensor-product) power series.*

When the regressors are β-mixing (see Section 4 for definition), the following Lemma shows that Assumption 5(b) is still easily satisfied.

Lemma 2.2 *Under Assumption 4(iii), if $\{X_i\}_{i=-\infty}^{\infty}$ is strictly stationary and β-mixing with mixing coefficients such that one can choose an integers $q = q(n) \leq n/2$ with $\beta(q)n/q = o(1)$, then*

$$\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| = O_p\left(\zeta_{K,n}\lambda_{K,n}\sqrt{(q\log K)/n}\right) = o_p(1)$$

provided $\zeta_{K,n}\lambda_{K,n}\sqrt{(q\log K)/n} = o(1)$.

Remark 2.3 *Consider the compact support case from Remark 2.2, with $\{X_i\}_{i=-\infty}^{\infty}$ strictly stationary and β-mixing.*

(i) Exponential β-mixing: Assumption 5(b) is satisfied with $\sqrt{K(\log n)^2/n} = o(1)$ for (tensor-product) spline, trigonometric polynomial or wavelet bases, and with $K\sqrt{(\log n)^2/n} = o(1)$ for (tensor-product) power series.

(ii) Algebraic β-mixing at rate γ: Assumption 5(b) is satisfied with $\sqrt{(K\log K)/n^{\gamma/(1+\gamma)}} = o(1)$ for (tensor-product) spline, trigonometric polynomial or wavelet bases, and with $K\sqrt{(\log K)/n^{\gamma/(1+\gamma)}} = o(1)$ for (tensor-product) power series.

2.2 A general upper bound on uniform convergence rates

Let $B_{K,w} = clsp\{b_{K1}w_n, \ldots, b_{KK}w_n\}$ be a general weighted linear sieve space. Let $\widetilde{h}$ denote the projection of h_0 onto $B_{K,w}$ under the empirical measure, that is,

$$\widetilde{h}(x) = b_w^K(x)'(B_w'B_w)^- B_w'H_0 = \widetilde{b}_w^K(x)'(\widetilde{B}_w'\widetilde{B}_w)^-\widetilde{B}_w'H_0 \tag{12}$$

where $H_0 = (h_0(X_1), \ldots, h_0(X_n))'$. The sup-norm distance $\|\widehat{h} - h_0\|_{\infty,w}$ may be trivially bounded using

$$\begin{aligned} \|\widehat{h} - h_0\|_{\infty,w} &\leq \|h_0 - \widetilde{h}\|_{\infty,w} + \|\widehat{h} - \widetilde{h}\|_{\infty,w} && (13) \\ &=: \text{bias term} + \text{variance term}\,. && (14) \end{aligned}$$

Sharp bound on the sup-norm variance term. The following result establishes a sharp uniform convergence rate of the variance term for an arbitrary linear sieve space. Convergence is established in sup norm rather than the weighted sup norm $\|\cdot\|_{\infty,w}$ because both $\widehat{h}$ and $\widetilde{h}$ have support $\mathcal{D}_n$. Therefore, $\|\widehat{h}-\widetilde{h}\|_\infty = \sup_{x\in\mathcal{D}_n}|\widehat{h}(x)-\widetilde{h}(x)| = \|\widehat{h}-\widetilde{h}\|_{\infty,w}$.

Lemma 2.3 *Let Assumptions 1(i)(ii), 2(i)(ii)(iii), 3, 4, and 5 hold. Then*

$$\|\widehat{h}-\widetilde{h}\|_\infty = O_p\left(\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right) = o_p(1)$$

as $n, K\to\infty$ provided the following are satisfied:

(i) $(\zeta_{K,n}\lambda_{K,n})^{(2+\delta)/\delta} \lesssim \sqrt{(n/\log n)}$;

(ii) either: (a) $\{(X_i,Y_i)\}_{i=1}^n$ *are i.i.d., or (b)* $\sqrt{\frac{K}{\log n}}\times\|(\widetilde{B}_w'\widetilde{B}_w/n)-I_K\| = O_p(1)$.

Remark 2.4 *Weak dependence of the regressor process $\{X_i\}$ is implicitly captured by the speed of convergence of $\|(\widetilde{B}_w'\widetilde{B}_w/n)-I_K\|$. If $\{X_i\}$ is exponentially β-mixing (respectively algebraically β-mixing at rate γ), condition (ii)(b) in Lemma 2.3 is satisfied provided $\zeta_{K,n}\lambda_{K,n}\sqrt{(K\log n)/n} = O(1)$ (respectively $\zeta_{K,n}\lambda_{K,n}\sqrt{K/n^{\gamma/(1+\gamma)}} = O(1)$); see Lemma 2.2.*

General bound on the sup-norm bias term. With our sharp bound on the variance term $\|\widehat{h}-\widetilde{h}\|_{\infty,w}$ in hand it remains to provide a calculation for the bias term $\|h_0-\widetilde{h}\|_{\infty,w}$. Let $P_{K,w,n}$ be the (empirical) projection operator onto $B_{K,w}\equiv clsp\{b_{K1}w_n,\ldots,b_{KK}w_n\}$, namely

$$P_{K,w,n}h(x) = b_w^K(x)'\left(\frac{B_w'B_w}{n}\right)^{-}\frac{1}{n}\sum_{i=1}^n b_w^K(X_i)h(X_i) = \widetilde{b}_w^K(x)'(\widetilde{B}_w'\widetilde{B}_w)^{-}\widetilde{B}_w'H \tag{15}$$

where $H=(h(X_1),\ldots,h(X_n))'$. $P_{K,w,n}$ is a well defined operator: if $L^2_{w,n}(X)$ denotes the space of functions with norm $\|\cdot\|_{w,n}$ where $\|f\|^2_{w,n} = \frac{1}{n}\sum_{i=1}^n f(X_i)^2 w_n(X_i)$, then $P_{K,w,n}: L^2_{w,n}(X)\to L^2_{w,n}(X)$ is an orthogonal projection onto $B_{K,w}$ whenever $B_w'B_w$ is invertible (which it is wpa1 under Assumptions 4(iii) and 5).

One way to control the bias term $\|h_0-\widetilde{h}\|_{\infty,w}$ is to bound $P_{K,w,n}$ in sup norm. Note that $\widetilde{h} = P_{K,w,n}h_0$. Let $L^\infty_{w,n}(X)$ denote the space of functions for which $\sup_x|f(x)w_n(x)|<\infty$ and let

$$\|P_{K,w,n}\|_{\infty,w} = \sup_{h\in L^\infty_{w,n}(X):\|h\|_{\infty,w}\neq 0}\frac{\|P_{K,w,n}h\|_{\infty,w}}{\|h\|_{\infty,w}}$$

denote the (weighted sup) operator norm of $P_{K,w,n}$. The following crude bound on $\|P_{K,w,n}\|_\infty$ is valid for general linear sieve bases and weakly dependent regressors.

Remark 2.5 *Let Assumptions 4(iii) and 5 hold. Then:* $\|P_{K,w,n}\|_\infty \le \sqrt{2}\zeta_{K,n}\lambda_{K,n}$ *wpa1.*

More refined bounds on $\|P_{K,w,n}\|_{\infty,w}$ may be derived for particular linear sieves with local properties, such as splines and wavelets stated below. These more refined bounds, together with the following Lemma, lead to the optimal uniform convergence rates of series LS estimators with the particular linear sieves.

Lemma 2.4 *Let the assumptions and conditions of Lemma 2.3 hold. Then: (1)*

$$\|\widehat{h} - h_0\|_\infty \le O_p\left(\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right) + (1 + \|P_{K,w,n}\|_{\infty,w}) \inf_{h \in B_{K,w}} \|h_0 - h\|_{\infty,w}.$$

(2) Further, if the linear sieve satisfies $\zeta_{K,n}\lambda_{K,n} \lesssim \sqrt{K}$ *and* $\|P_{K,w,n}\|_{\infty,w} = O_p(1)$*, then*

$$\|\widehat{h} - h_0\|_\infty \le O_p\left(\sqrt{(K \log n)/n} + \inf_{h \in B_{K,w}} \|h_0 - h\|_{\infty,w}\right).$$

2.3 Attainability of optimal uniform convergence rates

We now turn to attainability of the optimal uniform rate of Stone (1982) by specific series LS estimators. To fix ideas, in what follows we take $\mathcal{D}_n = \mathcal{D} = [0,1]^d \subseteq \mathcal{X}$ for all n, whence $\|f\|_{\infty,w} = \sup_{x \in \mathcal{D}} |f(x)|$. Let $\Lambda^p([0,1]^d)$ denote a Hölder space of smoothness p on the domain $[0,1]^d$ (see, e.g. Chen (2007) for definition). Let $\mathrm{BSpl}(K,[0,1]^d,\gamma)$ denote a B-spline sieve of degree γ and dimension K on the domain $[0,1]^d$, and let $\mathrm{Wav}(K,[0,1]^d,\gamma)$ denote a Wavelet sieve basis of regularity γ and dimension K on the domain $[0,1]^d$ (see Section 6 for details on construction of these sieve bases). Because our bases have been constructed to have support $[0,1]^d$ we trivially have $b_{Kk}(x) = b_{Kk}(x)w_n(x)$ for all $k = 1,\ldots,K$ and all n and K. Recall $B_K \equiv clsp\{b_{K1},\ldots,b_{KK}\}$ is the linear sieve space.

The following assumptions on the conditional mean function and the sieve basis functions are sufficient for attaining the optimal uniform convergence rate.

Assumption 1 (continued) *(iii)* $\mathcal{D}_n = \mathcal{D} = [0,1]^d \subseteq \mathcal{X}$ *for all* n*, (iv) the unconditional density of* X_i *is uniformly bounded away from zero and infinity on* $\mathcal{D}$*.*

Assumption 6 *The restriction of* h_0 *to* $[0,1]^d$ *belongs to* $\Lambda^p([0,1]^d)$ *for some* $p > 0$*.*

Assumption 7 *The sieve* B_K *is BSpl*$(K,[0,1]^d,\gamma)$ *or Wav*$(K,[0,1]^d,\gamma)$ *with* $\gamma > \max\{p,1\}$*.*

Assumptions 1 and 6 are standard regularity conditions used in derivation of optimal uniform convergence rates (Stone, 1982; Tsybakov, 2009). Assumption 1(iii) implies Assumption 3. Assumptions 1 and 7 imply Assumption 4 with $\zeta_{K,n} \lesssim \sqrt{K}$ and $\lambda_{K,n} \lesssim 1$.

Let $h_{0,K}^* \in B_K$ solve $\inf_{h \in B_K} \|h_0 - h\|_{\infty,w}$. Assumptions 1, 6 and 7 imply that $\|h_0 - h_{0,K}^*\|_{\infty,w} \lesssim K^{-p/d}$ (see, e.g. DeVore and Lorentz (1993), Huang (1998), Chen (2007)). Previously Huang (2003b) showed that $\|P_{K,w,n}\|_{\infty,w} \lesssim 1$ wpa1 for spline bases with i.i.d. data. In the proof of Theorem 2.1 we extend his result to allow for weakly dependent regressors. In addition, Theorem 5.2 in Section 5 shows that $\|P_{K,w,n}\|_{\infty,w} \lesssim 1$ wpa1 for wavelet bases with i.i.d. or weakly dependent regressors.

Theorem 2.1 *Let Assumptions 1, 2(i)(ii)(iii) (with $\delta \geq d/p$), 6 and 7 hold. If $K \asymp (n/\log n)^{d/(2p+d)}$, then*

$$\|\widehat{h} - h_0\|_{\infty,w} = O_p((n/\log n)^{-p/(2p+d)})$$

provided that either (a), (b), or (c) is satisfied:

(a) $\{(X_i, Y_i)\}_{i=1}^n$ is i.i.d.;

(b) $\{X_i\}_{i=1}^n$ is exponentially β-mixing and $d < 2p$;

(c) $\{X_i\}_{i=1}^n$ is algebraically β-mixing at rate γ and $(2+\gamma)d < 2\gamma p$.

Theorem 2.1 states that the optimal uniform convergence rates of Stone (1982) are achieved by spline and wavelet series LS estimators with i.i.d. data whenever $\delta \geq d/p$. If the regressors are exponentially β-mixing the optimal rate of convergence is achieved with $\delta \geq d/p$ and $d < 2p$. The restrictions $\delta \geq d/p$ and $(2+\gamma)d < 2\gamma p$ for algebraically β-mixing (at a rate γ) reduces naturally towards the exponentially β-mixing restrictions as the dependence becomes weaker (i.e. γ becomes larger). In all cases, for a fixed dimension $d \geq 1$, a smoother function (i.e. bigger p) means a lower value of δ, and hence fatter-tailed error terms ϵ_i, are permitted while still obtaining the optimal uniform convergence rate. In particular this is achieved with $\delta = d/p < 2$.

Discussion of closely related results. Under Assumption 1 with i.i.d. data and compact $\mathcal{X}$, we can set the weight to be $w_n = 1$ for all n. Let P_K denote the $L^2(X)$ orthogonal projection operator onto B_K, given by

$$P_K h(x) = b^K(x)' \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} E[b^K(X_i)h(X_i)] \tag{16}$$

for any $h \in L^2(X)$, and define its L^∞ operator norm:

$$\|P_K\|_\infty := \sup_{h \in L^\infty(X): \|h\|_\infty \neq 0} \frac{\|P_K h\|_\infty}{\|h\|_\infty}. \tag{17}$$

Under i.i.d. data, Assumptions 1, 2(i)(ii) (i.e., $E[\epsilon_i|X_i] = 0$, $\sup_x E[|\epsilon_i|^2 |X_i = x] < \infty$), 6, and the conditions $\lambda_{K,n} \lesssim 1$, $\zeta_{K,n}^2 K/n = o(1)$ and $\|h_0 - h_{0,K}^*\|_\infty \lesssim K^{-p/d}$ on the series basis, Newey (1997) derived the following sup-norm convergence rates for series LS estimators with an arbitrary basis:

$$\|\widehat{h} - h_0\|_\infty \leq O_p\left(\zeta_{K,n}\sqrt{K/n} + \zeta_{K,n} K^{-p/d}\right). \tag{18}$$

By Remark 2.5, under the same set of mild conditions imposed in Newey (1997) (except allowing for weakly dependent regressors), the bound (18) can be slightly improved to

$$\|\widehat{h} - h_0\|_\infty \leq O_p\left(\zeta_{K,n}\sqrt{K/n} + \|P_{K,w,n}\|_{\infty,w} K^{-p/d}\right). \tag{19}$$

It is clear that the general bounds in (18) and (19) with an arbitrary basis are not optimal, but they are derived under the minimal moment restriction of Assumption 2(ii) without the existence of higher-than-second moments (i.e., $\delta = 0$ in Assumption 2(iii)).

Under the extra moment condition $\sup_x E[|\epsilon_i|^4 |X_i = x] < \infty$, de Jong (2002) obtained the following general bound on sup-norm rates for series LS estimators with an arbitrary basis:

$$\|\widehat{h} - h_0\|_\infty \leq O_p\left(\zeta_{K,n}\sqrt{(\log n)/n} + K^{-p/d} + \|P_K h_0 - h_{0,K}^*\|_\infty\right). \tag{20}$$

de Jong (2002) did not provide sharp bounds for $\|P_K h_0 - h_{0,K}^*\|_\infty$ for any particular basis, and was therefore unable to attain the optimal convergence rate $\|\widehat{h} - h_0\|_\infty = O_p((n/\log n)^{-p/(2p+d)})$ of Stone (1982). Note that

$$\|P_K h_0 - h_{0,K}^*\|_\infty = \|P_K(h_0 - h_{0,K}^*)\|_\infty \leq \|P_K\|_\infty \|h_0 - h_{0,K}^*\|_\infty \lesssim \|P_K\|_\infty K^{-p/d}. \tag{21}$$

Given the newly derived sharp bounds of $\|P_K\|_\infty \lesssim 1$ in Huang (2003b) for splines, in Belloni et al. (2014) for the local polynomial partition series, and in our paper (Theorem 5.1) for wavelets, one could now apply de Jong (2002)'s result (20) to conclude the attainability of the optimal sup-norm rate by spline, local polynomial partition and wavelet series LS estimators for i.i.d. data. However, de Jong (2002)'s result (20) is proved under the strong bounded conditional fourth moment condition of $\sup_x E[|\epsilon_i|^4 |X_i = x] < \infty$ and the side condition $\zeta_{K,n}^2 K/n = o(1)$.

In a rough note, Chen and Huang (2003) derived $\|\widehat{h}-h_0\|_\infty = O_p(\sqrt{(K\log n)/n}+K^{-p/d})$ for spline LS estimators under i.i.d. data and condition $E[|\epsilon_i|^{2+\delta}] < \infty$ for some $\delta > d/p$, but they carelessly set $\delta = 2$ to attain Stone (1982)'s optimal rate and concluded that the finite fourth moment condition $E[|\epsilon_i|^4] < \infty$ is too strong. Cattaneo and Farrell (2013) proved that a local polynomial partitioning regression estimator can attain the optimal sup-norm rate under i.i.d. data and the conditional moment condition $\sup_x E[|\epsilon_i|^{2+\delta} |X_i = x] < \infty$ for some $\delta \geq \max(1, d/p)$. Belloni et al. (2014) show that spline and local polynomial partition LS estimators attain the optimal sup-norm rate under i.i.d. data and the conditional moment condition $\sup_x E[|\epsilon_i|^{2+\delta} |X_i = x] < \infty$ for some $\delta > d/p$. By contrast, we require a weaker unconditional moment condition $E[|\epsilon_i|^{2+(d/p)}] < \infty$ for spline and wavelet LS estimators to attain the optimal uniform convergence rate, allowing for both i.i.d. data and weakly dependent regressors. It remains an open question whether one could obtain the optimal sup-norm convergence rate without imposing a finite higher-than-second unconditional moment of the error term, however.

2.4 A general sharp bound on L^2 convergence rates

In this subsection, we present a simple but sharp upper bound on the L^2 (or root mean square) convergence rates of series LS estimators with an arbitrary basis and weakly dependent regressors.

Recall that $B_{K,w} = clsp\{b_{K1}w_n, \ldots, b_{KK}w_n\}$ is a general weighted linear sieve space and $\widetilde{h} = P_{K,w,n}h_0$ is defined in (12). Let $h_{0,K}$ be the $L^2(X)$ orthogonal projection of the conditional mean function h_0 onto $B_{K,w}$.

Lemma 2.5 *Let Assumptions 1(i), 2(i)(ii), 4(iii) and 5 hold. Then:*

$$\|\widehat{h} - \widetilde{h}\|_{L^2(X)} = O_p\left(\sqrt{K/n}\right) \text{ and } \|\widetilde{h} - h_0\|_{L^2(X)} = O_p\left(\|h_0 - h_{0,K}\|_{L^2(X)}\right) .$$

Lemmas 2.5, 2.1 and 2.2 immediately imply the following result.

Remark 2.6 *Let Assumptions 1(i), 2(i)(ii) and* $\lambda_{K,n} \lesssim 1$ *hold. Then: (1)*

$$\|\widehat{h} - h_0\|_{L^2(X)} = O_p\left(\sqrt{K/n} + \|h_0 - h_{0,K}\|_{L^2(X)}\right), \tag{22}$$

provided that either (1.a), (1.b) or (1.c) is satisfied:

(1.a) $\{X_i\}_{i=1}^n$ *is i.i.d., and* $\zeta_{K,n}\sqrt{(\log K)/n} = o(1)$*;*

(1.b) $\{X_i\}_{i=1}^n$ *is exponentially* β*-mixing, and* $\zeta_{K,n}\sqrt{(\log n)^2/n} = o(1)$*;*

(1.c) $\{X_i\}_{i=1}^n$ *is algebraically* β*-mixing at rate* γ*, and* $\zeta_{K,n}\sqrt{(\log K)/n^{\gamma/(1+\gamma)}} = o(1)$*.*

(2) Further, if Assumptions 1 and 6 hold and $K \asymp (n/\log n)^{d/(2p+d)}$*, then*

$$\|\widehat{h} - h_0\|_{L^2(X)} = O_p(n^{-p/(2p+d)})$$

provided that either (2.a) or (2.b) is satisfied:

(2.a) $\{X_i\}_{i=1}^n$ *is i.i.d. or exponentially* β*-mixing: with* $p > 0$ *for trigonometric polynomial, spline or wavelet series, and* $p > d/2$ *for power series;*

(2.b) $\{X_i\}_{i=1}^n$ *is algebraically* β*-mixing at rate* γ*: with* $p > d/(2\gamma)$ *for trigonometric polynomial, spline or wavelet series; and* $p > d(2+\gamma)/(2\gamma)$ *for power series.*

With i.i.d. data under the condition $\lambda_{K,n} \lesssim 1$, Newey (1997) derived the same sharp L^2 rate in (22) for series LS estimators under the restriction $\zeta_{K,n}^2 K/n = o(1)$. Huang (2003a) showed that spline LS estimator has the same L^2 rate under the much weaker condition $K(\log K)/n = o(1)$. Both our Remark 2.6 part (1.a) and Belloni et al. (2014) extend Huang (2003a)'s weakened condition to other bases satisfying $\zeta_{K,n} \lesssim \sqrt{K}$ (such as trigonometric polynomial and wavelet) for series LS regression with i.i.d. data. In addition, Remark 2.6 part (1.b) shows that the mild condition $K(\log K)^2/n = o(1)$ suffices for trigonometric polynomial, wavelet, spline and other bases satisfying $\zeta_{K,n} \lesssim \sqrt{K}$ for exponentially β-mixing regressors.

With weakly-dependent data, Chen and Shen (1998) derived L^2 rates for LS regression using various linear or nonlinear sieves with beta-mixing sequence under higher-than-second moment restriction (see Proposition 5.1 in Chen and Shen (1998)). Huang and Yang (2004) and others derived the optimal L^2 rate for spline LS regression with strongly mixing sequence assuming a uniformly bounded higher-than-second conditional moment. Thanks to our Lemma 2.2, we are able to show that series LS estimators with arbitrary bases attain the optimal L^2 convergence rate with beta-mixing regressors under a uniformly bounded second conditional moment condition on the residuals. Our result should be very useful to nonparametric series regression for financial time series data with heavy-tailed errors.

3 Inference on possibly nonlinear functionals

We now study inference on possibly nonlinear functionals $f : L^2(X) \cap L^\infty(X) \to \mathbb{R}$ of the regression function h_0. Examples of functionals include, but are not limited to, the pointwise evaluation functional, the partial mean functional, and consumer surplus (see, e.g., Newey (1997) for examples). The functional $f(h_0)$ may be estimated using the plug-in series LS estimator $f(\widehat{h})$, for which we now establish feasible limit theory.

As with Newey (1997) and Chen, Liao, and Sun (2014), our results allow researchers to perform inference on nonlinear functionals f of h_0 without needing to know whether or not $f(h_0)$ is regular (i.e., $\sqrt{n}$-estimable). However, there is already a large literature on the $\sqrt{n}$-asymptotic normality and the consistent variance estimation for series estimators of regular functionals of conditional mean functions with weakly dependent data (see, e.g., Chen and Shen (1998), Chen (2007), Li and Racine (2006)). To save space and to illustrate the usefulness of our new sup-norm convergence rate results, we focus on asymptotic normality of $f(\widehat{h})$ and the corresponding sieve t statistic when the functional is irregular (i.e., slower than $\sqrt{n}$-estimable) in this section.

We borrow some notation and definitions from Chen et al. (2014). Denote the pathwise derivative of f at h_0 in the direction $v \in \mathcal{V} := (L^2(X) - \{h_0\})$ by

$$\frac{\partial f(h_0)}{\partial h}[v] := \lim_{\tau \to 0^+} \frac{f(h_0 + \tau v)}{\tau} \tag{23}$$

and assume it is linear. Let $v_K^* \in \mathcal{V}_K := (B_{K,w} - \{h_{0,K}\})$ be the sieve Riesz representer of $\frac{\partial f(h_0)}{\partial h}[\cdot]$ on V_K, i.e. v_K^* is the unique element of $\mathcal{V}_K$ such that

$$\frac{\partial f(h_0)}{\partial h}[v] = E[v_K^*(X_i)v(X_i)] \quad \text{for all} \quad v \in \mathcal{V}_K. \tag{24}$$

It is straightforward to verify that

$$v_K^*(\cdot) = b_w^K(\cdot)' \left(E[b_w^K(X_i)b_w^K(X_i)']\right)^{-1} \frac{\partial f(h_0)}{\partial h}[b_w^K] = \widetilde{b}_w^K(\cdot)' \frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K] \tag{25}$$

where $\frac{\partial f(h_0)}{\partial h}[b_w^K]$ is understood to be the vector formed by evaluating $\frac{\partial f(h_0)}{\partial h}[\cdot]$ at each element of $b_w^K(\cdot)$. Let $\|v_K^*\|_{L^2(X)}^2 = E[v_K^*(X_i)^2]$. It is clear that $\|v_K^*\|_{L^2(X)}^2 = (\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K])'(\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K])$.

Following Chen et al. (2014), we say that f is a regular (or L^2-norm bounded) functional if

$\|v_K^*\|_{L^2(X)} \nearrow \|v^*\|_{L^2(X)} < \infty$ where $v^* \in \mathcal{V}$ is the unique solution to

$$\frac{\partial f(h_0)}{\partial h}[v] = E[v^*(X_i)v(X_i)] \quad \text{for all} \quad v \in \mathcal{V}.$$

We say that f is an irregular (or L^2-norm unbounded) functional if $\|v_K^*\|_{L^2(X)} \nearrow +\infty$. Note that a functional could be irregular but still sup-norm bounded (see Remark 3.1 below).

Given the martingale difference errors (Assumption 2(i)), we can define the sieve variance associated with $f(\widehat{h})$ as $V_K := \|v_K^*\|_{sd}^2 := E[(\epsilon_i v_K^*(X_i))^2]$. It is clear that

$$\begin{aligned} V_K &= \left(\frac{\partial f(h_0)}{\partial h}[b_w^K]\right)' \left(E[b_w^K(X_i)b_w^K(X_i)']\right)^{-1} E[\epsilon_i^2 b_w^K(X_i)b_w^K(X_i)'] \left(E[b_w^K(X_i)b_w^K(X_i)']\right)^{-1} \left(\frac{\partial f(h_0)}{\partial h}[b_w^K]\right) \\ &= \left(\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]\right)' E[\epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'] \left(\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]\right). \end{aligned}$$

The sieve variance $V_K = \|v_K^*\|_{sd}^2$ is estimated with the simple plug-in estimator $\widehat{V}_K = \widehat{\|v_K^*\|}_{sd}^2$, where

$$\begin{aligned} \widehat{\|v_K^*\|}_{sd}^2 &= \frac{1}{n}\sum_{i=1}^n \widehat{v}_K^*(X_i)^2 (Y_i - \widehat{h}(X_i))^2 \\ \widehat{v}_K^*(X_i) &= b_w^K(X_i)'(B_w'B_w/n)^- \tfrac{\partial f(\widehat{h})}{\partial h}[b_w^K]\,. \end{aligned} \tag{26}$$

We first introduce a slight variant of Assumption 2(ii).

Assumption 2 *(iv)* $\inf_{x\in\mathcal{X}} E[\epsilon_i^2|X_i = x] > 0$, *(v)* $\sup_{x\in\mathcal{X}} E[\epsilon_i^2\{|\epsilon_i| > \ell(n)\}|X_i = x] \to 0$ *as* $n \to \infty$ *for any positive sequence* $\ell : \mathbb{N} \to \mathbb{R}_+$ *with* $\ell(n) \to \infty$ *as* $n \to \infty$.

Assumption 2(ii) and (iv) together imply that $\|v_K^*\|_{L^2(X)}^2 \asymp \|v_K^*\|_{sd}^2 = V_K$. Assumption 2(v) is a standard uniform integrability condition, which is not needed for the asymptotic normality of $f(\widehat{h})$ with i.i.d. data when f is a regular functional (see, e.g., Chen (2007))

Before we establish the asymptotic normality of $f(\widehat{h})$ under general weak dependence, we need an additional assumption on the joint dependence of X_i and ϵ_i^2, since this is not captured by the martingale difference property of $\{\epsilon_i\}$ (Assumption 2(i)). Define the $K \times K$ matrices

$$\begin{aligned} \widehat{\Omega} &= n^{-1}\textstyle\sum_{i=1}^n \epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)' \\ \Omega &= E[\epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'] \end{aligned} \tag{27}$$

Assumption 8 $\|\widehat{\Omega} - \Omega\| = o_p(1)$.

The following Lemma is a useful technical result that is again derived using our new exponential inequality for sums of weakly dependent random matrices.

Lemma 3.1 *Let Assumptions 1(i), 2(ii)(iii), and 4(iii) hold. Then Assumption 8 is satisfied provided that either (a), (b) or (c) is satisfied:*

(a) $\{(X_i, Y_i)\}_{i=1}^n$ *is i.i.d. and* $(\zeta_{K,n}\lambda_{K,n})^{(2+\delta)/\delta}\sqrt{(\log K)/n} = o(1)$*;*

(b) $\{(X_i, Y_i)\}_{i=1}^n$ *is exponentially* β*-mixing and* $(\zeta_{K,n}\lambda_{K,n})^{(2+\delta)/\delta}\sqrt{(\log n)^2/n} = o(1)$*;*

(c) $\{(X_i, Y_i)\}_{i=1}^n$ *is algebraically* β*-mixing at rate* γ *and* $(\zeta_{K,n}\lambda_{K,n})^{(2+\delta)/\delta}\sqrt{(\log K)/n^{\gamma/(1+\gamma)}} = o(1)$.

3.1 Asymptotic normality of $f(\widehat{h})$ for general irregular functionals

Let $N_{K,n}$ denote a convex neighborhood of h_0 such that $\widehat{h}, \widetilde{h} \in N_{K,n}$ wpa1. The appropriate neighborhood will depend on the properties of the functional under consideration. For regular and irregular functionals we can typically take $N_{K,n}$ to be of the form $N_{K,n} = \{h \in B_{K,w} : \|h - h_0\|_{L^2(X)} \leq (\sqrt{K/n} + \|h_0 - h_{0,K}\|_{L^2(X)}) \times \log\log n\}$. However, for sup-norm bounded nonlinear functionals (see Remark 3.1) it may suffice to take $N_{K,n} = \{h \in B_{K,w} : \|h - h_0\|_{L^\infty(X)} < \epsilon\}$ for some fixed $\epsilon > 0$, or even $N_{K,n} = L^\infty(X) \cap B_{K,w}$ for sup-norm bounded linear functionals. Our sup-norm and L^2 rate results are clearly useful in defining an appropriate neighborhood.

We now introduce some primitive regularity conditions on the functional f.

Assumption 9 *(i)* $v \mapsto \frac{\partial f(h_0)}{\partial h}[v]$ *is a linear functional;*
(ii) $\sup_{h \in N_{K,n}} \sqrt{n}\|v_K^*\|_{L^2(X)}^{-1} \left| f(h) - f(h_0) - \frac{\partial f(h_0)}{\partial h}[h - h_0] \right| = o(1)$ *where* $\widehat{h}, \widetilde{h} \in N_{K,n}$ *wpa1;*
(iii) $\|v_K^*\|_{L^2(X)} \nearrow +\infty$, $\sqrt{n}\|v_K^*\|_{L^2(X)}^{-1} \left| \frac{\partial f(h_0)}{\partial h}[\widetilde{h} - h_0] \right| = o_p(1)$.

Assumption 9 corresponds to Assumption 3.1 in Chen et al. (2014) and Assumption 2.1 in Chen and Liao (2014) for irregular functionals. We refer the reader to these papers for a detailed discussion and verification of Assumption 9. Note that parts (i) and (ii) of Assumption 9 are automatically satisfied when f is a linear functional.

Remark 3.1 *Certain linear and nonlinear functionals may be irregular yet may still be bounded with respect to the sup norm. Alternative sufficient conditions for Assumption 9 may be provided for such functionals:*

(a) *Suppose f is a linear, irregular functional but that f is sup-norm bounded, i.e. $|f(h)| \lesssim \|h\|_\infty$ (e.g. the evaluation functional $f(h) = h(x)$ for some fixed $x \in \mathcal{X}$ is sup-norm bounded because $|f(h)| = |h(x)| \leq \|h\|_\infty$). Then a sufficient condition for Assumption 9 is*

$$\sqrt{n} V_K^{-1/2} \|\widetilde{h} - h_0\|_\infty \lesssim_p \sqrt{n} V_K^{-1/2} \|P_{K,w,n}\|_\infty \|h_0 - h_{0,K}^*\|_\infty = o_p(1).$$

When $\|P_{K,w,n}\|_\infty \lesssim 1$ and $\|h_0 - h_{0,K}^*\|_\infty = O(K^{-p/d})$ then Assumption 9 is satisfied provided $\sqrt{n} V_K^{-1/2} K^{-p/d} = o(1)$.

(b) *Suppose f is a nonlinear, irregular functional whose derivative is sup-norm bounded. Then Assumption 9 may be replaced with:*
(i') $v \mapsto \frac{\partial f(h_0)}{\partial h}[v]$ is a linear functional;
(ii') $\left| f(h) - f(h_0) - \frac{\partial f(h_0)}{\partial h}[h - h_0] \right| \lesssim \|h - h_0\|_\infty^2$ uniformly for $h \in N_{K,n}$;
(iii') $\left| \frac{\partial f(h_0)}{\partial h}[h - h_0] \right| \lesssim \|h - h_0\|_\infty$ uniformly for $h \in N_{K,n}$; and
(iv') $\widehat{h}, \widetilde{h} \in N_{K,n}$ wpa1, $\sqrt{n}\|v_K^*\|_{L^2(X)}^{-1} \left(\|\widetilde{h} - h_0\|_\infty + \|\widetilde{h} - h_0\|_\infty^2 + \|\widehat{h} - \widetilde{h}\|_\infty^2 \right) = o_p(1)$
where $N_{K,n} = \{h \in B_{K,w} : \|h - h_0\|_\infty \leq \epsilon\}$ for some fixed $\epsilon > 0$.
For example, Newey (1997) shows that conditions (i')(ii')(iii') are satisfied for consumer surplus functionals in demand estimation.

Theorem 3.1 *Let Assumptions 1(i), 2(i)(ii)(iv)(v), 4(iii), 5 and 9 hold. Then*

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{V_K^{1/2}} \to_d N(0,1)$$

as $n, K \to \infty$ provided that either (a) or (b) is satisfied:

(a) *$\{(X_i, Y_i)\}_{i=1}^n$ is i.i.d.;*

(b) *$\{X_i\}_{i=1}^n$ is weakly dependent: Assumption 8 holds, and $\|\widetilde{B}_w' \widetilde{B}_w / n - I_K\| = o_p(K^{-1/2})$.*

We now consider the special case of irregular but sup-norm bounded linear or nonlinear functionals as discussed in Remark 3.1. The sup-norm convergence rates for series LS estimators in Section 2 are employed to derive asymptotic normality of plug-in estimators of such functionals under weak conditions. To save space, for the weakly dependent case we only present sufficient conditions for asymptotic normality of $f(\widehat{h})$ when the regression error has no more than a finite 4th absolute moment (i.e., $E[|\epsilon_i|^{2+\delta}] < \infty$ for some $0 \leq \delta \leq 2$). We also take $\mathcal{X} = [0,1]^d$ and $w_n = 1$ for all n for simplicity.

Corollary 3.1 *Let f be an irregular but sup-norm bounded linear functional, and let Assumptions 1 (with $\mathcal{X} = [0,1]^d$), 2(i)(ii)(iv)(v), 6, and 7 hold. Then*

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{V_K^{1/2}} \to_d N(0,1)$$

as $n, K \to \infty$ provided that either (a), (b) or (c) is satisfied:

(a) $\{(X_i, Y_i)\}_{i=1}^n$ is i.i.d.: $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$ and $(K \log K)/n = o(1)$;

(b) $\{(X_i, Y_i)\}_{i=1}^n$ is exponentially β-mixing: Assumption 2(iii) also holds, $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$, and $K^{(2+\delta)/\delta}(\log n)^2/n = o(1)$ with $\delta \leq 2$;

(c) $\{(X_i, Y_i)\}_{i=1}^n$ is algebraically β-mixing at rate γ: Assumption 2(iii) also holds, $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$, and $K^{(2+\delta)/\delta}(\log K)/n^{\gamma/(1+\gamma)} = o(1)$ with $\delta \leq 2$.

Corollary 3.1 part (a) extends the weakest known result on pointwise asymptotic normality of spline LS estimators in Huang (2003b) to general sup-norm bounded linear functionals of spline or wavelet series LS estimators.[8]

Corollary 3.2 *Let f be an irregular but sup-norm bounded nonlinear functional, and let Assumptions 1 (with $\mathcal{X} = [0,1]^d$), 2, 6, 7 and 9(i')(ii')(iii') hold. Then*

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{V_K^{1/2}} \to_d N(0,1)$$

as $n, K \to \infty$ provided that either (a), (b) or (c) is satisfied:

(a) $\{(X_i, Y_i)\}_{i=1}^n$ is i.i.d.: $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$ and $K^{(2+\delta)/\delta}(\log n)/n \lesssim 1$ with $\delta < 2$;

(b) $\{(X_i, Y_i)\}_{i=1}^n$ is exponentially β-mixing: $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$ and $K^{(2+\delta)/\delta}(\log n)^2/n = o(1)$ with $\delta \leq 2$;

[8] Under the assumption of empirical identifiability (see equation (30)) and other conditions similar to the ones listed in Corollary 3.1 part (a), Chen and Huang (2003) derived the asymptotic normality of plug-in spline LS estimators of sup-norm bounded linear functionals (see their Theorem 4).

(c) $\{(X_i, Y_i)\}_{i=1}^n$ *is algebraically* β*-mixing at rate* γ*:* $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$ *and* $K^{(2+\delta)/\delta}(\log K)/n^{\gamma/(1+\gamma)} = o(1)$ *with* $\delta \leq 2$*.*

Conditions for weakly dependent data in Corollary 3.2 parts (b) and (c) are natural extensions of those in part (a) for i.i.d. data, which in turn are much weaker than the well-known conditions in Newey (1997) for the asymptotic normality of nonlinear functionals of spline LS estimators, namely $\sqrt{n}K^{-p/d} = o(1)$, $K^4/n = o(1)$ and $\sup_x E[|\epsilon_i|^4 \, |X_i = x] < \infty$.

3.2 Asymptotic normality of sieve t statistics for general functionals

We now turn to the consistent estimation of $V_K = \|v_K^*\|_{sd}^2$ and feasible asymptotic inference for $f(h_0)$.

Assumption 10 $\|v_K^*\|_{L^2(X)}^{-1} \left\| \frac{\partial f(h)}{\partial h}[\widetilde{b}_w^K] - \frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K] \right\| = o(1)$ *uniformly over* $h \in N_{K,n}$ *or* $B_{\epsilon,\infty}(h_0)$*.*

Note that Assumption 10 is automatically satisfied when f is a linear functional. It is only required to establish consistency of $\widehat{\|v_K^*\|}_{sd}$ for a nonlinear functional, and corresponds to Assumption 3.1(iii) of Chen and Liao (2014).

The first part of the following Lemma establishes the consistency of the sieve variance estimator under both the i.i.d. and general weakly dependent data.

Lemma 3.2 *Let Assumptions 1(i)(ii), 2(i)(ii)(iv), 4(iii), 5, 8 and 10 hold and* $\|\widehat{h} - h_0\|_{\infty,w} = o_p(1)$*. Then:*

(1) $\left| \frac{\widehat{\|v_K^*\|}_{sd}}{\|v_K^*\|_{sd}} - 1 \right| = o_p(1)$ *as* $n, K \to \infty$*.*

(2) Further, if Assumptions 2(v) and 9 hold, then:

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{\widehat{V}_K^{1/2}} \to_d N(0,1)$$

as $n, K \to \infty$ *provided that either (a)* $\{(X_i, Y_i)\}_{i=1}^n$ *is i.i.d., or (b)* $\{X_i\}_{i=1}^n$ *is weakly dependent with* $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = o_p(K^{-1/2})$ *is satisfied.*

Lemma 3.2 can be combined with different sufficient conditions for Assumptions 5, 8 and 9 to yield different special cases of the asymptotic normality of sieve t statistics for general (possibly) nonlinear functionals. We state three special cases below. The following Theorem is applicable to series LS estimators with an arbitrary basis.

Theorem 3.2 *Let Assumptions 1(i)(ii), 2, 3, 4, 9 and 10 hold and* $\|\widetilde{h} - h_0\|_\infty = o_p(1)$*. Then*

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{\widehat{V}_K^{1/2}} \to_d N(0,1)$$

as $n, K \to \infty$ *provided that either (a), (b) or (c) is satisfied:*

- *(a)* $\{(X_i, Y_i)\}_{i=1}^n$ *is i.i.d.:* $(\zeta_{K,n}\lambda_{K,n})^{(2+\delta)/\delta}\sqrt{(\log n)/n} = o(1)$*;*

- *(b)* $\{(X_i, Y_i)\}_{i=1}^n$ *is exponentially* β*-mixing:* $\max(\sqrt{K}, (\zeta_{K,n}\lambda_{K,n})^{2/\delta}) \times (\zeta_{K,n}\lambda_{K,n})\sqrt{\frac{(\log n)^2}{n}} = o(1)$*;*

- *(c)* $\{(X_i, Y_i)\}_{i=1}^n$ *is algebraically* β*-mixing at rate* γ*:* $\max(\sqrt{K}, (\zeta_{K,n}\lambda_{K,n})^{2/\delta}) \times (\zeta_{K,n}\lambda_{K,n})\sqrt{\frac{\log n}{n^{\gamma/(1+\gamma)}}} = o(1)$*.*

The following Corollaries are direct consequences of Theorem 3.2 for linear and nonlinear sup-norm bounded functionals, with spline or wavelet bases. For simplicity, we take $w_n = 1$ for all n and $\mathcal{X} = [0,1]^d$.

Corollary 3.3 *Let Assumptions 1 (with* $\mathcal{X} = [0,1]^d$*), 2, 6, and 7 hold for a sup-norm bounded linear functional. Then*

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{\widehat{V}_K^{1/2}} \to_d N(0,1)$$

as $n, K \to \infty$ *provided that either (a), (b) or (c) is satisfied:*

- *(a)* $\{(X_i, Y_i)\}_{i=1}^n$ *is i.i.d.:* $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$ *and* $K^{(2+\delta)/\delta}(\log n)/n = o(1)$*;*

- *(b) part (b) of Corollary 3.1;*

- *(c) part (c) of Corollary 3.1.*

Corollary 3.4 *Let Assumptions 1 (with* $\mathcal{X} = [0,1]^d$*), 2, 6, 7, 9(i')(ii')(iii') and 10 hold for a nonlinear functional. Then*

$$\frac{\sqrt{n}(f(\widehat{h}) - f(h_0))}{\widehat{V}_K^{1/2}} \to_d N(0,1)$$

as $n, K \to \infty$ *provided that either (a), (b) or (c) is satisfied:*

- *(a)* $\{(X_i, Y_i)\}_{i=1}^n$ *is i.i.d.:* $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$ *and* $K^{(2+\delta)/\delta}(\log n)/n = o(1)$ *with* $\delta < 2$*;*

- *(b) part (b) of Corollary 3.2;*

(c) part (c) of Corollary 3.2.

Previously, Newey (1997) required that $\sup_x E[\epsilon_i^4|X_i = x] < \infty$ and $K^4/n = o(1)$ in order to establish asymptotic normality of student t statistics for nonlinear functionals with i.i.d. data. Our sufficient conditions are weaker and allow for weakly dependent data with heavy-tailed errors.

4 Useful results on random matrices

4.1 An exponential inequality for sums of weakly dependent random matrices

In this section we derive a new Bernstein-type inequality for sums of random matrices formed from absolutely regular (β-mixing) sequences, where the dimension, norm, and variance measure of the random matrices are allowed to grow with the sample size. This inequality is particularly useful for establishing sharp convergence rates for semi/nonparametric sieve estimators with weakly dependent data. We first recall an inequality of Tropp (2012) for independent random matrices.

Theorem 4.1 (Tropp (2012)) *Let $\{\Xi_i\}_{i=1}^n$ be a finite sequence of independent random matrices with dimensions $d_1 \times d_2$. Assume $E[\Xi_i] = 0$ for each i and $\max_{1\leq i\leq n} \|\Xi_i\| \leq R_n$, and define*

$$\sigma_n^2 = \max\left\{ \left\| \sum_{i=1}^n E[\Xi_i \Xi_i'] \right\|, \left\| \sum_{i=1}^n E[\Xi_i' \Xi_i] \right\| \right\}.$$

Then for all $t \geq 0$,

$$\mathbb{P}\left(\left\| \sum_{i=1}^n \Xi_i \right\| \geq t \right) \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{\sigma_n^2 + R_n t/3} \right).$$

Corollary 4.1 *Under the conditions of Theorem 4.1, if $R_n\sqrt{\log(d_1 + d_2)} = o(\sigma_n)$ then*

$$\left\| \sum_{i=1}^n \Xi_{i,n} \right\| = O_p(\sigma_n \sqrt{\log(d_1 + d_2)}).$$

When $\{X_i\}_{i=-\infty}^{\infty}$ is i.i.d., Corollary 4.1 is used to provide weak low-level sufficient conditions under which $\|\widetilde{B}_w' \widetilde{B}_w/n - I_K\| = o_p(1)$ holds (see Lemma 2.1).

We now provide an extension of Theorem 4.1 and Corollary 4.1 for matrix-valued functions of β-mixing sequences. The β-mixing coefficient between two σ-algebras $\mathcal{A}$ and $\mathcal{B}$ is defined as

$$\beta(\mathcal{A}, \mathcal{B}) = \frac{1}{2} \sup \sum_{(i,j) \in I \times J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| \tag{28}$$

with the supremum taken over all finite partitions $\{A_i\}_{i\in I} \subset \mathcal{A}$ and $\{B_j\}_{j\in J} \subset \mathcal{B}$ of Ω (see, e.g., Bradley (2005)). The qth β-mixing coefficient of $\{X_i\}_{i=-\infty}^{\infty}$ is defined as

$$\beta(q) = \sup_i \beta(\sigma(\ldots, X_{i-1}, X_i), \sigma(X_{i+q}, X_{i+q+1}, \ldots)) \,. \tag{29}$$

The process $\{X_i\}_{i=-\infty}^{\infty}$ is said to be *algebraically β-mixing* at rate γ if $q^\gamma \beta(q) = o(1)$ for some $\gamma > 1$, and *exponentially β-mixing* if $\beta(q) \leq c\exp(-\gamma q)$ for some $\gamma > 0$ and $c \geq 0$. The following extension of Theorem 4.1 is made using Berbee's Lemma and a coupling argument.

Theorem 4.2 *Let $\{X_i\}_{i=-\infty}^{\infty}$ be a β-mixing sequence and let $\Xi_{i,n} = \Xi_n(X_i)$ for each i where $\Xi_n : \mathcal{X} \to \mathbb{R}^{d_1 \times d_2}$ is a sequence of measurable $d_1 \times d_2$ matrix-valued functions. Assume $E[\Xi_{i,n}] = 0$ and $\|\Xi_{i,n}\| \leq R_n$ for each i and define $s_n^2 = \max_{1\leq i,j\leq n} \max\{\|E[\Xi_{i,n}\Xi'_{j,n}]\|, \|E[\Xi'_{i,n}\Xi_{j,n}]\|\}$. Let q be an integer between 1 and $n/2$ and let $I_r = q[n/q]+1, \ldots, n$ when $q[n/q] < n$ and $I_r = \emptyset$ when $q[n/q] = n$. Then for all $t \geq 0$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geq 6t\right) \leq \frac{n}{q}\beta(q) + \mathbb{P}\left(\left\|\sum_{i\in I_r} \Xi_{i,n}\right\| \geq t\right) + 2(d_1+d_2)\exp\left(\frac{-t^2/2}{nqs_n^2 + qR_n t/3}\right)$$

(where $\|\sum_{i\in I_r} \Xi_{i,n}\| := 0$ whenever $I_r = \emptyset$).

Corollary 4.2 *Under the conditions of Theorem 4.2, if $q = q(n)$ is chosen s.t. $\frac{n}{q}\beta(q) = o(1)$ and $R_n\sqrt{q\log(d_1+d_2)} = o(s_n\sqrt{n})$ then*

$$\left\|\sum_{i=1}^n \Xi_{i,n}\right\| = O_p(s_n\sqrt{nq\log(d_1+d_2)}) \,.$$

When the regressors $\{X_i\}_{i=-\infty}^{\infty}$ are β-mixing, Corollary 4.2 is used to provide weak low-level sufficient conditions under which $\|\widetilde{B}'_w\widetilde{B}_w/n - I_K\| = o_p(1)$ holds (see Lemma 2.2).

We note that both Theorem 4.2 and Corollary 4.2 allow for non-identically distributed beta-mixing sequences. So the convergence rate and the inference results in previous sections could be extended to non-identically distributed regressors $\{X_i\}_{i=-\infty}^{\infty}$ as well, except that notation and regularity conditions will be slightly more complicated.

4.2 Empirical identifiability

We now provide a readily verifiable condition under which the theoretical and empirical L^2 norms are equivalent over a (weighted) linear sieve space wpa1. This equivalence, referred to by Huang (2003b) as *empirical identifiability*, has several applications in nonparametric sieve estimation. In nonparametric series LS estimation, empirical identifiability ensures that the estimator is the orthogonal projection of Y onto the linear sieve space under the empirical inner product and is uniquely defined wpa1 (Huang, 2003b). Empirical identifiability is also used to establish the large-sample properties of sieve conditional moment estimators (see, e.g., Chen and Pouzo (2012)). A sufficient condition for empirical identifiability is now cast in terms of convergence of a random matrix, which we verify for i.i.d. and β-mixing sequences.

Recall that $L^2(X)$ denotes the space of functions $f : \mathcal{X} \to \mathbb{R}$ such that $E[f(X_i)^2] < \infty$. A (linear) subspace $\mathcal{A} \subseteq L^2(X)$ is said to be *empirically identifiable* if $\frac{1}{n}\sum_{i=1}^n b(X_i)^2 = 0$ implies $b = 0$. A sequence of spaces $\{\mathcal{A}_K : K \geq 1\} \subseteq L^2(X)$ is empirically identifiable wpa1 as $K = K(n) \to \infty$ with n if

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{a \in A_K} \left| \frac{\frac{1}{n}\sum_{i=1}^n a(X_i)^2 - E[a(X_i)^2]}{E[a(X_i)^2]} \right| > t \right) = 0 \tag{30}$$

for any $t > 0$. Huang (1998) verifies (30) for i.i.d. data using a chaining argument. Chen and Pouzo (2012) use this result to establish convergence of sieve conditional moment estimators. However, it may be difficult to verify (30) via chaining arguments for certain types of weakly dependent sequences.

To this end, the following is a readily verifiable sufficient condition for empirical identifiability for (weighted) linear sieve spaces given by $B_{K,w} = clsp\{b_{K1}w_n, \ldots, b_{KK}w_n\}$.

Condition 4.1 *$\lambda_{\min}(E[b_w^K(X_i)b_w^K(X_i)']) > 0$ for each K and $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = o_p(1)$.*

Lemma 4.1 *If $\lambda_{\min}(E[b_w^K(X_i)b_w^K(X_i)']) > 0$ for each K then*

$$\sup_{b \in B_{K,w}} \left| \frac{\frac{1}{n}\sum_{i=1}^n b(X_i)^2 - E[b(X_i)^2]}{E[b(X_i)^2]} \right| = \|\widetilde{B}_w'\widetilde{B}_w/n - I_K\|^2 \,.$$

Corollary 4.3 *If Condition 4.1 holds then $B_{K,w}$ is empirically identifiable wpa1.*

Condition 4.1 is therefore a sufficient condition for (30) to hold for the linear sieve space $B_{K,w}$.

Remark 4.1 *Consider the compact support case in which $\mathcal{X} = [0,1]^d$ and $w_n(x) = 1$ for all $x \in \mathcal{X}$ and all n (so that $b_{Kk}(x)w_n(x) = b_{Kk}(x)$ for all n and K) and suppose the density of X_i is* *uniformly bounded away from zero and infinity over $\mathcal{X}$. **(1)** For i.i.d. regressors (and $\lambda_{K,n} \lesssim 1$), previously Huang (1998) establishes equivalence of the theoretical and empirical L^2 norms over the sieve space via a chaining argument with $\zeta_{K,n}^2 K/n = o(1)$. Huang (2003b) relaxes this to $K(\log n)/n = o(1)$ for a polynomial spline basis. Our Lemma 2.1 shows that, in fact, $\zeta_{K,n}\sqrt{(\log K)/n} = o(1)$ is sufficient with an arbitrary linear sieve (provided $\lambda_{K,n} \lesssim 1$). **(2)** For strictly stationary beta-mixing regressors (and $\lambda_{K,n} \lesssim 1$), Lemma 2.2 shows the equivalence of the theoretical and empirical L^2 norms over any linear sieve space under either $\zeta_{K,n}\sqrt{(\log n)^2/n} = o(1)$ for exponential beta-mixing, or $\zeta_{K,n}\sqrt{(\log K)/n^{\gamma/(1+\gamma)}} = o(1)$ for algebraic beta-mixing.*

5 Sup-norm stability of $L^2(X)$ projection onto wavelet sieves

In this section we show that the $L^2(X)$ orthogonal projection onto (tensor product) compactly supported wavelet bases is stable in sup norm as the dimension of the space increases. Consider the orthogonal projection operator P_K defined in expression (16) where the elements of b^K span the tensor products of d univariate wavelet spaces $\text{Wav}(K_0, [0,1])$. We show that its L^∞ operator norm $\|P_K\|_\infty$ (see expression (17)) is stable, in the sense that $\|P_K\|_\infty \lesssim 1$ as $K \to \infty$. We also show that the empirical L^2 projection $P_{K,n}$ onto the wavelet sieve is stable in sup norm wpa1. This result is used to establish that series LS estimators with (tensor-product) wavelet bases attain their optimal sup-norm rates. A variant of this result for projections arising in series two-stage LS was used in an antecedent of this paper (Chen and Christensen, 2013) but its proof was omitted for brevity.

The following Theorem presents our result for the stability of the projection with respect to the $L^2(X)$ inner product.

Theorem 5.1 *Let $\mathcal{X} \supseteq [0,1]^d$ and let the density f_X of X_i be such that $0 < \inf_{x\in[0,1]^d} f_X(x) \leq \sup_{x\in[0,1]^d} f_X(x) < \infty$. Let B_K be the tensor product of d univariate wavelet spaces* $\text{Wav}(K_0, [0,1])$ *where* $\text{Wav}(K_0, [0,1])$ *is as described in Section 6 and $K = 2^{dJ}$ and $K_0 = 2^J > 2N$. Then: $\|P_K\|_\infty \lesssim 1$.*

We now present conditions under which the empirical projection onto a tensor-product wavelet basis is stable wpa1. Here the projection operator is

$$P_{K,n}h(x) = b^K(x)' \left(\frac{B'B}{n}\right)^- \frac{1}{n}\sum_{i=1}^n b^K(X_i)h(X_i)$$

where the elements of b^K span the tensor products of d univariate spaces $\mathrm{Wav}(K_0, [0,1])$. The following Theorem states simple sufficient conditions for $\|P_{K,n}\|_\infty \lesssim 1$ wpa1.

Theorem 5.2 *Let conditions stated in Theorem 5.1 hold. Then $\|P_{K,n}\|_\infty \lesssim 1$ wpa1 provided that either (a), (b), or (c) is satisfied:*

(a) $\{X_i\}_{i=1}^n$ are i.i.d. and $\sqrt{(K \log n)/n} = o(1)$

(b) $\{X_i\}_{i=1}^n$ are exponentially β-mixing and $\sqrt{K(\log n)^2/n} = o(1)$, or

(c) $\{X_i\}_{i=1}^n$ are algebraically β-mixing at rate γ and $\sqrt{(K \log n)/n^{\gamma/(1+\gamma)}} = o(1)$.

6 Brief review of B-spline and wavelet sieve spaces

We first outline univariate B-spline and wavelet sieve spaces on $[0,1]$, then deal with the multivariate case by constructing a tensor-product sieve basis.

B-splines B-splines are defined by their order $r \geq 1$ (or degree $r-1 \geq 0$) and number of interior knots $m \geq 0$. Define the knot set

$$0 = t_{-(r-1)} = \ldots = t_0 \leq t_1 \leq \ldots \leq t_m \leq t_{m+1} = \ldots = t_{m+r} = 1\,. \tag{31}$$

We generate a L^∞-normalized B-spline basis recursively using the De Boor relation (see, e.g., Chapter 5 of DeVore and Lorentz (1993)) then appropriately rescale the basis functions. Define the interior intervals $I_1 = [t_0, t_1), \ldots, I_m = [t_m, t_{m+1}]$ and generate a basis of order 1 by setting

$$N_{j,1}(x) = 1_{I_j}(x) \tag{32}$$

for $j = 0, \ldots m$, where $1_{I_j}(x) = 1$ if $x \in I_j$ and $1_{I_j}(x) = 0$ otherwise. Bases of order $r > 1$ are generated recursively according to

$$N_{j,r}(x) = \frac{x - t_j}{t_{j+r-1} - t_j} N_{j,r-1}(x) + \frac{t_{j+r} - x}{t_{j+r} - t_{j+1}} N_{j+1,r-1}(x) \tag{33}$$

for $j = -(r-1), \ldots, m$ where we adopt the convention $\frac{1}{0} := 0$. Finally, we rescale the basis by multiplying each $N_{j,r}$ by $(m+r)^{1/2}$ for $j = -(r-1), \ldots, m$. This results in a total of $K = m + r$

splines of order r. Each spline is a polynomial of degree $r-1$ on each interior interval $I_1, \dots, I_m$ and is $(r-2)$-times continuously differentiable on $(0,1)$ whenever $r > 2$. The mesh ratio is defined as

$$\operatorname{mesh}(K) = \frac{\max_{0 \le j \le m}(t_{j+1} - t_j)}{\min_{0 \le j \le m}(t_{j+1} - t_j)} . \tag{34}$$

We let the space $\mathrm{BSpl}(K, [0,1])$ be the closed linear span of these $K = m + r$ splines. The space $\mathrm{BSpl}(K, [0,1])$ has uniformly bounded mesh ratio if $\operatorname{mesh}(K) \le \kappa$ for all $N \ge 0$ and some $\kappa \in (0, \infty)$. We let $\mathrm{BSpl}(K, [0,1], \gamma)$ denote the space $\mathrm{BSpl}(K, [0,1])$ with degree γ and uniformly bounded mesh ratio. See De Boor (2001) and Schumaker (2007) for further details.

Wavelets We construct a wavelet basis with support $[0,1]$ following Cohen, Daubechies, and Vial (1993). Let (φ, ψ) be a Daubechies pair such that φ has support $[-N+1, N]$. Given j such that $2^j - 2N > 0$, the orthonormal (with respect to the $L^2([0,1])$ inner product) basis for the space V_j consists of $2^j - 2N$ interior scaling functions of the form $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$, each of which has support $[2^{-j}(-N+1+k), 2^{-j}(N+k)]$ for $k = N, \dots, 2^j - N - 1$. These are augmented with N left scaling functions of the form $\varphi^0_{j,k}(x) = 2^{j/2}\varphi^l_k(2^j x)$ for $k = 0, \dots, N-1$ (where $\varphi^l_0, \dots, \varphi^l_{N-1}$ are fixed independent of j), each of which has support $[0, 2^{-j}(N+k)]$, and N right scaling functions of the form $\varphi_{j,2^j-k}(x) = 2^{j/2}\varphi^r_{-k}(2^j(x-1))$ for $k = 1, \dots, N$ (where $\varphi^r_{-1}, \dots, \varphi^r_{-N}$ are fixed independent of j), each of which has support $[1 - 2^{-j}(1-N-k), 1]$. The resulting 2^j functions $\varphi^0_{j,0}, \dots, \varphi^0_{j,N-1}, \varphi_{j,N}, \dots, \varphi_{j,2^j-N-1}, \varphi^1_{j,2^j-N}, \dots, \varphi^1_{j,2^j-1}$ form an orthonormal basis (with respect to the $L^2([0,1])$ inner product) for the subspace they span, denoted V_j.

An orthonormal wavelet basis for the space W_j, defined as the orthogonal complement of V_j in V_{j+1}, is similarly constructed form the mother wavelet. This results in an orthonormal basis of 2^j functions $\psi^0_{j,0}, \dots, \psi^0_{j,N-1}, \psi_{j,N}, \dots, \psi_{j,2^j-N-1}, \psi^1_{j,2^j-N}, \dots, \psi^1_{j,2^j-1}$. To simplify notation we ignore the 0 and 1 superscripts on the left and right wavelets and scaling functions henceforth.

Let J_0 and J be integers such that $2^{J_0} \le 2^J < 2N$. A wavelet space at resolution level J is the set of 2^J functions given by

$$\mathrm{Wav}(J) = \left\{ \sum_{k=0}^{2^{J_0}-1} a_{J_0,k}\varphi_{J_0,k} + \sum_{j=J_0}^{J} \sum_{k=0}^{2^j-1} b_{j,k}\psi_{j,k} : a_{J_0,k}, b_{j,k} \in \mathbb{R} \right\} . \tag{35}$$

The spaces V_j and W_j are constructed so that $V_{j+1} = V_j \oplus W_j$ for all j with $2^j - 2N > 0$. Therefore, we can reexpress $\mathrm{Wav}(J)$ as

$$\mathrm{Wav}(J) = \left\{ \sum_{k=0}^{2^J-1} a_{J,k}\varphi_{J,k} : a_{J,k} \in \mathbb{R} \right\} . \tag{36}$$

The orthogonal projection onto $\mathrm{Wav}(J)$ is therefore the same, irrespective of whether we use the bases for V_J or $V_{J_0} \oplus W_{J_0} \oplus \ldots \oplus W_J$. Note that, by the support of the $\varphi_{J,0}, \ldots, \varphi_{J,2^J-1}$, the support of at most $2N-1$ basis functions overlaps on a set of positive Lebesgue measure. We use this local support to bound the orthogonal projection operator onto (tensor product) wavelet bases below.

We say that $\mathrm{Wav}(K, [0,1])$ has *regularity* γ if $N \geq \gamma$, and write $\mathrm{Wav}(K, [0,1], \gamma)$ for a wavelet space of regularity γ with continuously differentiable basis functions. See Johnstone (2013) for further details.

Tensor products We construct tensor product B-spline or wavelet bases for $[0,1]^d$ as follows. First, for $x = (x_1, \ldots, x_d) \in [0,1]^d$ we construct d B-spline or wavelet bases for $[0,1]$. We then form the tensor product basis by taking the product of the elements of each of the univariate bases. Therefore, $b^K(x)$ may be expressed as

$$b^K(x) = \bigotimes_{l=1}^{d} b^{K_0}(x_l) \tag{37}$$

where the elements of each vector $b^{K_0}(x_l)$ span $\mathrm{BSpl}(K_0, [0,1], \gamma)$ with $K_0 = m + r$ for $l = 1, \ldots, d$, or span $\mathrm{Wav}(K_0, [0,1], \gamma)$ with $K_0 = 2^J$ for $l = 1, \ldots, d$. We let $\mathrm{BSpl}(K, [0,1]^d, \gamma)$ and $\mathrm{Wav}(K, [0,1]^d, \gamma)$ denote the resulting tensor-product spaces spanned by the $K = (m+r)^d$ or $K = 2^{dJ}$ elements of b^K.

7 Proofs

7.1 Proofs for Section 2

Proof of Lemma 2.1. Follows from Corollary 4.1 by setting $\Xi_{i,n} = n^{-1}(\widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)' - I_K)$ and noting that $R_n \leq n^{-1}(\zeta_{K,n}^2\lambda_{K,n}^2 + 1)$, and $\sigma_n^2 \leq n^{-1}(\zeta_{K,n}^2\lambda_{K,n}^2 + 1)$. ■

Proof of Lemma 2.2. Follows from Corollary 4.2 by setting $\Xi_{i,n} = n^{-1}(\widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)' - I_K)$ and noting that $R_n \leq n^{-1}(\zeta_{K,n}^2\lambda_{K,n}^2 + 1)$, and $\sigma_n^2 \leq n^{-2}(\zeta_{K,n}^2\lambda_{K,n}^2 + 1)$. ■

Proof of Lemma 2.3. By rotational invariance, we may rescale $\widehat{h}$ and $\widetilde{h}$ to yield

$$\widehat{h}(x) - \widetilde{h}(x) = \widetilde{b}_w^K(x)'(\widetilde{B}_w'\widetilde{B}_w/n)^-\widetilde{B}_w'e/n \tag{38}$$

where $e = (\epsilon_1, \ldots, \epsilon_n)'$.

Let $\breve{h} = \widehat{h} - \widetilde{h}$ to simplify notation. By the mean value theorem, Assumptions 3(i) and 4(i)(iii), for any $(x, x^*) \in \mathcal{D}_n^2$ we have

$$\begin{aligned} |\breve{h}(x) - \breve{h}(x^*)| &= |(\widetilde{b}_w^K(x) - \widetilde{b}_w^K(x^*))'(\widetilde{B}_w'\widetilde{B}_w/n)^-\widetilde{B}_w'e/n| && (39)\\ &= |(x - x^*)'\nabla\widetilde{b}_w^K(x^{**})'(\widetilde{B}_w'\widetilde{B}_w/n)^-\widetilde{B}_w'e/n| && (40)\\ &\leq C_\nabla \lambda_{K,n} n^{\omega_1} K^{\omega_2}\|x - x^*\|\|(\widetilde{B}_w'\widetilde{B}_w/n)^-\|\|\widetilde{B}_w'e/n\| && (41) \end{aligned}$$

for some x^{**} in the segment between x and x^* and some finite constant C_∇ (independent of x, x^*, n, K). Now, $\|(\widetilde{B}_w'\widetilde{B}_w/n)^{-1}\| = O_p(1)$ by Assumption 5, and we may deduce by Markov's inequality (under Assumptions 2(i)(ii)) that $\|\widetilde{B}_w'e/n\| = O_p(\sqrt{K/n})$. It follows that

$$\limsup_{n\to\infty} \mathbb{P}\left(C_\nabla\|(\widetilde{B}_w'\widetilde{B}_w/n)^-\|\|\widetilde{B}_w'e/n\| > \bar{M}\right) = 0 \tag{42}$$

for any fixed $\bar{M} > 0$ (since condition (i) implies $K/n = o(1)$). Let $\mathcal{B}_n$ denote the event on which $C_\nabla\|(\widetilde{B}_w'\widetilde{B}_w/n)^-\|\|\widetilde{B}_w'e/n\| \leq \bar{M}$ and observe that $\mathbb{P}(\mathcal{B}_n^c) = o(1)$. On $\mathcal{B}_n$, for any $C \geq 1$, a finite positive $\eta_1 = \eta_1(C)$ and $\eta_2 = \eta_2(C)$ can be chosen such that

$$C_\nabla \lambda_{K,n} n^{\omega_1} K^{\omega_2}\|x - x^*\|\|(\widetilde{B}_w'\widetilde{B}_w/n)^-\|\|\widetilde{B}_w'e/n\| \leq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n} \tag{43}$$

whenever $\|x - x^*\| \leq \eta_1 n^{-\eta_2}$, by Assumption 4(ii). Let $\mathcal{S}_n$ be the smallest subset of $\mathcal{D}_n$ such that for each $x \in \mathcal{D}_n$ there exists a $x_n \in \mathcal{S}_n$ with $\|x_n - x\| \leq \eta_1 n^{-\eta_2}$. For any $x \in \mathcal{D}_n$ let $x_n(x)$ denote the $x_n \in \mathcal{S}_n$ nearest (in Euclidean distance) to x. Then on $\mathcal{B}_n$ we have

$$|\breve{h}(x) - \breve{h}(x_n(x))| \leq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n} \tag{44}$$

for any $x \in \mathcal{D}_n$.

Using the fact that $\mathbb{P}(A) \leq \mathbb{P}(A \cap B) + \mathbb{P}(B^c)$, we obtain

$$
\begin{aligned}
& \mathbb{P}\left(\|\check{h}\|_\infty \geq 4C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right) \\
\leq\ & \mathbb{P}\left(\left\{\|\check{h}\|_\infty \geq 4C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{B}_n\right) + \mathbb{P}(\mathcal{B}_n^c) && (45) \\
\leq\ & \mathbb{P}\left(\left\{\sup_{x \in \mathcal{X}} |\check{h}(x) - \check{h}(x_n(x))| \geq 2C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{B}_n\right) \\
& + \mathbb{P}\left(\left\{\max_{x_n \in \mathcal{S}_n} |\check{h}(x_n)| \geq 2C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{B}_n\right) + \mathbb{P}(\mathcal{B}_n^c) && (46) \\
=\ & \mathbb{P}\left(\left\{\max_{x_n \in \mathcal{S}_n} |\check{h}(x_n)| \geq 2C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{B}_n\right) + o(1) && (47)
\end{aligned}
$$

where the final line is by (44) and the fact that $\mathbb{P}(\mathcal{B}_n^c) = o(1)$. The arguments used to control expression (47) differ depending upon whether or not $\{(X_i, Y_i)\}_{i=1}^n$ is i.i.d.

With **i.i.d. data**, first let $\mathcal{A}_n$ denote the event on which $\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| \leq \frac{1}{2}$ and observe that $\mathbb{P}(\mathcal{A}_n^c) = o(1)$ because $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = o_p(1)$. Let $\{\mathcal{A}_n\}$ denote the indicator function of $\mathcal{A}_n$, let $\{M_n : n \geq 1\}$ be an increasing sequence diverging to $+\infty$, and define

$$
\begin{aligned}
\epsilon_{1,i,n} &:= \epsilon_i\{|\epsilon_i| \leq M_n\} - E[\epsilon_i\{|\epsilon_i| \leq M_n\}|X_i] && (48) \\
\epsilon_{2,i,n} &:= \epsilon_i - \epsilon_{1,i,n} && (49) \\
G_{i,n}(x_n) &:= \tilde{b}_w^K(x_n)'(\widetilde{B}_w'\widetilde{B}_w/n)^{-}\tilde{b}_w^K(X_i)\{\mathcal{A}_n\}\,. && (50)
\end{aligned}
$$

Since $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ and $\mathbb{P}(A) \leq \mathbb{P}(A \cap B) + \mathbb{P}(B^c)$, we have

$$
\begin{aligned}
& \mathbb{P}\left(\left\{\max_{x_n \in \mathcal{S}_n} |\check{h}(x_n)| \geq 2C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{B}_n\right) \\
\leq\ & \mathbb{P}\left(\max_{x_n \in \mathcal{S}_n} |\check{h}(x_n)| \geq 2C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right) \\
\leq\ & \mathbb{P}\left(\left\{\max_{x_n \in \mathcal{S}_n} |\check{h}(x_n)| \geq 2C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{A}_n\right) + \mathbb{P}(\mathcal{A}_n^c) \\
\leq\ & (\#\mathcal{S}_n) \max_{x_n \in \mathcal{S}_n} \mathbb{P}\left(\left\{\left|\frac{1}{n}\sum_{i=1}^n G_{i,n}(x_n)\epsilon_{1,i,n}\right| \geq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{A}_n\right) && (51a) \\
& + \mathbb{P}\left(\left\{\max_{x_n \in \mathcal{S}_n} \left|\frac{1}{n}\sum_{i=1}^n G_{i,n}(x_n)\epsilon_{2,i,n}\right| \geq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\} \cap \mathcal{A}_n\right) + \mathbb{P}(\mathcal{A}_n^c)\,. && (51b)
\end{aligned}
$$

Control of (51a): Note that $\|(\widetilde{B}_w'\widetilde{B}_w/n)^{-1}\| \leq 2$ on $\mathcal{A}_n$. Therefore, by the Cauchy-Schwarz inequality and definition of $\epsilon_{1,i,n}$ and $\zeta_{K,n}$, $\lambda_{K,n}$, we have:

$$|n^{-1}G_{i,n}(x_n)\epsilon_{1,i,n}| \lesssim \frac{\zeta_{K,n}^2\lambda_{K,n}^2 M_n}{n}\,. \tag{52}$$

Let $E[\,\cdot\,|X_1^n]$ denote expectation conditional on $X_1,\ldots,X_n$. Assumption 2(ii) in the i.i.d. data case implies that $\sup_x E[\epsilon_i^2|X_i = x] < \infty$. Therefore,

$$\begin{aligned}
& \sum_{i=1}^n E[(n^{-1}G_{i,n}(x_n)\epsilon_{1,i,n})^2|X_1^n] \\
= \;& \frac{1}{n^2}\sum_{i=1}^n E[\epsilon_{1,i,n}^2|X_i]\widetilde{b}_w^K(x_n)'(\widetilde{B}_w'\widetilde{B}_w/n)^-\widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\{\mathcal{A}_n\}(\widetilde{B}_w'\widetilde{B}_w/n)^-\widetilde{b}_w^K(x_n) && (53)\\
\lesssim \;& \frac{1}{n^2}\sum_{i=1}^n \widetilde{b}_w^K(x_n)'(\widetilde{B}_w'\widetilde{B}_w/n)^-\widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\{\mathcal{A}_n\}(\widetilde{B}_w'\widetilde{B}_w/n)^-\widetilde{b}_w^K(x_n) && (54)\\
= \;& \frac{1}{n}\widetilde{b}_w^K(x_n)'E[(\widetilde{B}_w'\widetilde{B}_w/n)^-(\widetilde{B}_w'\widetilde{B}_w/n)\{\mathcal{A}_n\}(\widetilde{B}_w'\widetilde{B}_w/n)^-]\widetilde{b}_w^K(x_n) \;\lesssim\; \frac{\zeta_{K,n}^2\lambda_{K,n}^2}{n}\,. && (55)
\end{aligned}$$

Bernstein's inequality for independent random variables (see, e.g., pp. 192–193 of Pollard (1984)) then provides that

$$\begin{aligned}
& (\#\mathcal{S}_n)\max_{x_n\in\mathcal{S}_n}\mathbb{P}\left(\left\{\left|\frac{1}{n}\sum_{i=1}^n G_{i,n}(x_n)\epsilon_{1,i,n}\right| \geq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\}\cap\mathcal{A}_n\,\middle|\, X_1^n\right) \\
\lesssim \;& n^{\nu_1+\eta_2\nu_2}\exp\left\{-\frac{C^2\zeta_{K,n}^2\lambda_{K,n}^2(\log n)/n}{C_1\zeta_{K,n}^2\lambda_{K,n}^2/n + C_2\zeta_{K,n}^2\lambda_{K,n}^2 M_n/n \times C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}}\right\} && (56)\\
\lesssim \;& \exp\left\{\log n - \frac{C^2\zeta_{K,n}^2\lambda_{K,n}^2(\log n)/n}{C_3\zeta_{K,n}^2\lambda_{K,n}^2/n}\right\} + \exp\left\{\log n - \frac{C\sqrt{n\log n}}{C_4\zeta_{K,n}\lambda_{K,n}M_n}\right\} && (57)
\end{aligned}$$

for finite positive constants $C_1,\ldots,C_4$ (independent of $X_1,\ldots,X_n$). Thus (51a) vanishes asymptotically for all sufficiently large C provided $M_n = O(\zeta_{K,n}^{-1}\lambda_{K,n}^{-1}\sqrt{n/(\log n)})$.

Control of the leading term in (51b): First note that $|G_{i,n}| \leq 2\zeta_{K,n}^2\lambda_{K,n}^2$ by the Cauchy-Schwarz inequality and Assumption 4(iii). This, together with Markov's inequality and Assumption 2(iii) yields

$$\begin{aligned}
& \mathbb{P}\left(\max_{x_n\in\mathcal{S}_n}\left|\frac{1}{n}\sum_{i=1}^n G_{i,n}\epsilon_{2,i,n}\right| \geq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right) \\
\lesssim \;& \frac{\zeta_{K,n}^2\lambda_{K,n}^2 E[|\epsilon_i|\{|\epsilon_i| > M_n\}]}{\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}} \leq \frac{\zeta_{K,n}\lambda_{K,n}\sqrt{n}}{\sqrt{\log n}}\frac{E[|\epsilon_i|^{2+\delta}\{|\epsilon_i| > M_n\}]}{M_n^{1+\delta}}
\end{aligned}$$

which is $o(1)$ provided $\zeta_{K,n}\lambda_{K,n}\sqrt{n/\log n} = O(M_n^{1+\delta})$. Setting $M_n^{1+\delta} \asymp \zeta_{K,n}\lambda_{K,n}\sqrt{n/\log n}$ trivially satisfies the condition $\zeta_{K,n}\lambda_{K,n}\sqrt{n/\log n} = O(M_n^{1+\delta})$. The condition $M_n = O(\zeta_{K,n}^{-1}\lambda_{K,n}^{-1}\sqrt{n/(\log n)})$ is satisfied for this choice of M_n provided $\zeta_{K,n}^2\lambda_{K,n}^2 \lesssim (n/\log n)^{\delta/(2+\delta)}$ (cf. condition (i)). Finally, it is straightforward to verify that $M_n \to \infty$ as a consequence of condition (i). Thus, both (51a) and (51b) vanish asymptotically. This completes the proof in the i.i.d. case.

With **weakly dependent data** we use $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ to bound remaining term on the right-hand side of (47) by

$$\begin{aligned} & \mathbb{P}\left(\left\{\max_{x_n \in \mathcal{S}_n} |\check{h}(x_n)| \geq 2C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right\}\right) \\ \leq\ & \mathbb{P}\left(\max_{x_n \in \mathcal{S}_n} |\widetilde{b}_w^K(x_n)'\{(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\}\widetilde{B}_w'e/n| \geq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right) \qquad (58) \\ & +\mathbb{P}\left(\max_{x_n \in \mathcal{S}_n} |\widetilde{b}_w^K(x_n)'\widetilde{B}_w'e/n| \geq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}\right) . \qquad (59) \end{aligned}$$

It is now shown that a sufficiently large C can be chosen to control terms (58) and (59).

Control of (58): The Cauchy-Schwarz inequality and Assumption 4(iii) yield

$$|\widetilde{b}_w^K(x_n)'\{(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\}\widetilde{B}_w'e/n| \lesssim \zeta_{K,n}\lambda_{K,n}\|(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\| \times O_p(\sqrt{K/n}) \qquad (60)$$

uniformly for $x_n \in \mathcal{S}_n$ (since $\|\widetilde{B}_w'e/n\| = O_p(\sqrt{K/n})$ under Assumption 2(i)(ii)). On $\mathcal{A}_n$ we have

$$\|(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\| = \|(\widetilde{B}_w'\widetilde{B}_w/n)^{-1}((\widetilde{B}_w'\widetilde{B}_w/n) - I_K)\| \leq 2\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| . \qquad (61)$$

Thus $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = O_p(\sqrt{(\log n)/K})$ (i.e. condition (ii)) ensures that (58) can be made arbitrarily small for large enough C.

Control of (59): Let M_n be as in the i.i.d. case and define

$$\begin{aligned} \epsilon_{1,i,n} &:= \epsilon_i\{|\epsilon_i| \leq M_n\} - E[\epsilon_i\{|\epsilon_i| \leq M_n\}|\mathcal{F}_{i-1}] \qquad (62) \\ \epsilon_{2,i,n} &:= \epsilon_i - \epsilon_{1,i,n} \qquad (63) \\ g_{i,n}(x_n) &:= \widetilde{b}_w^K(x_n)'\widetilde{b}_w^K(X_i)\{\mathcal{A}_n\} . \qquad (64) \end{aligned}$$

The relation $\mathbb{P}(A) \leq \mathbb{P}(A \cap B) + \mathbb{P}(B^c)$ and the triangle inequality together yield

$$\begin{aligned}
& \mathbb{P}\left(\max_{x_n \in \mathcal{S}_n} |\widetilde{b}_w^K(x_n)' \widetilde{B}_w' e/n| \geq C\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n} \right) - \mathbb{P}(\mathcal{A}_n^c) \\
\leq \quad & (\#\mathcal{S}_n) \max_{x_n \in \mathcal{S}_n} \mathbb{P}\left(\left\{ \left| \frac{1}{n}\sum_{i=1}^n g_{i,n}\epsilon_{1,i,n} \right| > \frac{C}{2}\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n} \right\} \cap \mathcal{A}_n \right) \qquad (65a) \\
& + \mathbb{P}\left(\max_{x_n \in \mathcal{S}_n} \left| \frac{1}{n}\sum_{i=1}^n g_{i,n}\epsilon_{2,i,n} \right| \geq \frac{C}{2}\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n} \right) . \qquad (65b)
\end{aligned}$$

Control of (65b): First note that $|g_{i,n}| \leq \zeta_{K,n}^2 \lambda_{K,n}^2$ by the Cauchy-Schwarz inequality and Assumption 4(iii). This, together with Markov's inequality and Assumption 2(iii) yields

$$\begin{aligned}
& \mathbb{P}\left(\max_{x_n \in \mathcal{S}_n} \left| \frac{1}{n}\sum_{i=1}^n g_{i,n}\epsilon_{2,i,n} \right| \geq \frac{C}{2}\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n} \right) \\
\lesssim \quad & \frac{\zeta_{K,n}^2\lambda_{K,n}^2 E[|\epsilon_i|\{|\epsilon_i| > M_n\}]}{\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}} \leq \frac{\zeta_{K,n}\lambda_{K,n}\sqrt{n}}{\sqrt{\log n}} \frac{E[|\epsilon_i|^{2+\delta}\{|\epsilon_i| > M_n\}]}{M_n^{1+\delta}}
\end{aligned}$$

which is $o(1)$ provided $\zeta_{K,n}\lambda_{K,n}\sqrt{n/\log n} = O(M_n^{1+\delta})$.

Control of (65a): By Assumption 2(ii), the predictable variation of the summands in (65a) may be bounded by

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n E[(g_{i,n}\epsilon_{1,i,n})^2|\mathcal{F}_{i-1}] \quad & \lesssim \quad n^{-1}\widetilde{b}_w^K(x_n)' \left(\widetilde{B}_w'\widetilde{B}_w/n \right) \widetilde{b}_w^K(x_n) \qquad (66) \\
& \lesssim \quad \zeta_{K,n}^2\lambda_{K,n}^2/n \quad \text{on } \mathcal{A}_n \qquad (67)
\end{aligned}$$

uniformly for $x_n \in \mathcal{S}_n$. Moreover,

$$|n^{-1}g_{i,n}\epsilon_{1,i,n}| \lesssim \frac{\zeta_{K,n}^2\lambda_{K,n}^2 M_n}{n} \qquad (68)$$

uniformly for $x_n \in \mathcal{S}_n$. An tail bound for martingales (Freedman, 1975, Proposition 2.1) then provides that

$$
\begin{aligned}
& (\#\mathcal{S}_n) \max_{x_n \in \mathcal{S}_n} \mathbb{P}\left(\left\{ \left| \frac{1}{n} \sum_{i=1}^n g_{i,n} \epsilon_{1,i,n} \right| > \frac{C}{2} \zeta_{K,n} \lambda_{K,n} \sqrt{(\log n)/n} \right\} \cap \mathcal{A}_n \right) \\
\lesssim \;& n^{\nu_1 + \eta_2 \nu_2} \exp\left\{ - \frac{C^2 \zeta_{K,n}^2 \lambda_{K,n}^2 (\log n)/n}{c_1 \zeta_{K,n}^2 \lambda_{K,n}^2 / n + c_2 \zeta_{K,n}^2 \lambda_{K,n}^2 M_n / n \times C \zeta_{K,n} \lambda_{K,n} \sqrt{(\log n)/n}} \right\} && (69) \\
\lesssim \;& \exp\left\{ \log n - \frac{C^2 \zeta_{K,n}^2 \lambda_{K,n}^2 (\log n)/n}{c_3 \zeta_{K,n}^2 \lambda_{K,n}^2 / n} \right\} + \exp\left\{ \log n - \frac{C \sqrt{n \log n}}{c_4 \zeta_{K,n} \lambda_{K,n} M_n} \right\} && (70)
\end{aligned}
$$

for finite positive constants $c_1, \ldots, c_4$. Thus (65a) vanishes asymptotically for all sufficiently large C provided $M_n = O(\zeta_{K,n}^{-1} \lambda_{K,n}^{-1} \sqrt{n/(\log n)})$. Choosing M_n as in the i.i.d. case completes the proof. ■

Proof of Remark 2.5. Take any $h \in L^\infty_{w,n}$ with $\|h\|_{\infty,w} \neq 0$. By the Cauchy-Schwarz inequality we have

$$
\begin{aligned}
|P_{K,w,n}(x)| &\leq \|\widetilde{b}^K_w(x)\| \|(\widetilde{B}_w' \widetilde{B}_w / n)^- \widetilde{B}_w' H / n\| && (71) \\
&\leq \zeta_{K,n} \lambda_{K,n} \|(\widetilde{B}_w' \widetilde{B}_w / n)^- \widetilde{B}_w' H / n\| && (72)
\end{aligned}
$$

uniformly over x, where $H = (h(X_1) w_n(X_1), \ldots, h(X_n) w_n(X_n))'$. When $\lambda_{\min}(\widetilde{B}_w' \widetilde{B}_w / n) \geq \frac{1}{2}$ (which it is wpa1 since $\|\widetilde{B}_w' \widetilde{B}_w / n - I_K\| = o_p(1)$), we have:

$$
\begin{aligned}
\|(\widetilde{B}_w' \widetilde{B}_w / n)^- \widetilde{B}_w' H / n\|^2 &= (H' \widetilde{B}_w / n)(\widetilde{B}_w' \widetilde{B}_w / n)^{-1} (\widetilde{B}_w' \widetilde{B}_w / n)^{-1} \widetilde{B}_w' H / n \\
&\leq 2 (H' \widetilde{B}_w / n)(\widetilde{B}_w' \widetilde{B}_w / n)^{-1} \widetilde{B}_w' H / n \\
&\leq 2 \|h\|_{w,n}^2 \leq 2 \|h\|_{\infty,w}^2
\end{aligned}
$$

where the second last line is because $\widetilde{B}_w (\widetilde{B}_w' \widetilde{B}_w)^{-1} \widetilde{B}_w'$ is a projection matrix. Thus $\|P_{K,w,n} h\|_{\infty,w} / \|h\|_{\infty,w} \leq \sqrt{2} \zeta_{K,n} \lambda_{K,n}$ wpa1 (uniformly in h). Taking the sup over h yields the desired result. ■

Proof of Lemma 2.4. It suffices to control the bias term. Note that $\widetilde{h} = P_{K,w,n} h_0$. Therefore, for any $h \in B_{K,w}$ we have, by the usual argument,

$$\begin{aligned}
\|\widetilde{h} - h_0\|_{\infty,w} &= \|\widetilde{h} - h + h - h_0\|_{\infty,w} && (73)\\
&= \|P_{K,w,n}(h_0 - h) + h - h_0\|_{\infty,w} && (74)\\
&\leq \|P_{K,w,n}(h_0 - h)\|_{\infty,w} + \|h - h_0\|_{\infty,w} && (75)\\
&\leq (1 + \|P_{K,w,n}\|_{\infty,w})\|h - h_0\|_{\infty,w} \,. && (76)
\end{aligned}$$

Taking the infimum over $h \in B_{K,w}$ yields the desired result. ■

Proof of Theorem 2.1. The variance term is $O_p(\sqrt{K(\log n)/n})$ by Lemma 2.3: condition (i) of Lemma 2.3 is satisfied by virtue of the condition $\delta \geq d/p$; condition (ii) is satisfied for $K \asymp (n/\log n)^{d/(2p+d)}$ directly in the i.i.d. case, and by Lemma 2.2 and the conditions on p for the β-mixing cases.

For the bias term, it is well known that $\inf_{h \in B_{K,w}} \|h_0 - h\|_{\infty,w} = O(K^{-p/d})$ under Assumptions 1, 6 and 7 (e.g. Huang (1998) and Chen (2007)). It therefore remains to show that $\|P_{K,w,n}\|_\infty \lesssim 1$ wpa1.

When $B_K = \text{BSpl}(K, [0,1]^d, \gamma)$, we may slightly adapt Corollary A.1 of Huang (2003b) to show that $\|P_{K,w,n}\|_{\infty,w} \lesssim 1$ wpa1, using the fact that the empirical and true $L^2(X)$ norms are equivalent over $B_{K,w}$ wpa1 by virtue of the condition $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = o_p(1)$ (see our Lemma 4.1). This condition is satisfied with $K \asymp (n/\log n)^{d/(2p+d)}$ for i.i.d. data (see Lemma 2.1), and is satisfied in the β-mixing case by Lemma 2.2 and the conditions on p.

When $B_K = \text{Wav}(K, [0,1]^d, \gamma)$, the conditions on K in Theorem 5.2 are satisfied with $K \asymp (n/\log n)^{d/(2p+d)}$ under the conditions on p in the Theorem. Therefore, $\|P_{K,w,n}\|_{\infty,w} \lesssim 1$ wpa1. ■

Proof of Lemma 2.5. By similar arguments to the proof of Lemma 2.3:

$$\|\widehat{h} - \widetilde{h}\|_{L^2(X)} = \|(\widetilde{b}_w^K)'(\widetilde{B}_w'\widetilde{B}_w/n)^- \widetilde{B}_w' e/n\|_{L^2(X)} \leq \|(\widetilde{B}_w'\widetilde{B}_w/n)^-\| \|\widetilde{B}_w' e/n\| \,. \tag{77}$$

Chebyshev's inequality and Assumption 2(i)(ii) yield $\|\widetilde{B}_w' e/n\| = O_p(\sqrt{K/n})$. Moreover, it follows from Assumption 5 that $\|(\widetilde{B}_w'\widetilde{B}_w/n)^-\| = O_p(1)$.

For the remaining term it suffices to show that $\|\widetilde{h} - h_0\|_{L^2(X)} = O_p(\|h_0 - h_{0,K}\|_{L^2(X)})$. By the triangle inequality we bound

$$\|\widetilde{h} - h_0\|_{L^2(X)} \leq \|\widetilde{h} - h_{0,K}\|_{L^2(X)} + \|h_{0,K} - h_0\|_{L^2(X)} \,. \tag{78}$$

Recall the definition of the empirical projection $P_{K,w,n}$ from expression (15), and observe that $\widetilde{h} = P_{K,w,n}h_0$ and that $P_{K,w,n}h = h$ for all $h \in B_{K,w}$. Also recall the definition of $L^2_{w,n}(X)$ as the space of functions with finite norm $\|\cdot\|_{w,n}$ where $\|f\|^2_{w,n} = \frac{1}{n}\sum_{i=1}^n f(X_i)^2 w_n(X_i)$. Since the empirical and theoretical $L^2(X)$ norms are equivalent over $B_{K,w}$ wpa1 under the condition $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = o_p(1)$ (see Section 4.2). Therefore, we have

$$\begin{aligned}
\|\widetilde{h} - h_{0,K}\|^2_{L^2(X)} &= \|P_{K,w,n}(h_0 - h_{0,K})\|^2_{L^2(X)} && (79)\\
&\asymp \|P_{K,w,n}(h_0 - h_{0,K})\|^2_{w,n} \quad \text{wpa1} && (80)\\
&\leq \|(h_0 - h_{0,K})\|^2_{w,n} && (81)
\end{aligned}$$

where the second line is by equivalence of the empirical and theoretical $L^2(X)$ norms wpa1, and the final line is because $P_{K,w,n}$ is an orthogonal projection on $L^2_{w,n}(X)$. Finally, Markov's inequality yields $\|(h_0 - h_{0,K})\|^2_{w,n} = O_p(\|h_0 - h_{0,K}\|^2_{L^2(X)})$. ■

7.2 Proofs for Section 3

Proof of Lemma 3.1. We use a truncation argument together with exponential inequalities for random matrices. Let $M_n \asymp (\zeta_{K,n}\lambda_{K,n})^{(2+\delta)/\delta}$ (with δ as in Assumption 2(iii) be a sequence of positive numbers and let

$$\begin{aligned}
\widehat{\Omega}_1 &= \frac{1}{n}\sum_{i=1}^n (\Xi_{1,i} - E[\Xi_{1,i}]) && (82)\\
\widehat{\Omega}_2 &= \frac{1}{n}\sum_{i=1}^n (\Xi_{2,i} - E[\Xi_{2,i}]) && (83)\\
\Xi_{1,i} &= \epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\{\|\epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\| \leq M_n^2\} && (84)\\
\Xi_{2,i} &= \epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\{\|\epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\| > M_n^2\}\,. && (85)
\end{aligned}$$

Clearly $\widehat{\Omega} - \Omega = \widehat{\Omega}_1 + \widehat{\Omega}_2$, so it is enough to show that $\|\widehat{\Omega}_1\| = o_p(1)$ and $\|\widehat{\Omega}_2\| = o_p(1)$.

Control of $\|\widehat{\Omega}_1\|$: By definition, $\|\Xi_{1,i}\| \leq M_n^2$. It follows by the triangle inequality and Jensen's inequality ($\|\cdot\|$ is convex) that $\|\Xi_{1,i} - E[\Xi_{1,i}]\| \leq 2M_n$. Moreover, by Assumption 2(ii)

$$\begin{aligned} E[(\Xi_{1,i} - E[\Xi_{1,i}])^2] &\leq E[\epsilon_i^4 \|\widetilde{b}_w^K(X_i)\|^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\{\|\epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\| \leq M_n^2\}] & (86)\\ &\leq M_n^2 E[\epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\{\|\epsilon_i^2 \widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'\| \leq M_n^2\}] & (87)\\ &\leq M_n^2 E[E[\epsilon_i^2|X_i]\widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'] & (88)\\ &\lesssim M_n^2 E[\widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'] \;=\; M_n^2 I_K & (89) \end{aligned}$$

where the inequalities are understood in the sense of positive semi-definite matrices. It follows that $\|E[(\Xi_{1,i} - E[\Xi_{1,i}])^2]\| \lesssim M_n^2$. In the i.i.d. case, Corollary 4.1 yields $\|\widehat{\Omega}_1\| = O_p(M_n\sqrt{(\log K)/n}) = o_p(1)$. In the β-mixing case, Corollary 4.2 yields $\|\widehat{\Omega}_1\| = O_p(M_n\sqrt{q(\log K)/n})$, and the result follows by taking $q = \gamma^{-1}\log n$ in the exponentially β-mixing case and $q \asymp n^{1/(1+\gamma)}$ in the algebraically β-mixing case.

Control of $\|\widehat{\Omega}_2\|$: The simple bound $\|\Xi_{2,i}\| \leq (\zeta_{K,n}\lambda_{K,n})^2\epsilon_i^2\{\epsilon_i^2 > M_n^2/(\zeta_{K,n}\lambda_{K,n})^2\}$ together with the triangle inequality and Jensen's inequality ($\|\cdot\|$ is convex) yield

$$\begin{aligned} E[\|\widehat{\Omega}_2\|] &\leq 2(\zeta_{K,n}\lambda_{K,n})^2 E[\epsilon_i^2\{|\epsilon_i| > M_n/(\zeta_{K,n}\lambda_{K,n})\}] & (90)\\ &\leq 2\frac{(\zeta_{K,n}\lambda_{K,n})^{2+\delta}}{M_n^\delta} E[|\epsilon_i|^{2+\delta}\{|\epsilon_i| > M_n/(\zeta_{K,n}\lambda_{K,n})\}] \;=\; o(1) & (91) \end{aligned}$$

by Assumption 2(iii) because $M_n/(\zeta_{K,n}\lambda_{K,n}) \asymp (\zeta_{K,n}\lambda_{K,n})^{2/\delta} \to \infty$ and $(\zeta_{K,n}\lambda_{K,n})^{2+\delta}/M_n^\delta \asymp 1$. Therefore, $\|\widehat{\Omega}_2\| = o_p(1)$ by Markov's inequality. ■

Proof of Theorem 3.1. First define $u_K^*(x) = v_K^*(x)/\|v_K^*\|_{sd}$. Note that $E[(u_K^*(X_i)\epsilon_i)^2] = 1$, $E[u_K^*(X_i)^2] = \|v\|_{L^2(X)}^2/\|v\|_{sd}^2 \asymp 1$ (by Assumptions 2(ii)(iv)), and $\|u_K^*\|_\infty \lesssim \zeta_{K,n}\lambda_{K,n}$ by the relation between the L^2 and sup norms on $B_{K,w}$.

By Assumption 9(i)(ii) and the fact that $\widehat{h}, \widetilde{h} \in N_{K,n}$ wpa1, we obtain:

$$\begin{aligned} \frac{\sqrt{n}(f(\widehat{h}) - f(\widetilde{h}))}{V_K^{1/2}} &= \frac{\sqrt{n}\frac{\partial f(h_0)}{\partial h}[\widehat{h} - \widetilde{h}]}{V_K^{1/2}} + o_p(1) & (92)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n u_K^*(X_i)\epsilon_i + \frac{\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]'((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)(\widetilde{B}_w' e/\sqrt{n})}{V_K^{1/2}} + o_p(1)\,. & (93) \end{aligned}$$

The leading term is now shown to be asymptotically $N(0,1)$ and the second term is shown to be asymptotically negligible. The proof of this differs depending upon whether the data are i.i.d. or weakly dependent.

With **i.i.d. data**, we first show that the second term on the right-hand side of (93) is $o_p(1)$. Let $\eta > 0$ be arbitrary. Let C_η be such that $\limsup \mathbb{P}(\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| > C_\eta \zeta_{K,n}\lambda_{K,n}\sqrt{(\log K)/n}) \leq \eta$ (we may always choose such a C_η by Lemma 2.1), let $\mathcal{C}_{n,\eta}$ denote the event $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| \leq C_\eta \zeta_{K,n}\lambda_{K,n}\sqrt{(\log K)/n}$ and let $\{\mathcal{C}_{n,\eta}\}$ denote its indicator function. Observe that $V_K^{1/2} \asymp \left\|\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]\right\|$ (under Assumption 2(ii)(iv)). Let $E[\,\cdot\,|X_1^n]$ denote expectation conditional on $X_1, \ldots, X_n$ and let $\partial\widetilde{b}_w^K$ denote $\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]$. By iterated expectations,

$$
\begin{aligned}
& E\left[\left(\frac{(\partial\widetilde{b}_w^K)'((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)(\widetilde{B}_w'e/\sqrt{n})}{V_K^{1/2}}\right)^2 \{\mathcal{C}_{n,\eta}\}\right] \\
= \; & \frac{(\partial\widetilde{b}_w^K)'E[((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)E[(\widetilde{B}_w'ee'\widetilde{B}_w/n)|X_1^n]((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)\{\mathcal{C}_{n,\eta}\}]\partial\widetilde{b}_w^K}{V_K} \\
= \; & \frac{(\partial\widetilde{b}_w^K)'E[((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)(\frac{1}{n}\sum_{i=1}^n E[\epsilon_i^2\widetilde{b}_w^K(X_i)\widetilde{b}_w^K(X_i)'|X_i])((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)\{\mathcal{C}_{n,\eta}\}]\partial\widetilde{b}_w^K}{V_K} \\
\lesssim \; & \frac{(\partial\widetilde{b}_w^K)'E[((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)(\widetilde{B}_w'\widetilde{B}_w/n)((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)\{\mathcal{C}_{n,\eta}\}]\partial\widetilde{b}_w^K}{V_K} \\
\lesssim \; & C_\eta^2\zeta_{K,n}^2\lambda_{K,n}^2(\log K)/n \quad = \quad o(1) \qquad (94)
\end{aligned}
$$

for all n sufficiently large, where the second last line is by Assumption 2(ii) and the final line is because both $\|(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\| \lesssim \|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\|$ and $\|(\widetilde{B}_w'\widetilde{B}_w)/n\| \lesssim 1$ hold on $\mathcal{C}_{n,\eta}$ for all n sufficiently large under Assumption 5. As $\liminf \mathbb{P}(\mathcal{C}_{n,\eta}) \geq 1 - \eta$ and η is arbitrary, the second term in (93) is therefore $o_p(1)$.

Now consider the leading term in (93). The summands are i.i.d. with mean zero and unit variance. The Lindeberg condition is easily verified:

$$
\begin{aligned}
E[\epsilon_i^2 u_K^*(X_i)^2\{|\epsilon_i u_K^*(X_i)| > \eta\sqrt{n}\}] \;&=\; E[\epsilon_i^2 u_K^*(X_i)^2\{|\epsilon_i| > \eta(\sqrt{n}/\zeta_{K,n}\lambda_{K,n})\}] && (95) \\
&\leq\; \sup_x E[\epsilon_i^2\{|\epsilon_i| > \eta(\sqrt{n}/\zeta_{K,n}\lambda_{K,n})\}|X_i = x] \;=\; o(1) && (96)
\end{aligned}
$$

by Assumption 2(v) because $\zeta_{K,n}^2\lambda_{K,n}^2/n = o(1)$. Thus the leading term is asymptotically $N(0,1)$ by the Lindeberg-Feller theorem.

With **weakly dependent data** we apply the Cauchy-Schwarz inequality to the second term in expression (93) to obtain

$$\begin{aligned}
& \left|\frac{\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]'((\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K)(\widetilde{B}_w'e/\sqrt{n})}{V_K^{1/2}}\right| && (97)\\
\leq\ & \frac{\left\|\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]\right\| \|(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\| \|\widetilde{B}_w'e/n\| \times \sqrt{n}}{V_K^{1/2}} && (98)\\
\lesssim\ & \|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| \|\widetilde{B}_w'e/n\| \times \sqrt{n} && (99)
\end{aligned}$$

wpa1, because $\left\|\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]\right\| \asymp V_K^{1/2}$ and $\|(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\| \leq 2\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\|$ wpa1 by Assumption 5. Assumption 2(i)(ii) implies $\|\widetilde{B}_w'e/n\| = O_p(\sqrt{K/n})$, whence the second term in expression (93) is $o_p(1)$ by the condition $\|(\widetilde{B}_w'\widetilde{B}_w/n) - I_K\| = o_p(K^{-1/2})$.

To show the leading term in (93) is asymptotically $N(0,1)$ we use a martingale CLT (Corollary 2.8 of McLeish (1974)). This verifying the conditions (a) $\max_{i\leq n} |u_K^*(X_i)\epsilon_i/\sqrt{n}| \to_p 0$ and (b) $\frac{1}{n}\sum_{i=1}^n u_K^*(X_i)^2\epsilon_i^2 \to_p 1$. To verify condition (a), let $\eta > 0$ be arbitrary. Then,

$$\begin{aligned}
\mathbb{P}(\max_{i\leq n} |\epsilon_i u_K^*(X_i)/\sqrt{n}| > \eta) & \leq \sum_{i=1}^n \mathbb{P}(|\epsilon_i u_K^*(X_i)/\sqrt{n}| > \eta) && (100)\\
& \leq \frac{1}{n\eta^2}\sum_{i=1}^n E[\epsilon_i^2 u_K^*(X_i)^2\{|\epsilon_i u_K^*(X_i)/\sqrt{n}| > \eta\}] && (101)\\
& = \frac{1}{\eta^2} E[\epsilon_i^2 u_K^*(X_i)^2\{|\epsilon_i u_K^*(X_i)/\sqrt{n}| > \eta\}] && (102)
\end{aligned}$$

which again is $o(1)$ by Assumption 2(v) since $\zeta_{K,n}^2\lambda_{K,n}^2/n = o(1)$. For condition (b), note that $\|\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]/\|v_K^*\|_{sd}\| = \|v_K^*\|_{L^2(X)}/\|v_K^*\|_{sd} \asymp 1$. Then by the Cauchy-Schwarz inequality, we have

$$\left|\frac{1}{n}\sum_{i=1}^n u_K^*(X_i)^2\epsilon_i^2 - 1\right| = \left|\left(\frac{\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]}{\|v_K^*\|_{sd}}\right)'(\widehat{\Omega} - \Omega)\left(\frac{\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]}{\|v_K^*\|_{sd}}\right)\right| \lesssim \|\widehat{\Omega} - \Omega\| \qquad (103)$$

which is $o_p(1)$ by Assumption 8. Therefore, the leading term in (93) is asymptotically $N(0,1)$.

It remains to show that

$$\frac{\sqrt{n}(f(\widetilde{h}) - f(h_0))}{V_K^{1/2}} = o_p(1)\,. \qquad (104)$$

Assumption 9(ii) and the fact that $\widetilde{h} \in N_{K,n}$ wpa1 together yield

$$\frac{\sqrt{n}(f(\widetilde{h}) - f(h_0))}{V_K^{1/2}} = \sqrt{\frac{n}{V_K}}\frac{\partial f(h_0)}{\partial h}[\widetilde{h} - h_0] + o_p(1)\,. \qquad (105)$$

which is $o_p(1)$ by Assumption 9(iii). ■

Proof of Corollary 3.1. The result follows from Theorem 3.1. Assumption 4(iii) is satisfied for these bases under Assumptions 1 and 7. Moreover, Assumption 5 is satisfied under the restrictions on K (see Lemmas 2.1 and 2.2). Assumption 9 is satisfied provided $\sqrt{n}\|\widetilde{h} - h_0\|_\infty = o(V_K^{-1/2})$. But $\|\widetilde{h} - h_0\|_\infty = O_p(K^{-p/d})$ by the proof of Theorem 2.1, so $\sqrt{n}V_K^{-1/2}K^{-p/d} = o(1)$ is sufficient for Assumption 9 to hold. Moreover, under Assumption 2(iii), Lemma 3.1 shows that Assumption 8 and the condition $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = o_p(K^{-1/2})$ are satisfied for weakly dependent data under the respective conditions on K (see Lemma 2.2). ■

Proof of Corollary 3.2. The result follows by Theorem 3.1 with Assumption 9(i')–(iv') in place of Assumption 9 (see Remark 3.1). Most conditions of Theorem 3.1 can be verified in the same way as those for Corollary 3.1. Assumption 9(iv') is satisfied under the conditions on K because $\|\widetilde{h} - h_0\|_\infty = O_p(K^{-p/d})$ by the proof of Theorem 2.1, and $\|\widehat{h} - \widetilde{h}\|_\infty = O_p(\sqrt{(K\log n)/n})$ by Lemma 2.3. ■

Proof of Lemma 3.2. Result (2) follows from Theorem 3.1 and Result (1) of Lemma 3.2 by the continuous mapping theorem. It remains to show Result (1). By addition and subtraction of terms,

$$\begin{aligned}
\frac{\widehat{\|v_K^*\|_{sd}}^2}{\|v_K^*\|_{sd}^2} &= \frac{1}{n}\sum_{i=1}^n \frac{\epsilon_i^2 v_K^*(X_i)^2}{\|v_K^*\|_{sd}^2} + \frac{1}{n}\sum_{i=1}^n \frac{\epsilon_i^2(\widehat{v}_K^*(X_i)^2 - v_K^*(X_i)^2)}{\|v_K^*\|_{sd}^2} \qquad (106)\\
&\quad + \frac{1}{n}\sum_{i=1}^n \frac{(\widehat{h}(X_i) - h_0(X_i))^2 v_K^*(X_i)^2}{\|v_K^*\|_{sd}^2} + \frac{1}{n}\sum_{i=1}^n \frac{(\widehat{h}(X_i) - h_0(X_i))^2(\widehat{v}_K^*(X_i)^2 - v_K^*(X_i)^2)}{\|v_K^*\|_{sd}^2}\\
&\quad - \frac{2}{n}\sum_{i=1}^n \frac{\epsilon_i(\widehat{h}(X_i) - h_0(X_i))v_K^*(X_i)^2}{\|v_K^*\|_{sd}^2} - \frac{2}{n}\sum_{i=1}^n \frac{\epsilon_i(\widehat{h}(X_i) - h_0(X_i))(\widehat{v}_K^*(X_i)^2 - v_K^*(X_i)^2)}{\|v_K^*\|_{sd}^2}\\
&=: T_1 + T_2 + T_3 + T_4 + T_5 + T_6\,.
\end{aligned}$$

Control of T_1: $T_1 \to_p 1$ by Assumption 8.

Control of T_2: Let

$$\partial = \frac{\frac{\partial f(h_0)}{\partial h}[\widetilde{b}_w^K]}{\|v_K^*\|_{sd}} \qquad (107)$$

$$\widehat{\partial} = \frac{\frac{\partial f(\widehat{h})}{\partial h}[\widetilde{b}_w^K]}{\|v_K^*\|_{sd}} \qquad (108)$$

$$\widehat{\widehat{\partial}} = (\widetilde{B}_w'\widetilde{B}_w/n)^{-1}\widehat{\partial}\,. \qquad (109)$$

Then with this notation,

$$|T_2| = \left|(\widehat{\widehat{\partial}})'\widehat{\Omega}\widehat{\widehat{\partial}} - \partial'\widehat{\Omega}\partial\right| = |(\widehat{\widehat{\partial}} + \partial)'\widehat{\Omega}(\widehat{\widehat{\partial}} - \partial)| \leq \|(\widehat{\widehat{\partial}} + \partial)'\|\|\widehat{\Omega}\|\|(\widehat{\widehat{\partial}} - \partial)\|\,. \tag{110}$$

Note that $\|\widehat{\Omega}\| = O_p(1)$ by Assumption 8(ii). By the triangle inequality and definition of ∂, $\widehat{\partial}$, and $\widehat{\widehat{\partial}}$,

$$\|\widehat{\widehat{\partial}} - \widehat{\partial}\| \leq \|(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\|(\|\widehat{\partial} - \partial\| + \|\partial\|)\,. \tag{111}$$

Assumption 5 implies $\|(\widetilde{B}_w'\widetilde{B}_w/n)^- - I_K\| = o_p(1)$; $\|\widehat{\partial} - \partial\| = o_p(1)$ by Assumption 9(iv), because $\widehat{h} \in N_{K,n}$ wpa1; and $\|\partial\| \asymp 1$ because $\|v_K^*\|_{L^2(X)} \asymp \|v_K^*\|_{sd}$ under Assumption 2(ii)(iv). Therefore, $\|\widehat{\widehat{\partial}} - \widehat{\partial}\| = o_p(1)$, $\|\widehat{\widehat{\partial}} + \widehat{\partial}\| = O_p(1)$, and so $|T_2| = o_p(1)$.

Control of T_3: First note that

$$|T_3| \leq \|\widehat{h} - h_0\|_{\infty,w}^2 \times \frac{1}{n}\sum_{i=1}^n \frac{v_K(X_i)^2}{\|v_K^*\|_{sd}^2} = o_p(1) \times O_p(1) = o_p(1) \tag{112}$$

where $\|\widehat{h} - h_0\|_{\infty,w} = o_p(1)$ by hypothesis and $n^{-1}\sum_{i=1}^n v_K(X_i)^2/\|v_K^*\|_{sd}^2$ by Markov's inequality and the fact that $\|v_K^*\|_{L^2(X)} \asymp \|v_K^*\|_{sd}$ under Assumption 2(ii)(iv).

Control of T_4: by the triangle inequality definition of $\widehat{v}_K^*$ and v_K^*:

$$\begin{aligned} |T_4| &\leq \|\widehat{h} - h_0\|_{\infty,w}^2 \times \left(\frac{1}{n}\sum_{i=1}^n \frac{\widehat{v}_K^*(X_i)^2}{\|v_K^*\|_{sd}^2} + \frac{1}{n}\sum_{i=1}^n \frac{v_K^*(X_i)^2}{\|v_K^*\|_{sd}^2}\right) \qquad (113)\\ &= o_p(1) \times \left(\widehat{\widehat{\partial}}'\widehat{\Omega}\widehat{\widehat{\partial}} + \partial'\widehat{\Omega}\partial\right) \leq o_p(1) \times \|\widehat{\Omega}\| \times \left(\|\widehat{\widehat{\partial}}\|^2 + \|\partial\|^2\right). \qquad (114)\end{aligned}$$

Moreover, $\|\widehat{\Omega}\| = O_p(1)$ by Assumption 8, $\|\partial\| \asymp 1$ by Assumption 2(ii)(iv), and $\|\widehat{\widehat{\partial}}\| \leq \|\widehat{\widehat{\partial}} - \widehat{\partial}\| + \|\widehat{\partial} - \partial\| + \|\partial\| = O_p(1)$ by Assumption 5 and 9(iv). It follows that $|T_4| = o_p(1)$.

Control of T_5: By the inequality $2|a| \leq 1 + a^2$, we have

$$|T_5| \leq \|\widehat{h} - h_0\|_{\infty,w}\frac{1}{n}\sum_{i=1}^n \frac{(1+\epsilon_i^2)v_K^*(X_i)^2}{\|v_K^*\|_{sd}^2} = o_p(1) \times O_p(1) = o_p(1) \tag{115}$$

where $\|\widehat{h} - h_0\|_{\infty,w} = o_p(1)$ by hypothesis, $n^{-1}\sum_{i=1}^n \epsilon_i^2 v_K^*(X_i)^2/\|v_K^*\|_{sd}^2 \to_p 1$ by Assumption 8, and the remaining term is $O_p(1)$ by the arguments for T_3.

Control of T_6: The proof is essentially the same as that for T_2, except we replace $\widehat{\Omega}$ by the matrix

$\widehat{\mho} = n^{-1}\sum_{i=1}^n \epsilon_i(\widehat{h}(X_i) - h_0(X_i))\widetilde{b}^K_w(X_i)\widetilde{b}^K_w(X_i)'$. By the inequality $2|a| \leq 1 + a^2$, it follows that

$$\begin{aligned} \|\widehat{\mho}\| &\leq \|\widehat{h} - h_0\|_{\infty,w} \times \left\| n^{-1} \sum_{i=1}^n (1+\epsilon_i^2)\widetilde{b}^K_w(X_i)\widetilde{b}^K_w(X_i)' \right\| && (116) \\ &= \|\widehat{h} - h_0\|_{\infty,w} \times \left\| \widetilde{B}_w'\widetilde{B}_w/n + \widehat{\Omega} \right\| \;=\; o_p(1) \times O_p(1) \;=\; o_p(1) && (117) \end{aligned}$$

because $\|\widehat{h} - h_0\|_{\infty,w} = o_p(1)$, $\|\widetilde{B}_w'\widetilde{B}_w/n\| = O_p(1)$ by Assumption 5, and $\|\widehat{\Omega}\| = O_p(1)$ by Assumption 8. ■

Proof of Theorem 3.2. This follows from Lemma 3.2.

First, Assumption 5 is satisfied for i.i.d. and β-mixing data under the respective conditions on K (see Lemmas 2.1 and 2.2). Moreover, Lemma 3.1 shows that Assumption 8 and the condition $\|\widetilde{B}_w'\widetilde{B}_w/n - I_K\| = o_p(K^{-1/2})$ is satisfied for weakly dependent data under the respective conditions on K (see Lemma 2.2). Therefore Theorem 3.1 may be applied for asymptotic normality of $f(\widehat{h})$.

To apply Lemma 3.2 it remains to show that $\|\widehat{h} - h_0\|_{\infty,w} = o_p(1)$. But $\|\widetilde{h} - h_0\|_\infty = o_p(1)$ by assumption, and $\|\widehat{h} - \widetilde{h}\|_\infty = O_p(\zeta_{K,n}\lambda_{K,n}\sqrt{(\log n)/n}) = o_p(1)$ by Lemmas 2.3, 2.1 and 2.2 under the conditions on K. ■

7.3 Proofs for Section 4

Proof of Corollary 4.1. Follows from Theorem 4.1 with $t = C\sigma_n\sqrt{\log(d_1+d_2)}$ for sufficiently large C, and applying the condition $R_n\sqrt{\log(d_1+d_2)} = o(\sigma_n)$. ■

Proof of Theorem 4.2. By Berbee's lemma (enlarging the probability space as necessary) the process $\{X_i\}$ can be coupled with a process X_i^* such that $Y_k := \{X_{(k-1)q+1}, \ldots, X_{kq}\}$ and $Y_k^* := \{X^*_{(k-1)q+1}, \ldots, X^*_{kq}\}$ are identically distributed for each $k \geq 1$, $\mathbb{P}(Y_k \neq Y_k^*) \leq \beta(q)$ for each $k \geq 1$ and $\{Y_1^*, Y_3^*, \ldots\}$ are independent and $\{Y_2^*, Y_4^*, \ldots\}$ are independent (see Lemma 2.1 of Berbee (1987)). Let I_e and I_o denote the indices of $\{1, \ldots, n\}$ corresponding to the odd- and even-numbered blocks, and I_r the indices in the remainder, so $I_r = q[n/q]+1, \ldots, n$ when $q[n/q] < n$ and $I_r = \emptyset$ when $q[n/q] = n$.

Let $\Xi^*_{i,n} = \Xi(X^*_{i,n})$. By the triangle inequality,

$$\begin{aligned}
& \mathbb{P}\left(\|\textstyle\sum_{i=1}^{n} \Xi_{i,n}\| \geq 6t\right) \\
\leq\;& \mathbb{P}(\|\textstyle\sum_{i=1}^{[n/q]q} \Xi^*_{i,n}\| + \|\sum_{i\in I_r} \Xi_{i,n}\| + \|\sum_{i=1}^{[n/q]q}(\Xi^*_{i,n} - \Xi_{i,n})\| \geq 6t) \qquad (118) \\
\leq\;& \tfrac{n}{q}\beta(q) + \mathbb{P}\left(\|\textstyle\sum_{i\in I_r} \Xi_{i,n}\| \geq t\right) + \mathbb{P}\left(\|\sum_{i\in I_e} \Xi^*_{i,n}\| \geq t\right) + \mathbb{P}\left(\|\sum_{i\in I_o} \Xi^*_{i,n}\| \geq t\right)
\end{aligned}$$

To control the last two terms we apply Theorem 4.1, recognizing that $\sum_{i\in I_e} \Xi^*_{i,n}$ and $\sum_{i\in I_o} \Xi^*_{i,n}$ are each the sum of fewer than $[n/q]$ independent $d_1 \times d_2$ matrices, namely $W^*_k = \sum_{i=(k-1)q+1}^{kq} \Xi^*_{i,n}$. Moreover each W^*_k satisfies $\|W^*_k\| \leq qR_n$ and $\max\{\|E[W^*_k W^{*\prime}_k]\|, \|E[W^{*\prime}_k W^*_k]\|\} \leq q^2 s_n$. Theorem 4.1 then yields

$$\mathbb{P}\left(\left\|\sum_{i\in I_e} \Xi^*_{i,n}\right\| \geq t\right) \leq (d_1+d_2)\text{exp}\left(\frac{-t^2/2}{nqs_n^2 + qR_n t/3}\right) \tag{119}$$

and similarly for I_o. ■

Proof of Corollary 4.2. Follows from Theorem 4.2 with $t = Cs_n\sqrt{nq\log(d_1+d_2)}$ for sufficiently large C, and the conditions $\frac{n}{q}\beta(q) = o(1)$ and $R_n\sqrt{q\log(d_1+d_2)} = o(s_n\sqrt{n})$. ■

Proof of Lemma 4.1. Let $G = E[b^K_w(X_i)b^K_w(X_i)']$. Since $B_{K,w} = clsp\{b_{K1}w_n, \ldots, b_{KK}w_n\}$, we have:

$$\begin{aligned}
& \sup\{|\tfrac{1}{n}\textstyle\sum_{i=1}^{n} b(X_i)^2 - 1| : b \in B_{K,w}, E[b(X)^2] = 1\} \\
=\;& \sup\{|c'(B_w'B_w/n - G)c| : c \in \mathbb{R}^K, \|G^{1/2}c\| = 1\} \qquad (120) \\
=\;& \sup\{|c'G^{1/2}(G^{-1/2}(B_w'B_w/n)G^{-1/2} - I_K)G^{1/2}c| : c \in \mathbb{R}^K, \|G^{1/2}c\| = 1\} \qquad (121) \\
=\;& \sup\{|c'(\widetilde{B}_w'\widetilde{B}_w/n - I_K)c| : c \in \mathbb{R}^K, \|c\| = 1\} \qquad (122) \\
=\;& \|\widetilde{B}_w'\widetilde{B}_w/n - I_K\|_2^2 \qquad (123)
\end{aligned}$$

as required. ■

7.4 Proofs for Section 5

We first present a general result that allows us to bound the L^∞ operator norm of the $L^2(X)$ projection P_K onto a linear sieve space $B_K \equiv clsp\{b_{K1}, \ldots, b_{KK}\}$ by the ℓ^∞ norm of the inverse of its corresponding Gram matrix.

Lemma 7.1 *If there exists a sequence of positive constants $\{c_K\}$ such that (i) $\sup_{x\in\mathcal{X}} \|b^K(x)\|_{\ell^1} \lesssim c_K$*

and (ii) $\max_{1\leq k\leq K}\|b_{Kk}\|_{L^1(X)} \lesssim c_K^{-1}$*, then*

$$\|P_K\|_\infty \lesssim \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty} .$$

Proof of Lemma 7.1. By Hölder's inequality (with (i)), definition of the operator norm, and Hölder's inequality again (with (ii)), we obtain:

$$\begin{aligned}
|P_K f(x)| &= |b^K(x)' \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} E[b^K(X_i)f(X_i)]| \\
&\leq \|b^K(x)\|_{\ell^1} \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} E[b^K(X_i)f(X_i)]|\|_{\ell^\infty} \\
&\lesssim c_K \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} E[b^K(X_i)f(X_i)]|\|_{\ell^\infty} \\
&\leq c_K \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty} \|E[b^K(X_i)f(X_i)]|\|_{\ell^\infty} \\
&= c_K \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty} \max_{1\leq k\leq K} E[|b_{Kk}(X_i)f(X_i)|] \\
&\leq c_K \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty} \max_{1\leq k\leq K} E[|b_{Kk}(X_i)|]\|f\|_\infty \\
&\lesssim \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_\infty \|f\|_\infty
\end{aligned}$$

uniformly in x. The result now follows by taking the supremum over $x \in \mathcal{X}$. ■

We will bound $\| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty}$ for (tensor product) wavelet bases using the following Lemma.

Lemma 7.2 *Let* $A \in \mathbb{R}^{K\times K}$ *be a positive definite symmetric matrix such that* $A_{i,j} = 0$ *whenever* $|i-j| > m/2$ *for* m *even. Then:* $\|A^{-1}\|_{\ell^\infty} \leq \frac{2C}{1-\lambda}$ *where*

$$\begin{aligned}
\kappa &= \lambda_{\max}(A)/\lambda_{\min}(A) \\
\lambda &= \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2/m} < 1 \\
C &= \|A^{-1}\| \max\{1, (1+\sqrt{\kappa})^2/(2\kappa)\} .
\end{aligned}$$

Proof of Lemma 7.2. By definition of the matrix infinity norm, we have

$$\|A^{-1}\|_\infty = \max_{j\leq K} \sum_{k=1}^K |(A^{-1})_{j,k}| .$$

The result now follows by Theorem 2.4 of Demko, Moss, and Smith (1984) (which states that $\left|(A^{-1})_{i,j}\right| \leq C\lambda^{|i-j|}$ for all i,j) and geometric summation. ■

Proof of Theorem 5.1. We first prove the univariate case (i.e. $d = 1$) before generalizing to the multivariate case.

By the definition of wavelet basis, we may assume without loss of generality that $b_{K1} = \varphi_{J,0}, \ldots, b_{KK} = \varphi_{J,2^J-1}$ with $K = 2^J$.

For any $x \in [0,1]$ the vector $b^K(x)$ has, at most, $2N$ elements that are nonzero (as a consequence of the compact support of the $\varphi_{J,k}$). It follows that

$$\|b^K(x)\|_{\ell^1} \leq (2N)2^{J/2}\max\{\|\varphi\|_\infty, \|\varphi_0^l\|_\infty, \ldots, \|\varphi_{N-1}^l\|_\infty, \|\varphi_{-1}^r\|_\infty, \ldots, \|\varphi_{-N}^r\|_\infty\} \lesssim 2^{J/2} \tag{124}$$

uniformly in x. Therefore $\sup_{x\in[0,1]}\|b^K(x)\|_{\ell^1} \lesssim \sqrt{K}$. Let k be such that $N \leq k \leq 2^J - N - 1$. By boundedness of f_X and a change of variables, we have (with μ denoting Lebesgue measure)

$$\begin{aligned} E[|\varphi_{J,k}(X_i)|] &\leq \sup_{x\in[0,1]} f_X(x)\int_{\mathbb{R}} 2^{J/2}|\varphi(2^Jx-k)|\,\mathrm{d}\mu(x) & (125)\\ &= \sup_{x\in[0,1]} f_X(x)2^{-J/2}\int_{\mathbb{R}}|\varphi(y)|\,\mathrm{d}\mu(y) & (126)\\ &= \sup_{x\in[0,1]} f_X(x)2^{-J/2}\|\varphi\|_{L^1(\mu)} & (127)\end{aligned}$$

where $\|\varphi\|_{L^1(\mu)} < \infty$ because φ has compact support and is continuous. Similar arguments can be used to show the same for the N left and right scaling functions. It follows that $\max_k\|b_{Kk}\|_{L^1(X)} \lesssim 2^{-J/2} = K^{-1/2}$. Therefore, the $b_{K1}, \ldots, b_{KK}$ satisfy the conditions of Lemma 7.1 and hence $\|P_K\|_\infty \lesssim \|\left(E[b^K(X_i)b^K(X_i)']\right)^{-1}\|_{\ell^\infty}$.

It remains to prove that $\|\left(E[b^K(X_i)b^K(X_i)']\right)^{-1}\|_{\ell^\infty} \lesssim 1$. We first verify the conditions of Lemma 7.2. Disjoint support of the $\varphi_{J,k}$ implies that $(E[b^K(X_i)b^K(X_i)'])_{k,j} = 0$ whenever $|k-j| > 2N-1$. For positive definiteness, we note that

$$\lambda_{\max}(E[b^K(X_i)b^K(X_i)']) \leq \left(\sup_{x\in[0,1]} f_X(x)\right)\lambda_{\max}\left(\int_{[0,1]} b^K(x)b^K(x)'\,\mathrm{d}\mu(x)\right) = \left(\sup_{x\in[0,1]} f_X(x)\right) \tag{128}$$

(where we understand the integral performed element wise) because $\varphi_{J,0}, \ldots, \varphi_{J,2^J-1}$ are an orthonormal basis for V_J with respect to the $L^2([0,1])$ inner product. Similarly, $\lambda_{\min}(E[b^K(X)b^K(X)']) \geq$

$\inf_{x\in[0,1]} f_X(x)$. Therefore

$$\kappa \leq (\sup_{x\in[0,1]} f_X(x))/(\inf_{x\in[0,1]} f_X(x)) < \infty$$

uniformly in K, and

$$\| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \| \leq 1/(\inf_{x\in[0,1]} f_X(x)) < \infty$$

uniformly in K. This verifies the conditions of Lemma 7.2 for $A = E[b^K(X_i)b^K(X_i)']$. It follows by Lemma 7.2 that $\| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty} \lesssim 1$, as required.

We now adapt the preceding arguments to the multivariate case. For any $x = (x_1,\ldots,x_d) \in [0,1]^d$ we define $b^K(x) = \otimes_{l=1}^d b^{K_0}(x_l)$ where $b^{K_0}(x_l) = (\varphi_{J,0}(x_l),\ldots,\varphi_{J,2^J-1}(x_l))'$ and $K_0 = 2^J$.

Recall that $K = 2^{Jd}$. For any $x = (x_1,\ldots,x_d) \in [0,1]^d$ we have

$$\begin{aligned}
\|b^K(x)\|_{\ell^1} &= \prod_{l=1}^d \|b^{K_0}(x_l)\|_{\ell^1} && (129)\\
&\leq \Big((2N)2^{J/2}\max\{\|\varphi\|_\infty, \|\varphi_0^l\|_\infty,\ldots,\|\varphi_{N-1}^l\|_\infty, \|\varphi_{-1}^r\|_\infty,\ldots,\|\varphi_{-N}^r\|_\infty\}\Big)^d && (130)\\
&= \lesssim (2^{J/2})^d \; = \; \sqrt{K}\,. && (131)
\end{aligned}$$

With slight abuse of notation we let $X_{i1},\ldots,X_{id}$ denote the d elements of X_i. For $0 \leq k_1,\ldots,k_d \leq 2^J-1$, Fubini's theorem and a change of variables yields

$$\begin{aligned}
E\left[\left|\prod_{l=1}^d \varphi_{J,k}(X_{il})\right|\right] &\leq \sup_{x\in[0,1]^d} f_X(x) \int_{\mathbb{R}^d} \left(\prod_{l=1}^d |\varphi_{J,k_l}(x_l)|\right) d\mu(x_1,\ldots,x_d) && (132)\\
&= \sup_{x\in[0,1]^d} f_X(x) \prod_{l=1}^d \left(\int_{\mathbb{R}} |\varphi_{J,k_l}(x_l)|\,\mathrm{d}\mu(x_l)\right) && (133)\\
&\lesssim (2^{-J/2})^d \; = \; K^{-1/2}\,. && (134)
\end{aligned}$$

This verifies the conditions of Lemma 7.1 and hence $\|P_K\|_\infty \lesssim \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty}$.

The tensor product basis is an orthonormal basis with respect to Lebesgue measure on $[0,1]^d$ (by Fubini's theorem). Therefore, the minimum and maximum eigenvalues of $E[b^K(X_i)b^K(X_i)']$ may be shown to be bounded below and above by $\inf_{x\in[0,1]^d} f_X(x)$ and $\sup_{x\in[0,1]^d} f_X(x)$ as in the univariate case. Again, compact support of the $\varphi_{J,k}$ and the tensor product construction implies that $E[b^K(X_i)b^K(X_i)']$ is banded: $(E[b^K(X_i)b^K(X_i)'])_{k,j} = 0$ whenever $|k-j| > (2N-1)^d$. This verifies the conditions of Lemma 7.2 for $E[b^K(X_i)b^K(X_i)']$. It follows by Lemma 7.2 that $\| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|_{\ell^\infty} \lesssim 1$, as required. ■

Theorem 7.1 *Under conditions stated in Theorem 5.1, we have $\|P_{K,n}\|_\infty \lesssim 1$ wpa1 provided the following are satisfied:*

(i) $\| (B'B/n) - E[b^K(X_i)b^K(X_i)']\| = o_p(1)$, *and*

(ii) $\max_{1\leq k\leq K} \left| \frac{\frac{1}{n}\sum_{i=1}^n |b_{Kk}(X_i)| - E[|b_{Kk}(X_i)|]}{E[|b_{Kk}(X_i)|]}\right| = o_p(1)$.

Proof of Theorem 7.1. Condition (ii) $\max_{1\leq k\leq K} \frac{\frac{1}{n}\sum_{i=1}^n |b_{Kk}(X_i)| - E[|b_{Kk}(X_i)|]}{E[|b_{Kk}(X_i)|]} = o_p(1)$ implies

$$\max_{1\leq k\leq K} \frac{1}{n}\sum_{i=1}^n |b_{Kk}(X_i)| \lesssim \max_{1\leq k\leq K} \|b_{Kk}\|_{L^1(X)} \lesssim K^{-1/2} \tag{135}$$

where the final inequality is by the proof of Theorem 5.1. Moreover, $\sup_x \|b^K(x)\|_{\ell^1} \lesssim \sqrt{K}$ by the proof of Theorem 5.1. It follows analogously to Lemma 7.1 that $\|P_{K,n}\|_\infty \lesssim \| (B'B/n)^{-1} \|_\infty$ wpa1 (noting that $B'B/n$ is invertible wpa1 because $\| (B'B/n) - E[b^K(X_i)b^K(X_i)']\| = o_p(1)$ and $\lambda_{K,n} \lesssim 1$).

Condition (i) $\| (B'B/n) - E[b^K(X_i)b^K(X_i)']\| = o_p(1)$ implies (1) $\lambda_{\min}(B'B/n) \gtrsim \lambda_{\min}(E[b^K(X_i)b^K(X_i)'])$, (2) $\lambda_{\max}(B'B/n) \lesssim \lambda_{\max}(E[b^K(X_i)b^K(X_i)'])$, and (3) $\| (B'B/n)^{-1} \| \lesssim \| \left(E[b^K(X_i)b^K(X_i)']\right)^{-1} \|$ all hold wpa1. Moreover, $\lambda_{\min}(E[b^K(X_i)b^K(X_i)']) \gtrsim 1$ and $\lambda_{\max}(E[b^K(X_i)b^K(X_i)']) \lesssim 1$ by the proof of Theorem 5.1. It follows by Lemma 7.2 that $\| (B'B/n)^{-1} \|_{\ell^\infty} \lesssim 1$ wpa1, as required. ■

Proof of Theorem 5.2. Condition (i) of Theorem 7.1 is satisfied because $\lambda_{K,n} \lesssim 1$ and the condition $\|(\widetilde{B}'\widetilde{B}/n) - I_K\| = o_p(1)$ under the conditions on K (see Lemma 2.1 for the i.i.d. case and Lemma 2.2 for the weakly dependent case). Therefore,

$$\begin{aligned} \|(B'B/n) - E[b^K(X_i)b^K(X_i)']\| &\leq [\lambda_{\min}(E[b^K(X_i)b^K(X_i)'])]^{-1}\|(\widetilde{B}'\widetilde{B}/n) - I_K\| \qquad (136) \\ &\lesssim \|(\widetilde{B}'\widetilde{B}/n) - I_K\| \;=\; o_p(1)\,. \qquad (137) \end{aligned}$$

It remains to verify condition (ii) of Theorem 7.1. Let $b_{K1} = \varphi^d_{J,0}, \ldots, b_{KK} = \varphi^d_{J,2^J-1}$ with $K = 2^{dJ}$ as in the proof of Theorem 5.1. Similar arguments to the proof of Theorem 5.1 yield the bounds $\|b_{Kk}\|_\infty \lesssim 2^{dJ/2} = \sqrt{K}$ uniformly for $1 \leq k \leq K$. Let $f_X(x)$ denote the density of X. Then by

$\inf_{x\in[0,1]^d}|f_X(x)| > 0$ and Fubini's theorem

$$\begin{aligned} E[|b_{Kk}(X)|] &\geq \left(\inf_{x\in[0,1]^d} f_X(x)\right)\int_{[0,1]^d}\left(\prod_{l=1^d}|\varphi_{J,k_l}(x_l)|\right)\mathrm{d}\mu(x_1,\ldots,x_d) && (138)\\ &= \left(\inf_{x\in[0,1]^d} f_X(x)\right)\prod_{l=1}^{d}\left(\int_{[0,1]}|\varphi_{J,k_l}(x_l)|\,\mathrm{d}\mu(x_l)\right). && (139)\end{aligned}$$

A change of variables argument yields $\int_{[0,1]}|\varphi_{J,k_l}(x_l)|\,\mathrm{d}\mu(x_l) \gtrsim 2^{-J/2}$ uniformly for $0 \leq k_l \leq 2^J - 1$, and so $E[|b_{Kk}(X)|] \gtrsim 2^{-dJ/2} = K^{-1/2}$ uniformly for $1 \leq k \leq K$.

For the **i.i.d. case**, define $b^*_{Kk}(X_i) = n^{-1}(|b_{Kk}(X_i)| - E[|b_{Kk}(X_i)|])/(E[|b_{Kk}(X_i)|])$ for each $1 \leq k \leq K$. It may be deduced from the preceding bounds and the fact that $E[b_{Kk}(X_i)^2] \asymp 1$ that $\|b^*_{Kk}\|_\infty \lesssim K/n$ and $E[b^*_{Kk}(X_i)^2] \lesssim K/n^2$. By the union bound and Bernstein's inequality (see, e.g., pp. 192–193 of Pollard (1984)) we obtain, for any $t > 0$,

$$\begin{aligned} &\mathbb{P}\left(\max_{1\leq k\leq K}\left|\frac{\frac{1}{n}\sum_{i=1}^n |b_{Kk}(X_i)| - E[|b_{Kk}(X_i)|]}{E[|b_{Kk}(X_i)|]}\right| > t\right)\\ \leq\ & \sum_{k=1}^{K}\mathbb{P}\left(\left|\frac{\frac{1}{n}\sum_{i=1}^n |b_{Kk}(X_i)| - E[|b_{Kk}(X_i)|]}{E[|b_{Kk}(X_i)|]}\right| > t\right) && (140)\\ \leq\ & 2\exp\left\{\log K - \frac{t^2/2}{c_1K/n + c_2K/nt}\right\} && (141)\end{aligned}$$

where c_1 and c_2 are finite positive constants independent of t. The right-hand side of (141) vanishes as $n \to \infty$ since $K\log n/n = o(1)$.

For the **beta-mixing regressors case**, we may extend the proof for the i.i.d. case using a coupling argument similar to the proof of Theorem 4.2 to deduce that

$$\begin{aligned} &\mathbb{P}\left(\max_{1\leq k\leq K}\left|\frac{\frac{1}{n}\sum_{i=1}^n |b_{Kk}(X_i)| - E[|b_{Kk}(X_i)|]}{E[|b_{Kk}(X_i)|]}\right| > t\right)\\ \lesssim\ & \frac{n}{q}\beta(q) + \exp\left\{\log n - \frac{t^2}{c_1Kq/n + c_2Kq/nt}\right\}. && (142)\end{aligned}$$

The right-hand side is $o(1)$ provided $\frac{n}{q}\beta(q) = o(1)$ and $(qK\log n)/n = o(1)$. Both these conditions are satisfied under the conditions on K, taking $q = \gamma^{-1}\log n$ in the exponentially β-mixing case and $q \asymp n^{\gamma/(1+\gamma)}$ in the algebraically β-mixing case. ■

References

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2014). Some new asymptotic theory for least squares series: Pointwise and uniform results. Preprint, arXiv:1212.0442v3 [stat.ME].

Berbee, H. (1987). Convergece rates in the strong law for bounded mixing sequences. *Probability Theory and Related Fields 74*, 225–270.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys 2*, 107–144.

Cattaneo, M. D. and M. H. Farrell (2013). Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics 174*(2), 127–143.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, Part B, Chapter 76, pp. 5549–5632. Elsevier.

Chen, X. (2013). Penalized sieve estimation and inference of semi-nonparametric dynamic models: A selective review. In M. A. Daron Acemoglu and E. Dekel (Eds.), *Advances in Economics and Econometrics. Tenth World Congress, Volume III.* Cambridge University Press, New York.

Chen, X. and T. M. Christensen (2013). Optimal uniform convergence rates for sieve nonparametric instrumental variables regression. Cemmap working paper CWP56/13.

Chen, X. and J. Z. Huang (2003). Sup norm convergence rate and asymptotic normality for a class of linear sieve estimators. Technical report, New York University and University of Pennsylvania.

Chen, X. and Z. Liao (2014). Sieve m inference on irregular parameters. *Journal of Econometrics 182*, 70–86.

Chen, X., Z. Liao, and Y. Sun (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics 178*, 639–658.

Chen, X. and D. Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica 80*(1), 277–321.

Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometrica 66*(2), 289–314.

Cohen, A., I. Daubechies, and P. Vial (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis 1*, 54–81.

De Boor, C. (2001). *A Practical Guide to Splines.* Springer-Verlag, New York.

de Jong, R. M. (2002). A note on "convergence rates and asymptotic normality for series estimators": Uniform convergence rates. *Journal of Econometrics 111*(1), 1–9.

Demko, S., W. F. Moss, and P. W. Smith (1984). Decay rates for inverses of band matrices. *Mathematics of Computation 43*, 491–499.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive Approximation.* Springer-Verlag, Berlin.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer-Verlag, New York.

Freedman, D. A. (1975). On tail probabilities for martingales. *The Annals of Probability 3*(1), 100–118.

Hansen, B. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*, 726–748.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics 26*(1), 242–272.

Huang, J. Z. (2003a). Asymptotics for polynomial spline regression under weak conditions. *Statistics & Probability Letters 65*(3), 207–216.

Huang, J. Z. (2003b). Local asymptotics for polynomial spline regression. *The Annals of Statistics 31*(5), 1600–1635.

Huang, J. Z. and H. Shen (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian Journal of Statistics 31*(4), 515–534.

Huang, J. Z. and L. Yang (2004). Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 463–477.

Johnstone, I. M. (2013). Gaussian estimation: Sequence and wavelet models. Manuscript.

Kristensen, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous dependent data. *Econometric Theory 25*, 1433–1445.

Lee, J. and P. Robinson (2013). Series estimation under cross-sectional dependence. Preprint, London School of Economics.

Li, Q., C. Hsiao, and J. Zinn (2003). Consistent specifiation tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics 112*, 295–325.

Li, Q. and J. S. Racine (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis 17*, 571–599.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *The Annals of Probability 2*(4), 620–628.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag New York.

Robinson, P. (1989). Hypothesis testing in semiparametric and nonparametric models for econometric time series. *The Review of Economic Studies 56*(4), 511–534.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, Cambridge.

Song, K. (2008). Uniform convergence of series estimators over function spaces. *Econometric Theory 24*, 1463–1499.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*(4), 1040–1053.

Tjøstheim, D. and B. H. Auestad (1994). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association 89*(428), 1398–1409.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics 12*, 389–434.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer, New York.