cemmap

centre for microdata methods and practice

On the asymptotic theory for least squares series: pointwise and uniform results

Alexandre Belloni
Victor Chernozhukov
Denis Chetverikov
Kengo Kato

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP73/13

An ESRC Research Centre

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE **Tel:** 44 (0)20 7291 4800 **Email:** info@cemmap.ac.uk **www.cemmap.ac.uk**

ON THE ASYMPTOTIC THEORY FOR LEAST SQUARES SERIES: POINTWISE AND UNIFORM RESULTS

ALEXANDRE BELLONI, VICTOR CHERNOZHUKOV, DENIS CHETVERIKOV, AND KENGO KATO

Abstract. In this work we consider series estimators for the conditional mean in light of three new ingredients: (i) sharp LLNs for matrices derived from the non-commutative Khinchin inequalities, (ii) bounds on the Lebesgue factor that controls the ratio between the L^∞ and L^2-norms, and (iii) maximal inequalities for processes whose entropy integrals diverge at some rate.

These technical tools allow us to contribute to the series literature, specifically the seminal work of Newey (1997), as follows. First, we weaken considerably the condition on the number k of approximating functions used in series estimation from the typical $k^2/n \to 0$ to $k/n \to 0$, up to log factors, which was available only for spline and local polynomial partition series before. Second, under the same weak conditions we derive L^2 rates and pointwise central limit theorems results when the approximation error vanishes. Under an incorrectly specified model, i.e. when the approximation error does not vanish, analogous results are also shown. Third, under stronger conditions we derive uniform rates and functional central limit theorems that hold if the approximation error vanishes or not. That is, we derive the strong approximation for the entire estimate of the nonparametric function. Finally, we derive uniform rates and inference results for linear functionals of interest of the conditional expectation function such as its partial derivative or conditional average partial derivative.

1. Introduction

Series estimators have been playing a central role on various fields. In econometric applications it is common that the exact form of a conditional expectation is unknown and having a flexible functional form can lead to improvements over a pre-specified functional form. Series estimation offers exactly that by approximating the unknown function based on k basic functions, where k is allowed to grow with the sample size n to balance the trade off between variance and bias.

Several asymptotic properties of series estimators have been investigated in the literature. The focus has been on convergence rates and asymptotic normality results (see Andrews,

Date: First version: May 2006, This version of December 27, 2013.

1991; Eastwood and Gallant, 1991; Gallant and Souza, 1991; Newey, 1997; Huang, 2003b; Chen, 2007; Cattaneo and Farell, 2013, and the references therein).

This work revisits the topic by making use of three critical ingredients:

1. The sharp LLNs for matrices derived from the non-commutative Khinchin inequalities.
2. The sharp bounds on the Lebesgue factor that controls the ratio between the L^∞ and L^2-norms of the least squares approximation of functions (which is bounded or grows like a $\log k$ in many cases).
3. Sharp maximal inequalities for processes whose entropy integrals diverge at some rate.

To the best of our knowledge, our results are the first applications of the first ingredient to statistical estimation problems. Regarding the second ingredient, it has already been used by Huang (2003a) but for splines only. The third ingredient was derived to allow for weak moment conditions. All of these ingredients are critical for generating sharp results.

This approach allows us to contribute to the series literature in several directions. First, we weaken considerably the condition on the number k of approximating functions used in series estimation from the typical $k^2/n \to 0$ (see Newey, 1997) to

$$k/n \to 0 \text{ (up to logs)}$$

for bounded or local bases which was previously available only for spline series (Huang, 2003a; Stone, 1994) and local polynomial partition series (Cattaneo and Farell, 2013). An example of a bounded basis is Fourier series; examples of local bases are spline, wavelet, and local polynomial partition series. To be more specific, for such bases we require $k \log n/n \to 0$. Note that the last condition is similar to that on the bandwidth value required for local polynomial (kernel) regression estimators ($h^{-d} \log n/n \to 0$ where h is the bandwidth value). Second, under the same weak conditions we derive L^2 rates and pointwise central limit theorems results when the approximation error vanishes. Under a misspecified model, i.e. when the approximation error does not vanish, analogous results are also shown. Third, under stronger conditions we derive uniform rates that hold if the approximation error vanishes or not. An important contribution here is that we show that the series estimator achieves the optimal uniform rate of convergence under quite general conditions. Previously, the same result was shown only for local polynomial partition series estimator (Cattaneo and Farell, 2013). In addition, we derive a functional central limit theorems. By the functional central limit theorem we mean here that the entire estimate of the nonparametric function is uniformly close to a Gaussian process that can change with n. That is, we derive the strong approximation for the entire estimate of the nonparametric function.

Another set of results established here pertains to the estimation and inference methods for linear functionals θ of the conditional mean function $g : \mathcal{X} \to \mathbb{R}$. Examples of linear functionals θ of interest include, for x_j denoting the j-th component of x and x_{-j} denoting all components of x excluding x_j,

1. the partial derivative: $\theta(x) = \partial_{x_j} g(x)$;
2. the average partial derivative: $\theta = \int \partial_{x_j} g(x) d\mu(x)$;
3. the conditional average partial derivative: $\theta(x_{-j}) = \int \partial_{x_j} g(x) d\mu(x_j | x_{-j})$.

where the measure μ entering the definitions above are taken as known; the result can be extended to include estimated measures. Under weak conditions we derive pointwise results for rates of convergence, large sample distributions and inference methods based on the Gaussian approximation. Under stronger conditions we derive new strong approximation for the entire estimate of the nonparametric functional. Specifically, we derive results for uniform rates of convergence, large sample distributions and inference methods based on the Gaussian approximation.

Notation. In what follows, all parameter values are indexed by the sample size n, but we omit the index whenever this does not cause confusion. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ℓ_2-norm of a vector v is denoted by $\|v\|$, while for a matrix Q the maximum eigenvalue is denoted by $\|Q\|$. We also use standard notation in the empirical process literature,

$$\mathbb{E}_n[f] = \mathbb{E}_n[f(w_i)] = \sum_{i=1}^{n} f(w_i)/n \text{ and } \mathbb{G}_n[f] = \mathbb{G}_n[f(w_i)] = \sum_{i=1}^{n} f(w_i)/\sqrt{n}$$

and we use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. Moreover, for two random variables X, Y we say that $X =_d Y$ if they have the same probability distribution.

2. Set-Up

Throughout the paper, we consider a sequence of models, indexed by the sample size n,

$$y_i = g(x_i) + \epsilon_i, \quad E[\epsilon_i | x_i] = 0, \quad x_i \in \mathcal{X} \subseteq \mathbb{R}^d, \quad i = 1, \ldots, n, \tag{2.1}$$

where y_i is a response variable, x_i a vector of covariates (basic regressors), ϵ_i noise, and $x \mapsto g(x) = E[y_i | x_i = x]$ a regression (conditional mean) function; that is, we consider a triangular array of models with $y_i = y_{i,n}$, $x_i = x_{i,n}$, $\epsilon_i = \epsilon_{i,n}$, and $g = g_n$. We assume that $g \in \mathcal{G}$ where $\mathcal{G}$ is some class of functions. Since we consider a sequence of models indexed by n, we allow the function class $\mathcal{G} = \mathcal{G}_n$, where the regression function g belongs to, to depend on n as well. In addition, we allow $\mathcal{X} = \mathcal{X}_n$ to depend on n but we assume for the sake of simplicity that the diameter of $\mathcal{X}$ is bounded from above uniformly over n (dropping the uniform boundedness condition is possible in expense of more technicalities; for example, without uniform boundedness condition, we would have an additional term $\log \mathrm{diam}(\mathcal{X})$ in (4.19) and (4.21) of Lemma 4.2). We denote $\sigma_i^2 = E[\epsilon_i^2 | x_i]$, $\bar{\sigma}^2 := \sup_{x \in \mathcal{X}} E[\epsilon_i^2 | x_i = x]$, and $\underline{\sigma}^2 := \inf_{x \in \mathcal{X}} E[\epsilon_i^2 | x_i = x]$. For notational convenience, we omit indexing by n where it does not lead to confusion.

Assumption A.1 (Sample) *For each n, random vectors $(y_i, x_i')'$, $i = 1, \ldots, n$, are i.i.d. and satisfy (2.1).*

We approximate the function $x \mapsto g(x)$ by linear forms $x \mapsto p(x)'b$, where

$$x \mapsto p(x) := (p_1(x), \ldots, p_k(x))'$$

is a vector of approximating functions that can change with n. We denote the regressors as

$$p_i := p(x_i) := (p_1(x_i), \ldots, p_k(x_i))'.$$

Throughout the paper, we assume that the number of series terms k is chosen so that $\log k \lesssim \log n$. The next assumption imposes regularity conditions on the regressors.

Assumption A.2 (Eigenvalues) *Uniformly over all n, eigenvalues of $Q := E[p_i p_i']$ are bounded above and away from zero.*

Condition A.2 imposes the restriction that p_i's are not too co-linear. Given this assumption, it is without loss of generality to impose the following normalization:

Normalization. *To simplify notation, we normalize $Q = I$, but we shall treat Q as unknown, that is we deal with random design.*

The following proposition establishes a simple sufficient condition for A.2 based on orthonormal bases with respect to some measure.

Proposition 2.1 (Stability of Bounds on Eigenvalues)**.** *Assume that $x_i \sim F$ where F is a probability measure on $\mathcal{X}$, and that the regressors p_i's are orthonormal on $(\mathcal{X}, \mu)$ for some measure μ. Then A.2 is satisfied if $dF/d\mu$ is bounded above and away from zero.*

It is well known that the least squares parameter β is defined by

$$\beta := \arg\min_{b\in\mathbb{R}^k} E\left[(y_i - p_i'b)^2\right],$$

which by (2.1) also implies that $\beta = \beta_g$ where β_g is defined by

$$\beta_g := \arg\min_{b\in\mathbb{R}^k} E\left[(g(x_i) - p_i'b)^2\right].$$

We call $x \mapsto g(x)$ the target function and $x \mapsto g_k(x) = p(x)'\beta$ the surrogate function. In this setting, the surrogate function provides the best linear approximation to the target function.

For all $x \in \mathcal{X}$, let

$$r(x) := r_g(x) := g(x) - p(x)'\beta_g \tag{2.2}$$

denote the approximation error at the point x, and let

$$r_i := r(x_i) = g(x_i) - p(x_i)'\beta_g$$

denote the approximation error for the observation i. Using this notation, we obtain a many regressors model

$$y_i = p_i'\beta + u_i, \quad E[u_i x_i] = 0, \quad u_i := r_i + \epsilon_i.$$

The least squares estimator of β is

$$\widehat{\beta} := \arg\min_{b\in\mathbb{R}^k} \mathbb{E}_n\left[(y_i - p_i'b)^2\right], \tag{2.3}$$

which induces the estimator $\widehat{g}(x) := p(x)'\widehat{\beta}$ for the target function $g(x)$. Then it follows from (2.2) that we can decompose the error in estimating the target function as

$$\widehat{g}(x) - g(x) = p(x)'(\widehat{\beta} - \beta) - r(x),$$

where the first term on the right-hand side is the estimation error and the second term is the approximation error.

We are also interested in various linear functionals θ of the conditional mean function. As discussed in the introduction, examples include the partial derivative function, the average partial derivative function, and the conditional average partial derivative. Importantly, in each example above we could be interested in estimating $\theta = \theta(w)$ simultaneously for many values $w \in \mathcal{W}$. Let $\mathcal{I} \subset \mathcal{W}$ denote the set of indices of interest. By the linearity of the series approximations, the above parameters can be seen as linear functions of the least squares coefficients β up to an approximation error, that is

$$\theta(w) = \ell_\theta(w)'\beta + r_\theta(w), \quad w \in \mathcal{I}, \tag{2.4}$$

where $\ell_\theta(w)'\beta$ is the series approximation, with $\ell_\theta(w)$ denoting the k-vector of loadings on the coefficients, and $r_\theta(w)$ is the remainder term, which corresponds to the approximation error. Indeed, the decomposition (2.4) arises from the application of different linear operators $\mathcal{A}$ to the decomposition $g(\cdot) = p(\cdot)'\beta + r(\cdot)$ and evaluating the resulting functions at w:

$$(\mathcal{A}g(\cdot))\,[w] = (\mathcal{A}p(\cdot))\,[w]'\beta + (\mathcal{A}r(\cdot))\,[w]. \tag{2.5}$$

Examples of the operator $\mathcal{A}$ corresponding to the cases enumerated in the introduction are given by, respectively,

1. a differential operator: $(\mathcal{A}f)[x] = (\partial_{x_j} f)[x]$, so that

$$\ell_\theta(x) = \partial_{x_j} p(x), \quad r_\theta(x) = \partial_{x_j} r(x);$$

2. an integro-differential operator: $\mathcal{A}f = \int \partial_{x_j} f(x) d\mu(x)$, so that

$$\ell_\theta = \int \partial_{x_j} p(x) d\mu(x), \quad r_\theta = \int \partial_{x_j} r(x) d\mu(x);$$

3. a partial integro-differential operator: $(\mathcal{A}f)[x_{-j}] = \int \partial_{x_j} f(x) d\mu(x_j|x_{-j})$, so that

$$\ell_\theta(x_{-j}) = \int \partial_{x_j} f(x) d\mu(x_j|x_{-j}), \quad r_\theta(x_{-j}) = \int \partial_{x_j} r(x) d\mu(x_j|x_{-j})$$

   where x_{-j} denotes all components of x excluding x_j.

For notational convenience, we use the formulation (2.4) in the analysis, instead of the motivational formulation (2.5).

We shall provide the inference tools that will be valid for inference on the series approximation

$$\ell_\theta(w)'\beta, \quad w \in \mathcal{I}.$$

If the approximation error $r_\theta(w)$, $w \in \mathcal{I}$, is small enough as compared to the estimation error, these tools will also be valid for inference on the functional of interest

$$\theta(w), \quad w \in \mathcal{I}.$$

In this case, the series approximation is an important intermediary target, whereas the functional θ is the ultimate target. The inference will be based on the plug-in estimator $\widehat{\theta}(w) := \ell_\theta(w)'\widehat{\beta}$ of the the series approximation $\ell_\theta(w)'\beta$ and hence of the final target $\theta(w)$.

3. Approximation Properties of Least Squares

Next we consider approximation properties of the least squares estimator. Not surprisingly, approximation properties must rely on the particular choice of approximating functions. At this point it is instructive to consider particular examples of relevant bases used in the literature. For each example, we state a bound on the following quantity:

$$\xi_k := \sup_{x \in \mathcal{X}} \|p(x)\|.$$

This quantity will play a key role in our analysis.

Example 3.1 (Polynomial series)**.** Let $\mathcal{X} = [0,1]$ and consider a polynomial series given by

$$\widetilde{p}(x) = (1, x, x^2, ..., x^{k-1})'.$$

In order to reduce collinearity problems, it is useful to orthonormalize the polynomial series with respect to the Lebesgue measure on $[0,1]$ to get the Legendre polynomial series

$$p(x) = (1, \sqrt{3}x, \sqrt{5/4}(3x^2 - 1), ...)'.$$

The Legendre polynomial series satisfies

$$\xi_k \lesssim k;$$

see, for example, Newey (1997). □

Example 3.2 (Fourier series)**.** Let $\mathcal{X} = [0,1]$ and consider a Fourier series given by

$$p(x) = (1, \cos(2\pi jx), \sin(2\pi jx), j = 1, 2, ..., k/2 - 1)',$$

for k even. Fourier series is orthonormal with respect to the Lebesgue measure on $[0,1]$ and satisfies

$$\xi_k \lesssim \sqrt{k},$$

which follows trivially from the fact that every element of $p(x)$ is bounded in absolute value by one. □

Example 3.3 (Spline series)**.** Let $\mathcal{X} = [0,1]$ and consider a linear regression spline series, or regression spline series of order 1, with a finite number of equally spaced knots $l_1, \dots, l_{k-2}$ in $\mathcal{X}$:

$$\widetilde{p}(x) = (1, x, (x - l_1)_+, \dots, (x - l_{k-2})_+)'.$$

The cubic regression spline series, or regression spline series of order 3, with a finite number of equally spaced knots $l_1, \dots, l_{k-4}$:

$$\widetilde{p}(x) = (1, x, x^2, x^3, (x - l_1)_+^3, ..., (x - l_{k-4})_+^3)'.$$

The function $x \mapsto \widetilde{p}(x)'b$ constructed using regression splines of order s_0 is $s_0 - 1$ times continuously differentiable in x for any b. Instead of regression splines, it is often helpful to consider B-splines $p(x) = (p_1(x), \dots, p_k(x))'$, which are linear transformations of the regression splines with lower multicellularity; see De Boor (2001) for the introduction to the theory of splines. B-splines are local in the sense that each B-spline $p_j(x)$ is supported on the interval $[l_{j(1)}, l_{j(2)}]$ for some $j(1)$ and $j(2)$ satisfying $j(2) - j(1) \lesssim 1$ and there is at most $s_0 + 1$ non-zero B-splines on each interval $[l_{j-1}, l_j]$. From this property of B-splines, it is easy to see that B-spline series satisfies

$$\xi_k \lesssim \sqrt{k};$$

see, for example, Newey (1997). □

Example 3.4 (Cohen-Deubechies-Vial wavelet series)**.** Let $\mathcal{X} = [0, 1]$ and consider Cohen-Deubechies-Vial (CDV) wavelet bases; see Section 4 in Cohen et al. (1993) and Chapter 7.5 in Mallat (2009) for details on CDV wavelet bases. CDV wavelet bases is a class of orthonormal with respect to the Lebesgue measure on $[0, 1]$ bases. Each such basis is built from a Daubechies scaling function ϕ (defined on $\mathbb{R}$) and the wavelet ψ of order s_0 starting from a fixed resolution level J_0 such that $2^{J_0} \geq 2s_0$. The functions ϕ and ψ are supported on $[0, 2s_0 - 1]$ and $[-s_0 + 1, s_0]$, respectively. Translate ϕ so that it has the support $[-s_0 + 1, s_0]$. Let

$$\phi_{l,m}(x) = 2^{l/2}\phi(2^l x - m),\ \psi_{l,m}(x) = 2^{l/2}\psi(2^l x - m),\ l, m \geq 0.$$

Then we can create the CDV wavelet basis from these functions as follows. Take all the functions $\phi_{J_0,m}, \psi_{l,m}$, $l \geq J_0$, that are supported in the interior of $[0, 1]$ (these are functions $\phi_{J_0,m}$ with $m = s_0 - 1, \dots, 2^{J_0} - s_0$ and $\psi_{l,m}$ with $m = s_0 - 1, \dots, 2^l - s_0, l \geq J_0$). Denote these functions $\widetilde{\phi}_{J_0,m}$, $\widetilde{\psi}_{l,m}$. To this set of functions, add suitable boundary corrected functions $\widetilde{\phi}_{J_0,0}, \dots, \widetilde{\phi}_{J_0,s_0-2}$, $\widetilde{\phi}_{J_0,2^{J_0}-s_0+1}, \dots, \widetilde{\phi}_{J_0,2^{J_0}-1}$, $\widetilde{\psi}_{l,0}, \dots, \widetilde{\psi}_{l,s_0-2}$, $\widetilde{\psi}_{l,2^{J_0}-s_0+1}, \dots, \widetilde{\psi}_{l,2^{J_0}-1}$, $l \geq J_0$, so that $\{\widetilde{\phi}_{J_0,m}\}_{0 \leq m < 2^{J_0}} \cup \{\widetilde{\psi}_{l,m}\}_{0 \leq m < 2^l, l \geq J_0}$ forms an orthonormal basis of $L^2[0, 1]$. Suppose that $k = 2^J$ for some $J > J_0$. Then the CDV series takes the form:

$$p(x) = (\widetilde{\phi}_{J_0,0}(x), \dots, \widetilde{\phi}_{J_0,2^{J_0}-1}(x), \widetilde{\psi}_{J_0,0}(x), \dots, \widetilde{\psi}_{J-1,2^{J-1}-1}(x))'.$$

This series satisfies

$$\xi_k \lesssim \sqrt{k}.$$

This bound can be derived by the same argument as that for B-splines. CDV wavelet bases is a flexible tool to approximate many different function classes. □

Example 3.5 (Local polynomial partition series)**.** Let $\mathcal{X} = [0,1]$ and define a local polynomial partition series as follows. Let s_0 be a nonnegative integer. Partition $\mathcal{X}$ as $0 = l_0 < l_1, \dots < l_{\widetilde{k}-1} < l_{\widetilde{k}} = 1$ where $\widetilde{k} := k/(s_0+1)$. For $j = 1, \dots, \widetilde{k}$, define $\delta_j : [0,1] \to \{0,1\}$ by $\delta_j(x) = 1$ if $x \in (l_{j-1}, l_j]$ and 0 otherwise. For $j = 1, \dots, k$, define

$$\widetilde{p}_j(x) := \delta_{[j/(s_0+1)]+1}(x) x^{j-1-(s_0+1)[j/(s_0+1)]}$$

for all $x \in \mathcal{X}$ where $[a]$ is the largest integer that is strictly smaller than a. Finally, define the local polynomial partition series $p_1(\cdot), \dots, p_k(\cdot)$ of order s_0 as an orthonormalization of $\widetilde{p}_1(\cdot), \dots, \widetilde{p}_k(\cdot)$ with respect to the Lebesgue (or some other) measure on $\mathcal{X}$. The local polynomial partition series estimator was analyzed in detail in Cattaneo and Farell (2013). Its properties are somewhat similar to those of local polynomial estimator of Stone (1982). When the partition $l_0, \dots, l_{\widetilde{k}}$ satisfies $p_j - p_{j-1} \asymp 1/\widetilde{k}$, that is there exist constants $c, C > 0$ independent of n and such that $c/\widetilde{k} \leq p_j - p_{j-1} \leq C/\widetilde{k}$ for all $j = 1, \dots, \widetilde{k}$, and the Lebesgue measure is used, the local polynomial partition series satisfies

$$\xi_k \lesssim \sqrt{k}.$$

This bound can be derived by the same argument as that for B-splines. □

Example 3.6 (Tensor Products)**.** Generalizations to multiple regressors are straightforward using tensor products of unidimensional series. Suppose that the basic regressors are

$$x_i = (x_{1i}, ..., x_{di})'.$$

Then we can create d series for each basic regressor. Then we take all interactions of functions from these d series, called tensor products, and collect them into vector of regressors p_i. If each series for a basic regressor has J terms, then the final regressor has dimension

$$k = J^d,$$

which explodes exponentially in the dimension d. The bounds on ξ_k in terms of k remain the same as in one-dimensional case. □

Each basis described in Examples 3.1-3.6 has different approximation properties which also depend on the particular class of functions $\mathcal{G}$. The following assumption captures the essence of this dependence into two quantities.

Assumption A.3 (Approximation) *For each n and k, there are finite constants c_k and ℓ_k such that for each $f \in \mathcal{G}$,*

$$\|r_f\|_{F,2} := \sqrt{\int_{x\in\mathcal{X}} r_f^2(x)dF(x)} \leq c_k \quad \textit{and} \quad \|r_f\|_{F,\infty} := \sup_{x\in\mathcal{X}} |r_f(x)| \leq \ell_k c_k.$$

Here r_f is defined by (2.2). We call ℓ_k the Lebesgue factor because of its relation to the Lebesgue constant defined in Section 3.2 below. Together c_k and ℓ_k characterize the approximation properties of the underlying class of functions under $L^2(\mathcal{X}, F)$ and uniform distances. Note that constants $c_k = c_k(\mathcal{G})$ and $\ell_k = \ell_k(\mathcal{G})$ are allowed to depend n but we omit indexing by n for simplicity of notation. Next we discuss primitive bounds on c_k and ℓ_k.

3.1. **Bounds on c_k.** In what follows, we call the case where $c_k \to 0$ as $k \to \infty$ the correctly specified case. In particular, if the series are formed from bases that span $\mathcal{G}$, then $c_k \to 0$ as $k \to \infty$. However, if series are formed from bases that do not span $\mathcal{G}$, then $c_k \not\to 0$ as $k \to \infty$. We call any case where $c_k \not\to 0$ the incorrectly specified (misspecified) case.

To give an example of the misspecified case, suppose that $d = 2$, so that $x = (x_1, x_2)'$ and $g(x) = g(x_1, x_2)$. Further, suppose that the researcher mistakenly assumes that $g(x)$ is additively separable in x_1 and x_2: $g(x_1, x_2) = g_1(x_1) + g(x_2)$. Given this assumption, the researcher forms the vector of approximating functions $p(x_1, x_2)$ such that each component of this vector depends either on x_1 or x_2 but not on both; see Newey (1997) and Newey et al. (1999) for the description of nonparametric series estimators of separately additive models. Then note that if the true function $g(x_1, x_2)$ is not separately additive, linear combinations $p(x_1, x_2)'b$ will not be able to accurately approximate $g(x_1, x_2)$ for any b, so that c_k does not converge to zero as $k \to \infty$. Since analysis of misspecified models plays an important role in econometrics, we include results both for correctly and incorrectly specified models.

To provide a bound on c_k, note that for any $f \in \mathcal{G}$,

$$\inf_b \|f - p'b\|_{F,2} \leq \inf_b \|f - p'b\|_{F,\infty},$$

so that it suffices to set c_k such that $c_k \geq \sup_{f\in\mathcal{G}} \inf_b \|f - p'b\|_{F,\infty}$. Next, the bounds for $\inf_b \|f - p'b\|_{F,\infty}$ are readily available from the Approximation Theory; see DeVore and Lorentz (1993). A typical example is based on the concept of s-smooth classes, namely Hölder classes of smoothness order s, $\Sigma_s(\mathcal{X})$. For $s \in (0, 1]$, the Hölder class of smoothness order s, $\Sigma_s(\mathcal{X})$, is defined as the set of all functions $f : \mathcal{X} \to \mathbb{R}$ such that for $C > 0$,

$$|f(x) - f(\widetilde{x})| \leq C(\sum_{j=1}^{d}(x_j - \widetilde{x}_j)^2)^{s/2}$$

for all $x = (x_1, \dots, x_d)'$ and $\widetilde{x} = (\widetilde{x}_1, \dots, \widetilde{x}_d)'$ in $\mathcal{X}$. The smallest C satisfying this property defines a norm of f in $\Sigma_s(\mathcal{X})$, which we denote by $\|f\|_s$. For $s > 1$, $\Sigma_s(\mathcal{X})$ can be defined as follows. For a d-tuple $\alpha = (\alpha_1, \dots, \alpha_d)$ of nonnegative integers, let

$$D^\alpha = \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d}.$$

Let $[s]$ denote the largest integer strictly smaller than s. Then $\Sigma_s(\mathcal{X})$ is defined as the set of all functions $f : \mathcal{X} \to \mathbb{R}$ such that f is $[s]$ times continuously differentiable and for some $C > 0$,

$$|D^\alpha f(x) - D^\alpha f(\widetilde{x})| \leq C(\sum_{j=1}^{d}(x_j - \widetilde{x}_j)^2)^{(s-[s])/2} \text{ and } |D^\beta f(x)| \leq C$$

for all $x = (x_1, \dots, x_d)'$ and $\widetilde{x} = (\widetilde{x}_1, \dots, \widetilde{x}_d)'$ in $\mathcal{X}$ and for all d-tuples $\alpha = (\alpha_1, \dots, \alpha_d)$ and $\beta = (\beta_1, \dots, \beta_d)$ of nonnegative integers satisfying $\alpha_1 + \dots + \alpha_d = [s]$ and $\beta_1 + \dots + \beta_d \leq [s]$. Again, the smallest C satisfying these properties defines a norm of f in $\Sigma_s(\mathcal{X})$, which we denote $\|f\|_s$.

If $\mathcal{G}$ is a set of functions f in $\Sigma_s(\mathcal{X})$ such that $\|f\|_s$ is bounded from above uniformly over all $f \in \mathcal{G}$, then we can take

$$c_k \lesssim k^{-s/d} \tag{3.6}$$

for the polynomial series and

$$c_k \lesssim k^{-(s \wedge s_0)/d}$$

for spline, CDV wavelet, and local polynomial partition series of order s_0. If in addition we assume that each element of $\mathcal{G}$ can be extended to a periodic function, then (3.6) also holds for the Fourier series. See, for example, Chen (2007) for references.

3.2. Bounds on ℓ_k.

We say that a least squares approximation by a particular series for the function class $\mathcal{G}$ is co-minimal if the Lebesgue factor ℓ_k is small in the sense of being a slowly varying function in k. A simple bound on ℓ_k, which is independent of $\mathcal{G}$, is established in the following proposition:

Proposition 3.1. *If c_k is chosen so that $c_k \geq \sup_{f \in \mathcal{G}} \inf_b \|f - p'b\|_{F,\infty}$, then Condition A.3 holds with some ℓ_k satisfying*

$$\ell_k \leq 1 + \xi_k.$$

The proof of this proposition is based on the ideas of Newey (1997) and is provided in the Appendix. The bound established in this proposition, however, is not sharp in most cases because ξ_k typically satisfies $\xi_k \gtrsim \sqrt{k}$; see our examples above. Much sharper bounds follow from Approximation Theory for some important cases. To apply these bounds, define the Lebesgue constant:

$$\widetilde{\ell}_k := \sup_{f \in \mathcal{G}} \frac{\|p'\beta_f\|_{F,\infty}}{\|f\|_{F,\infty}}.$$

The following proposition provides a bound on ℓ_k in terms of $\widetilde{\ell}_k$:

Proposition 3.2. *If c_k is chosen so that $c_k \geq \sup_{f \in \mathcal{G}} \inf_b \|f - p'b\|_{F,\infty}$, then Condition A.3 holds with*

$$\ell_k = 1 + \widetilde{\ell}_k.$$

Note that in all examples above, we provided c_k such that $c_k \geq \sup_{f \in \mathcal{G}} \inf_b \|f - p'b\|_{F,\infty}$, and so the results of Propositions 3.1 and 3.2 apply in our examples. We now provide bounds on $\widetilde{\ell}_k$.

Example 3.7 (Fourier series, continued)**.** For Fourier series on $\mathcal{X} = [0,1]$, $F = U(0,1)$, and $\mathcal{G} \subset C(\mathcal{X})$

$$\widetilde{\ell}_k \leq C_0 \log k + C_1,$$

where here and below C_0 and C_1 are some universal constants; see Zygmund (2002). □

Example 3.8 (Spline series, continued)**.** For B-spline series on $\mathcal{X} = [0,1]$, $F = U(0,1)$, and $\mathcal{G} \subset C(\mathcal{X})$

$$\widetilde{\ell}_k \leq C_0,$$

under approximately uniform placement of knots; see Huang (2003b). In fact, the result of Huang states that $\widetilde{\ell}_k \leq C$ whenever F has the pdf on $[0,1]$ bounded from above by $\bar{a}$ and below from zero by $\underline{a}$ where C is a constant that depends only on $\underline{a}$ and $\bar{a}$. □

Example 3.9 (Wavelet series, continued)**.** For CDV wavelet series on $\mathcal{X} = [0,1]$, $F = U(0,1)$, and $\mathcal{G} \subset C(\mathcal{X})$

$$\widetilde{\ell}_k \leq C_0.$$

The proof of this result was recently obtained by Chen and Christensen (2013) who extended the argument of Huang (2003b) for B-splines to cover wavelets. In fact, the result of Chen and Christensen also shows that $\widetilde{\ell}_k \leq C$ whenever F has the pdf on $[0,1]$ bounded from above by $\bar{a}$ and below from zero by $\underline{a}$ where C is a constant that depends only on $\underline{a}$ and $\bar{a}$. □

Example 3.10 (Local polynomial partition series, continued)**.** For local polynomial partition series on $\mathcal{X}$, $F = U(0,1)$, and $\mathcal{G} \subset C(\mathcal{X})$,

$$\widetilde{\ell}_k \leq C_0.$$

To prove this bound, note that first order conditions imply that for any $f \in \mathcal{G}$,

$$\beta_f = Q^{-1}E[p(x_1)f(x_1)] = E[p(x_1)f(x_1)].$$

Hence, for any $x \in \mathcal{X}$,

$$|p(x)'\beta_f| = |E[p(x)'p(x_1)f(x_1)]| \lesssim \|f\|_{F,\infty}$$

where the last inequality follows by noting that the sum $p(x)'p(x_1) = \sum_{j=1}^{k} p_j(x)p_j(x_1)$ contains at most $s_0 + 1$ nonzero terms, all nonzero terms in the sum are bounded by $\xi_k^2 \lesssim k$, and $p(x)'p(x_1) = 0$ outside of a set with probability bounded from above by $1/k$ up to a constant. The bound follows. □

Example 3.11 (Polynomial series, continued)**.** For Chebyshev polynomials with $\mathcal{X} = [0,1]$, $dF(x)/dx = 1/\sqrt{1-x^2}$, and $\mathcal{G} \subset C(\mathcal{X})$

$$\widetilde{\ell}_k \leq C_0 \log k + C_1.$$

This bound follows from a trigonometric representation of Chebyshev polynomials (see, for example, DeVore and Lorentz (1993)) and Example 3.7. □

Example 3.12 (Tailored Function Classes)**.** For each type of series approximations, it is possible to specify function classes for which the Lebesgue factors are small. □

Since the Lebesgue factor depends on the particular basis and on the underlying probability measure, it is important to have a stability result for the Lebesgue factor. The next proposition provides a bound on $\ell_k c_k$ for most functions in the α-ellipsoid class

$$\mathcal{F}(\alpha) = \left\{ \sum_{j \geq 1} p_j(\cdot) j^{-\alpha} \xi_j \ \ : \xi_j \in \mathbb{R}, j \geq 1 \right\}$$

according to a Gaussian measure on the coefficients ξ_j, $j \geq 1$, provided the basis functions are bounded and Lipschitz.

Proposition 3.3 (Generic Stability of Approximation Error for α-Ellipsoid)**.** *Consider an i.i.d. sequence of $N(0,1)$ coefficients ξ_j, $j \geq 1$, let $f = \sum_{j \geq 1} p_j(x) j^{-\alpha} \xi_j$ and let $\ell_k(f)$ and $c_k(f)$ denote respective the Lebesgue factor and the L_2 approximation rate associated with f.*

If the basis $\{p_j(x)\}_{j\geq 1}$ *obeys* $\sup_{x\in\mathcal{X}} |p_j(x)| \lesssim j^{\beta}$ *and* $\sup_{x\in\mathcal{X}} \|\nabla p_j(x)\| \leq M_j$*, with* $\beta > 0$ *and* $\sum_{j\geq 1} j^{-\alpha} M_j (\log j)^{1/2} < \infty$*, then*

$$P\left(\ell_k(f)c_k(f) \lesssim d^{1/2}\sqrt{(\alpha-\beta-1/2)\log k}k^{-\alpha+\beta+1/2}\right) = 1 - o(1) \text{ as } k \to \infty$$

for $\alpha > \beta + 1/2$*.*

In the case of orthogonal basis, most functions f in this class will have $c_k(f) = k^{-\alpha+\beta+1/2}$. Thus, Proposition 3.3 establishes that $\ell_k(f)$ is slow varying for those functions.

The following example illustrate the performance of the series estimator using different bases for a real data set.

Example 3.13 (Real Data)**.** Here $g(x)$ is the mean of log wage (y) conditional on education

$$x \in \{8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20\}.$$

The function $g(x)$ is computed using population data – the 1990 Census data for the U.S. men of prime age; see Angrist et al. (2006) for more details. So in this example, we know the true population function $g(x)$. We would like to know how well this function is approximated when common approximation methods are used to form the regressors. For simplicity we assume that x_i is uniformly distributed (otherwise we can weigh by the frequency). In population, least squares estimator solves the approximation problem: $\beta = \arg\min_b E[\{g(x_i) - p_i'b\}^2]$ for $p_i = p(x_i)$, where we form $p(x)$ as (a) linear spline (Figure 1, left) and (b) polynomial series (Figure 1, right), such that dimension of $p(x)$ is either $k = 3$ or $k = 8$. It is clear from these graphs that spline and polynomial series yield similar approximations.

In the table below, we also present L^2 and L^∞ norms of approximating errors:

	spline $k = 3$	spline $k = 8$	Poly $k = 3$	Poly $k = 8$
L^2 Error	0.12	0.08	0.12	0.05
L^∞ Error	0.29	0.17	0.30	0.12

We see from the table that in this example, the Lebesgue factor, which is defined as the ratio of L^∞ to L^2 errors, of the polynomial approximations is comparable to the Lebesgue factor of the spline approximations. □

4. Limit Theory

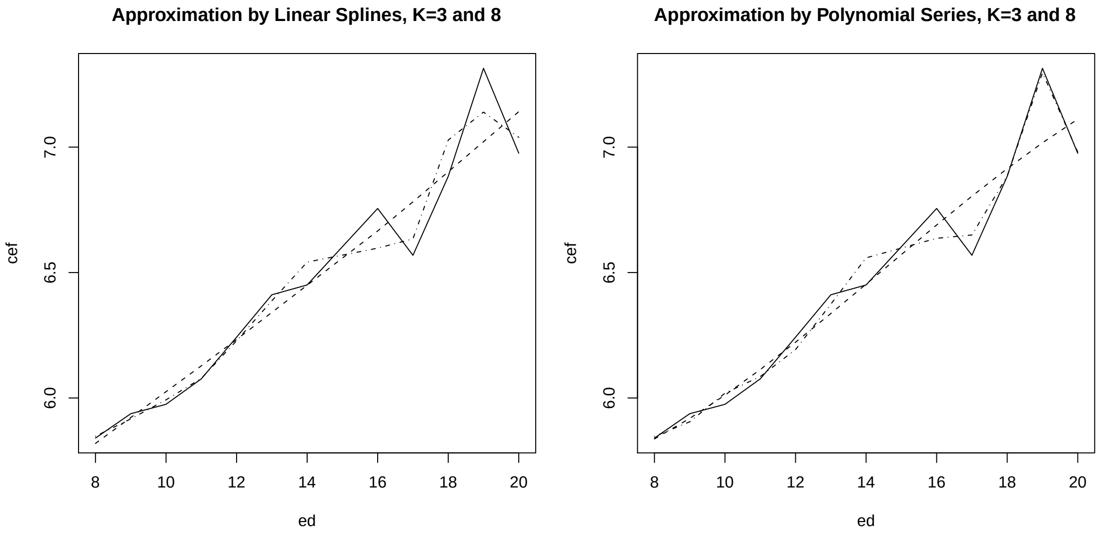

Figure 1. Conditional expectation function (cef) of log wage given education (ed) in the 1990 Census data for the U.S. men of prime age and its least squares approximation by spline (left panel) and polynomial series (right panel). Solid line - conditional expectation function; dashed line - approximation by $k = 3$ series terms; dash-dot line - approximation by $k = 8$ series terms

4.1. L^2 **Limit Theory.** After we have established the set-up, we proceed to derive our results. Recall that $\bar{\sigma}^2 = \sup_{x\in\mathcal{X}} E[\epsilon_i^2|x_i = x]$. In the theorem below, we assume that $\overline{\sigma}^2 \lesssim 1$. This is a mild regularity condition. We start with L^2 rate of convergence result.

Theorem 4.1 (L^2 rate of convergence)**.** *Assume that Conditions A.1-A.3 are satisfied. In addition, assume that $\xi_k^2 \log n/n \to 0$ and $\overline{\sigma}^2 \lesssim 1$. Then under $c_k \to 0$,*

$$\|\widehat{g} - g\|_{F,2} \lesssim_P \sqrt{k/n} + c_k, \tag{4.7}$$

and under $c_k \not\to 0$,

$$\|\widehat{g} - p'\beta\|_{F,2} \lesssim_P \sqrt{k/n} + (\ell_k c_k \sqrt{k/n}) \wedge (\xi_k c_k/\sqrt{n}), \tag{4.8}$$

Comment 4.1. (i) This is our first main result in this paper. The condition $\xi_k^2 \log n/n \to 0$, which we impose, is weaker than that imposed in Newey (1997) who required $k\xi_k^2/n \to 0$. For series satisfying $\xi_k \lesssim \sqrt{k}$, the condition $\xi_k^2 \log n/n \to 0$ amounts to

$$k \log n/n \to 0. \tag{4.9}$$

This condition is the same as that imposed in Stone (1994), Huang (2003a), and Cattaneo and Farell (2013) but the result (4.7) is obtained under the condition (4.9) in Stone (1994) and Huang (2003a) only for spline series and in Cattaneo and Farell (2013) only for local polynomial partition series. Therefore, our result improves on those in the literature by weakening the rate requirements on the growth of k (with respect to n) and/or by allowing for a wider set of series.

(ii) Under the correct specification ($c_k \to 0$), the fastest L^2 rate of convergence is achieved by setting the approximation error and the sampling error to be of the same order,

$$\sqrt{k/n} \asymp c_k.$$

One consequence of this result is that for Hölder classes of smoothness order s, $\Sigma_s(\mathcal{X})$, with $c_k \lesssim k^{-s/d}$, we obtain the optimal L^2 rate of convergence by setting $k \asymp n^{d/(d+2s)}$, which is allowed under our conditions for all $s > 0$ if $\xi_k \lesssim \sqrt{k}$ (Fourier, spline, wavelet, and local polynomial partition series). On the other hand, if ξ_k is growing faster than $\sqrt{k}$, then it is not possible to achieve optimal L^2 rate of convergence for all $s > 0$. For example, for polynomial series considered above, $\xi_k \lesssim k$, and so the condition $\xi_k^2 \log n/n \to 0$ becomes $k^2 \log n/n \to 0$. Hence, optimal L^2 rate of convergence is achieved by polynomial series only if $d/(d+2s) < 1/2$ or, equivalently, $s > d/2$. Even though this condition is somewhat restrictive, it is better than that obtained in Newey (1997) who required $k^3/n \to 0$ for polynomial series, so that optimal L^2 rate could be achieved only if $d/(d+2s) \leq 1/3$ or, equivalently, $s \geq d$. Therefore, our results allow to achieve optimal L^2 rate of convergence in a larger set of classes of functions for particular series.

(iii) The result (4.8) is concerned with the case when the model is misspecified ($c_k \not\to 0$). It shows that when $k/n \to 0$ and $(\ell_k c_k \sqrt{k/n}) \wedge (\xi_k c_k/\sqrt{n}) \to 0$, the estimator $\widehat{g}(\cdot)$ converges in L^2 to the surrogate function $p(\cdot)'\beta$ that provides the best linear approximation to the target function $g(\cdot)$. In this case, the estimator $\widehat{g}(\cdot)$ does not generally converge in L^2 to the target function $g(\cdot)$. □

4.2. **Pointwise Limit Theory.** Next we focus on pointwise limit theory (some authors refer to pointwise limit theory as local asymptotics; see Huang (2003b)). That is, we study asymptotic behavior of $\sqrt{n}\alpha'(\widehat{\beta} - \beta)$ and $\sqrt{n}(\widehat{g}(x) - g(x))$ for particular $\alpha \in S^{k-1}$ and $x \in \mathcal{X}$. Here S^{k-1} denotes the space of vectors α in $\mathbb{R}^k$ with unit Euclidean norm: $\|\alpha\| = 1$. Note that both α and x implicitly depend on n. As we will show, pointwise results can be achieved under weak conditions similar to those we required in Theorem 4.1. The following lemma plays a key role in our asymptotic pointwise normality result.

Lemma 4.1 (Pointwise Linearization)**.** *Assume that Conditions A.1-A.3 are satisfied. In addition, assume that $\xi_k^2 \log n/n \to 0$ and $\overline{\sigma}^2 \lesssim 1$. Then for any $\alpha \in S^{k-1}$,*

$$\sqrt{n}\alpha'(\widehat{\beta} - \beta) = \alpha' \mathbb{G}_n[p_i(\epsilon_i + r_i)] + R_{1n}(\alpha), \tag{4.10}$$

where the term $R_{1n}(\alpha)$, summarizing the impact of unknown design, obeys

$$R_{1n}(\alpha) \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}}(1 + \sqrt{k}\ell_k c_k). \tag{4.11}$$

Moreover,

$$\sqrt{n}\alpha'(\widehat{\beta} - \beta) = \alpha' \mathbb{G}_n[p_i\epsilon_i] + R_{1n}(\alpha) + R_{2n}(\alpha), \tag{4.12}$$

where the term $R_{2n}(\alpha)$, summarizing the impact of approximation error on the sampling error of the estimator, obeys

$$R_{2n}(\alpha) \lesssim_P \ell_k c_k. \tag{4.13}$$

Comment 4.2. (i) In summary, the only condition that generally matters for linearization (4.10)-(4.11) is that $R_{1n}(\alpha) \to 0$, which holds if $\xi_k^2 \log n/n \to 0$ and $k\xi_k^2 \ell_k^2 c_k^2 \log n/n \to 0$. In particular, linearization (4.10)-(4.11) allows for misspecification ($c_k \to 0$ is not required). In principle, linearization (4.12)-(4.13) also allows for misspecification but the bounds are only useful if the model is correctly specified, so that $\ell_k c_k \to 0$. As in the theorem on L^2 rate of convergence, our main condition is that $\xi_k^2 \log n/n \to 0$.

(ii) We conjecture that the bound on $R_{1n}(\alpha)$ can be improved for splines to

$$R_{1n}(\alpha) \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}}(1 + \sqrt{\log n} \cdot \ell_k c_k). \tag{4.14}$$

since it is attained by local polynomials and splines are also similarly localized. □

With the help of Lemma 4.1, we derive our asymptotic pointwise normality result. We will use the following additional notation:

$$\widetilde{\Omega} := Q^{-1}E[(\epsilon_i + r_i)^2 p_i p_i']Q^{-1} \text{ and } \Omega_0 := Q^{-1}E[\epsilon_i^2 p_i p_i']Q^{-1}.$$

In the theorem below, we will impose the condition that $\sup_{x\in\mathcal{X}} E\left[\epsilon_i^2 1\{|\epsilon_i| > M\}|x_i = x\right] \to 0$ as $M \to \infty$ uniformly over n. This is a mild uniform integrability condition. Specifically, it holds if for some $m > 2$, $\sup_{x\in\mathcal{X}} E[|\epsilon_i|^m|x_i = x] \lesssim 1$. In addition, we will impose the condition that $1 \lesssim \underline{\sigma}^2$. This condition is used to properly normalize the estimator.

Theorem 4.2 (Pointwise Normality)**.** *Assume that Conditions A.1-A.3 are satisfied. In addition, assume that (i)* $\sup_{x\in\mathcal{X}} E\left[\epsilon_i^2 1\{|\epsilon_i| > M\}|x_i = x\right] \to 0$ *as* $M \to \infty$ *uniformly over* n*, (ii)* $1 \lesssim \underline{\sigma}^2$*, and (iii)* $(\xi_k^2 \log n/n)^{1/2}(1 + k^{1/2}\ell_k c_k) \to 0$*. Then for any* $\alpha \in S^{k-1}$*,*

$$\sqrt{n}\frac{\alpha'(\widehat{\beta}-\beta)}{\|\alpha'\Omega^{1/2}\|} =_d N(0,1) + o_P(1), \tag{4.15}$$

where we set $\Omega = \widetilde{\Omega}$ *but if* $R_{2n}(\alpha) \to_P 0$*, then we can set* $\Omega = \Omega_0$*. Moreover, for any* $x \in \mathcal{X}$ *and* $s(x) := \Omega^{1/2}p(x)$*,*

$$\sqrt{n}\frac{p(x)'(\widehat{\beta}-\beta)}{\|s(x)\|} =_d N(0,1) + o_P(1), \tag{4.16}$$

and if the approximation error is negligible relative to the estimation error, namely $\sqrt{n}r(x) = o(\|s(x)\|)$*, then*

$$\sqrt{n}\frac{\widehat{g}(x)-g(x)}{\|s(x)\|} =_d N(0,1) + o_P(1). \tag{4.17}$$

Comment 4.3. (i) This is our second main result in this paper. The result delivers pointwise convergence in distribution for any sequences $\alpha = \alpha_n$ and $x = x_n$ with $\alpha \in S^{k-1}$ and $x \in \mathcal{X}$. In fact, the proof of the theorem implies that the convergence is uniform over all sequences. Note that the normalization factor $\|s(x)\|$ is the pointwise standard error, and it is of a typical order $\|s(x)\| \propto \sqrt{k}$ at most points. (If $\ell_k c_k \lesssim 1$ and $\xi_k \lesssim \sqrt{k}$, this holds uniformly across all points.) In this case the condition for negligibility of approximation error $\sqrt{n}r(x)/\|s(x)\| \to 0$, which can be understood as an undersmoothing condition, can be replaced by

$$\sqrt{n/k} \cdot \ell_k c_k \to 0.$$

When $\ell_k c_k \lesssim k^{-s}$, this condition substantially improves on Newey (1997) who required $\sqrt{n}k^{-s} \to 0$ in a similar set-up. Further, under the Newey's condition $\sqrt{n}k^{-s} \to 0$, our asymptotic pointwise normality (4.17) holds assuming that $\xi_k^2 \log n/n \to 0$ (if $k \leq n$) whereas Newey (1997) assumed that $k\xi_k^2/n \to 0$.

(ii) When applied to splines, our result is somewhat less sharp than that of Huang (2003b). Specifically, Huang required that $\xi_k^2 \log n/n \to 0$ and $(n/k)^{1/2} \cdot \ell_k c_k \to 0$ whereas we require $(k\xi_k^2 \log n/n)^{1/2}\ell_k c_k \to 0$ in addition to Huang's conditions. The difference can likely be explained by the fact that we use linearization bound (4.11) whereas for splines it is likely that (4.14) holds as well.

(iii) More generally, our asymptotic pointwise normality result, as well as other related results in this paper, applies to any problem where the estimator of $g(x) = p(x)'\beta + r(x)$ takes the form $p(x)'\widehat{\beta}$, where $\widehat{\beta}$ admits linearization of the form (4.10)-(4.13). □

4.3. **Uniform Limit Theory.** Finally, we turn to a uniform limit theory. Not surprising, stronger conditions are required for our results to hold when compared to the pointwise case. Let $m > 2$. We will need the following assumption on the tails of the regression errors.

Assumption A.4 (Disturbances) *Regression errors satisfy* $\sup_{x\in\mathcal{X}} E[|\epsilon_i|^m|x_i = x] \lesssim 1$.

It will be convenient to denote $\alpha(x) := p(x)/\|p(x)\|$ in this subsection. Moreover, denote

$$\xi_k^L := \sup_{x,x'\in\mathcal{X}:\, x\neq x'} \frac{\|\alpha(x)-\alpha(x')\|}{\|x-x'\|}$$

We will also need the following assumption on the basis functions to hold with the same $m > 2$ as that in Condition A.4.

Assumption A.5 (Basis) *Basis functions are such that (i)* $\xi_k^{2m/(m-2)} \log n/n \lesssim 1$ *and* $\log \xi_k^L \lesssim \log k$.

The following lemma provides uniform linearization of the series estimator and plays a key role in our derivation of the uniform rate of convergence.

Lemma 4.2 (Uniform Linearization). *Assume that Conditions A.1-A.5 are satisfied. Then*

$$\sqrt{n}\alpha(x)'(\widehat{\beta}-\beta) = \alpha(x)'\mathbb{G}_n[p_i(\epsilon_i + r_i)] + R_{1n}(\alpha(x)), \tag{4.18}$$

where $R_{1n}(\alpha(x))$*, summarizing the impact of unknown design, obeys*

$$R_{1n}(\alpha(x)) \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}}(n^{1/m}\sqrt{\log n} + \sqrt{k}\cdot \ell_k c_k) =: \bar{R}_{1n} \tag{4.19}$$

uniformly over $x \in \mathcal{X}$*. Moreover,*

$$\sqrt{n}\alpha(x)'(\widehat{\beta}-\beta) = \alpha(x)'\mathbb{G}_n[p_i\epsilon_i] + R_{1n}(\alpha(x)) + R_{2n}(\alpha(x)), \tag{4.20}$$

where $R_{2n}(\alpha(x))$*, summarizing the impact of approximation error on the sampling error of the estimator, obeys*

$$R_{2n}(\alpha(x)) \lesssim_P \sqrt{\log n}\cdot \ell_k c_k =: \bar{R}_{2n} \tag{4.21}$$

uniformly over $x \in \mathcal{X}$*.*

Comment 4.4. As in the case of pointwise linearization, our results on uniform linearization (4.18)-(4.19) allow for misspecification ($c_k \to 0$ is not required). In principle, linearization (4.20)-(4.21) also allows for misspecification but the bounds are most useful if the model is correctly specified so that $(\log n)^{1/2}\ell_k c_k \to 0$. We are not aware of any similar uniform linearization result in the literature. We believe that this result is useful in a variety of problems. Below we use this result to derive good uniform rate of convergence of the series estimator. Another application of this result would in be testing shape restrictions in the nonparametric model. □

The following theorem provides uniform rate of convergence of the series estimator:

Theorem 4.3 (Uniform Rate of Convergence)**.** *Assume that Conditions A.1-A.5 are satisfied. Then*

$$\sup_{x\in\mathcal{X}} |\alpha(x)'\mathbb{G}_n[p_i\epsilon_i]| \lesssim_P \sqrt{\log n}. \tag{4.22}$$

Moreover, for $\bar R_{1n}$ and $\bar R_{2n}$ given above we have

$$\sup_{x\in\mathcal{X}} |p(x)'(\widehat\beta-\beta)| \lesssim_P \frac{\xi_k}{\sqrt n}(\sqrt{\log n}+\bar R_{1n}+\bar R_{2n}) \tag{4.23}$$

and

$$\sup_{x\in\mathcal{X}} |\widehat g(x)-g(x)| \lesssim_P \frac{\xi_k}{\sqrt n}(\sqrt{\log n}+\bar R_{1n}+\bar R_{2n}) + \ell_k c_k. \tag{4.24}$$

Comment 4.5. (i) This is our third main result in this paper. Assume that $\ell_k c_k \lesssim k^{-s/d}$, $\xi_k \lesssim \sqrt{k}$, and $\bar R_{1n}+\bar R_{2n} \lesssim (\log n)^{1/2}$. Then the bound in (4.24) becomes

$$\sup_{x\in\mathcal{X}} |\widehat g(x)-g(x)| \lesssim_P \sqrt{\frac{k\log n}{n}} + k^{-s/d}.$$

Therefore, setting $k \asymp (\log n/n)^{-d/(2s+d)}$, we obtain

$$\sup_{x\in\mathcal{X}} |\widehat g(x)-g(x)| \lesssim_P \left(\frac{\log n}{n}\right)^{s/(2s+d)},$$

which is the optimal uniform rate of convergence in the function class $\Sigma_s(\mathcal{X})$; see Stone (1982). To the best of our knowledge, our paper is the first to show that the series estimator attains the optimal uniform rate of convergence under rather general conditions. We also note here that it has been known for a long time that a local polynomial (kernel) estimator achieves the same optimal uniform rate of convergence for a long time; see, for example, Tsybakov (2009), and it was also shown recently by Cattaneo and Farell (2013) that local polynomial partition series estimator also achieves the same rate.

(ii) One of the critical conditions to attain the optimal uniform rate of convergence is that we require $\bar R_{1n} \lesssim (\log n)^{1/2}$. Specifically, under our other assumptions, this condition requires that $k\log n/n^{1-2/m} \lesssim 1$ and $k^{2-2s/d}/n \lesssim 1$, and so we can set $k \asymp (\log n/n)^{-d/(2s+d)}$ if $d/(2s+d) < 1-2/m$ and $(2d-2s)/(2s+d) < 1$ or, equivalently, $m > 2+d/s$ and $d < 4s$.

(iii) If the errors have heavy tails (m is small) but one is only interested in estimating some location function, then one could use median regression estimator that will achieve faster uniform convergence rates, since the "errors" in the linearized version of this estimator are just Bernoulli and therefore are sub-Gaussian; see Belloni et al. (2011) for details. □

After establishing the uniform rate of convergence, we present two results on inference based on the series estimator. The first result on inference is concerned with the strong approximation of a series process by a Gaussian process and is a (relatively) minor extension of the result obtained by Chernozhukov et al. (2009). The extension is undertaken to allow for a non-vanishing specification error to cover misspecified models. In particular, we make a distinction between $\widetilde{\Omega} = Q^{-1}E[(\epsilon_i + r_i)^2 p_i p_i']Q^{-1}$, and $\Omega_0 = Q^{-1}E[\epsilon_i^2 p_i p_i']Q^{-1}$ which are potentially asymptotically different if $\bar{R}_{2n} \not\to_P 0$. To state the result, let a_n be some sequence of positive numbers satisfying $a_n \to \infty$.

Theorem 4.4 (Strong Approximation by a Gaussian Process)**.** *Assume that Conditions A.1-A.5 are satisfied with $m \geq 3$. In addition, assume that (i) $\bar{R}_{1n} = o_P(a_n^{-1})$, (ii) $1 \lesssim \underline{\sigma}^2$, and (iii) $a_n^6 k^4 \xi_k^2 (1 + \ell_k^3 c_k^3)^2 \log^2 n/n \to 0$. Then for some $\mathcal{N}_k \sim N(0, I_k)$,*

$$\sqrt{n}\frac{\alpha(x)'(\widehat{\beta} - \beta)}{\|\alpha(x)'\Omega^{1/2}\|} =_d \frac{\alpha(x)'\Omega^{1/2}}{\|\alpha(x)'\Omega^{1/2}\|}\mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}), \tag{4.25}$$

so that for $s(x) = \Omega^{1/2}p(x)$,

$$\sqrt{n}\frac{p(x)'(\widehat{\beta} - \beta)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|}\mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}), \tag{4.26}$$

and if $\sup_{x\in\mathcal{X}} \sqrt{n}|r(x)|/\|s(x)\| = o(a_n^{-1})$, then

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|}\mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}). \tag{4.27}$$

where we set $\Omega = \widetilde{\Omega}$ but if $\bar{R}_{2n} = o_P(a_n^{-1})$, then we can set $\Omega = \Omega_0$.

Comment 4.6. Ideally, in order to perform uniform in $x \in \mathcal{X}$ inference on $g(x)$, one would like to have a result of the form

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} =_d G(x) + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}) \tag{4.28}$$

or, equivalently,

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} \to_d G(x) \text{ in } \ell^\infty(\mathcal{X}) \tag{4.29}$$

where $\{G(x) : x \in \mathcal{X}\}$ is some zero-mean Gaussian process. However, one can show that the process on the left-hand side of (4.28) and (4.29) is not asymptotically equicontinuous, and so it does not have a limit distribution. Instead, Theorem 4.4 provides an approximation of the series process by *a sequence* of zero-mean Gaussian processes. Since $a_n \to \infty$, under our conditions the theorem implies that the series process is well approximated by a Gaussian process, and so the theorem can be interpreted that in large samples, the distribution of the series process depends on the distribution of the data only via covariance matrix Ω; hence, it allows to do inference based on the whole series process. Note that the conditions of the theorem are quite strong but the result of the theorem is also much stronger than the pointwise normality result: it asserts that the entire series process is uniformly close to a Gaussian process of the stated form. □

Our result on strong approximation by a Gaussian process plays an important role in our second result on inference that is concerned with the weighted bootstrap. Consider a set of weights $h_1, \ldots, h_n$ that are i.i.d. draws from the standard exponential distribution and are independent of the data. For each draw of such weights, define the weighted bootstrap draw of the least squares estimator as a solution to the least squares problem weighted by $h_1, \ldots, h_n$, namely

$$\widehat{\beta}^b \in \arg\min_{b\in\mathbb{R}^k} \mathbb{E}_n[h_i(y_i - p_i'b)^2]. \tag{4.30}$$

For all $x \in \mathcal{X}$, denote $\widehat{g}^b(x) = p(x)'\widehat{\beta}^b$. The following theorem establishes a new result that states that the weighted bootstrap distribution is valid for approximating the distribution of the series process.

Theorem 4.5 (Weighted Bootstrap Method)**.** *(1) Assume that Conditions A.1-A.5 are satisfied. Then the weighted bootstrap process satisfies*

$$\sqrt{n}\alpha(x)'(\widehat{\beta}^b - \widehat{\beta}) = \alpha(x)'\mathbb{G}_n[(h_i - 1)p_i(\epsilon_i + r_i)] + R^b_{1n}(\alpha(x)),$$

where $R^b_{1n}(\alpha(x))$ obeys

$$R^b_{1n}(\alpha(x)) \lesssim_P \sqrt{\frac{\xi_k^2 \log^3 n}{n}}(n^{1/m}\sqrt{\log n} + \sqrt{k}\cdot \ell_k c_k) = o(1/\log n) =: \bar{R}^b_{1n} \tag{4.31}$$

uniformly over $x \in \mathcal{X}$.

(2) If, in addition, Conditions A.4 and A.5 are satisfied with $m \geq 3$ and (i) $\bar{R}^b_{1n} = o_P(a_n^{-1})$, (ii) $1 \lesssim \underline{\sigma}^2$, and (iii) $a_n^6 k^4 \xi_k^2 (1+\ell_k^3 c_k^3)^2 \log^2 n/n \to 0$ hold, then for $s(x) = \Omega^{1/2}p(x)$ and some $\mathcal{N}_k \sim N(0, I_k)$,

$$\sqrt{n}\frac{p(x)'(\widehat{\beta}^b - \widehat{\beta})}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|}\mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}), \tag{4.32}$$

and so

$$\sqrt{n}\frac{\widehat{g}^b(x) - \widehat{g}(x)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|}\mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}). \tag{4.33}$$

where we set $\Omega = \widetilde{\Omega}$, but if $\bar{R}_{2n} = o_P(a_n^{-1})$, then we can set $\Omega = \Omega_0$. Moreover, the bounds (4.31), (4.32), and (4.33) continue to hold in P-probability if we replace the unconditional probability P by the conditional probability $P^*(\cdot|D)$ where $D = \{(x_i, y_i) : i = 1, \dots, n\}$.

Comment 4.7. (i) This is our fourth main result in this paper. The theorem implies that the weighted bootstrap process can be approximated by the same Gaussian process as that used to approximate original series process, and so weighted bootstrap process also approximates the original series process. Alternatively, one can think of this theorem as a result that shows equivalence of the weighted bootstrap process and the Gaussian process in approximating the original series process.

(ii) We emphasize that the first part of the theorem does not require the correct specification, that is the case $c_k \not\to 0$ is allowed. The second part implicitly require $c_k \to 0$ via the condition $a_n^6 k^4 \xi_k^2 (1 + \ell_k^3 c_k^3)^2 \log^2 n/n \to 0$. Also, in this theorem, symbol P refers to a joint probability measure with respect to the data $D = \{(x_i, y_i) : i = 1, \dots, n\}$ and the set of bootstrap weights $\{h_i : i = 1, \dots, n\}$. □

We close this section by establishing sufficient conditions for consistent estimation of Ω. Recall that $Q = E[p_i p_i']$. In addition, denote $\Sigma = E[(\epsilon_i + r_i)^2 p_i p_i'] = I$, $\widehat{Q} = \mathbb{E}_n[p_i p_i']$, and $\widehat{\Sigma} = \mathbb{E}_n[\widehat{\epsilon}_i^2 p_i p_i']$ where $\widehat{\epsilon}_i = y_i - p_i'\widehat{\beta}$, and let $v_n = (E[\max_{1 \leq i \leq n} |\epsilon_i|^2])^{1/2}$.

Theorem 4.6 (Matrices Estimation). *Assume that Conditions A.1-A.5 are satisfied. In addition, assume that $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log n)^{1/2}$. Then*

$$\|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} = o(1) \quad \textit{and} \quad \|\widehat{\Sigma} - \Sigma\| \lesssim_P (v_n \vee 1 + \ell_k c_k) \sqrt{\frac{\xi_k^2 \log n}{n}} = o(1).$$

Moreover, for $\widehat{\Omega} = \widehat{Q}^{-1} \widehat{\Sigma} \widehat{Q}^{-1}$,

$$\|\widehat{\Omega} - \Omega\| \lesssim_P (v_n \vee 1 + \ell_k c_k) \sqrt{\frac{\xi_k^2 \log n}{n}} = o(1).$$

Comment 4.8. Theorem 4.6 allows for consistent estimation of the matrix Q under the mild condition $\xi_k^2 \log n/n \to 0$ and for consistent estimation of the matrices Σ and Ω under slightly more restricted conditions. Not surprising, the estimation of Σ and Ω depends on the tail behavior of the error term via the value of v_n. Under Condition A.4, the following simple inequality can be used to bound v_n: $v_n \lesssim n^{1/m}$. □

5. Rates and Inference on Linear Functionals

In this section, we derive rates and inference results for linear functionals $\theta(w), w \in \mathcal{I}$ of the conditional expectation function such as its derivative or average derivative. To a large extent, with the exception of Theorem 5.6, the results presented in this section can be considered as an extension of results presented in Section 4, and so similar comments can be applied as those given in Section 4. Theorem 5.6 deals with construction of uniform confidence bands for linear functionals under weak conditions and is a new result.

By the linearity of the series approximations, the linear functionals can be seen as linear functions of the least squares coefficients β up to an approximation error, that is

$$\theta(w) = \ell_\theta(w)'\beta + r_\theta(w), \quad w \in \mathcal{I},$$

where $\ell_\theta(w)'\beta$ is the series approximation, with $\ell_\theta(w)$ denoting the k-vector of loadings on the coefficients, and $r_\theta(w)$ is the remainder term, which corresponds to the approximation error. Throughout this section, we assume that $\mathcal{I}$ is a subset of some Euclidean space $\mathbb{R}^l$ equipped with its usual norm $\|\cdot\|$. We allow $\mathcal{I} = \mathcal{I}_n$ to depend on n but for simplicity, we assume that the diameter of $\mathcal{I}$ is bounded from above uniformly over n. Results allowing for the case where $\mathcal{I}$ is expanding as n grows can be covered as well with slightly more technicalities.

In order to perform inference, we construct estimators of $\sigma_\theta^2(w) = \ell_\theta(w)'\Omega\ell_\theta(w)/n$, the variance of the associated linear functionals, as

$$\widehat{\sigma}_\theta^2(w) = \ell_\theta(w)'\widehat{\Omega}\ell_\theta(w)/n. \tag{5.34}$$

In what follows, it will be convenient to have the following result on consistency of $\widehat{\sigma}_\theta(w)$:

Lemma 5.1 (Variance Estimation for Linear Functionals)**.** *Assume that Conditions A.1-A.5 are satisfied. In addition, assume that (i) $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log n)^{1/2}$ and (ii) $1 \lesssim \underline{\sigma}^2$. Then*

$$\left|\frac{\widehat{\sigma}_\theta(w)}{\sigma_\theta(w)} - 1\right| \lesssim_P \|\widehat{\Omega} - \Omega\| \lesssim_P (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log n}{n}} = o(1)$$

uniformly over $w \in \mathcal{I}$.

By Lemma 5.1, under our conditions, (5.34) is uniformly consistent for $\sigma_\theta^2(w)$ in the sense that $\widehat{\sigma}_\theta^2(w)/\sigma_\theta^2(w) = 1 + o_P(1)$ uniformly over $w \in \mathcal{I}$.

5.1. **Pointwise Limit Theory for Linear Functionals.** We now present a result on pointwise rate of convergence for linear functionals. The rate we derive is $\|\ell_\theta(w)\|/\sqrt{n}$. Some examples with explicit bounds on $\|\ell_\theta(w)\|$ are given below.

Theorem 5.1 (Pointwise Rate of Convergence for Linear Functionals)**.** *Assume that Conditions A.1-A.3 are satisfied. In addition, assume that (i)* $\sqrt{n}|r_\theta(w)|/\|\ell_\theta(w)\| \to 0$*, (ii)* $\bar{\sigma}^2 \lesssim 1$*, (iii)* $(\xi_k^2 \log n/n)^{1/2}(1 + k^{1/2}\ell_k c_k) \to 0$*, and (iv)* $\ell_k c_k \to 0$*. Then*

$$|\widehat{\theta}(w) - \theta(w)| \lesssim_P \frac{\|\ell_\theta(w)\|}{\sqrt{n}}.$$

Comment 5.1. (i) This theorem shows in particular that $\widehat{\theta}(w)$ is $\sqrt{n}$-consistent whenever $\|\ell_\theta(w)\| \lesssim 1$. A simple example of this case is $\theta = \theta(w) = E[g(x_1)]$. In this example, $\ell = \ell(w) = E[p(x_1)]$, and so $\|\ell\| = \|E[p(x_1)]\| \lesssim 1$ where the last inequality follows from the argument used in the proof of Proposition 3.1. Another simple example is $\theta = \theta(w) = E[p(x_1)g(x_1)] = \beta_1$. In this example, $\ell = \ell(w)$ is a k-vector whose first component is 1 and all other components are 0, and so $\|\ell\| \lesssim 1$. This example trivially implies $\sqrt{n}$-consistency of the series estimator of the linear part of the partially linear model. Yet another example, which is discussed in Newey (1997), is the average partial derivative.

(ii) Condition $\sqrt{n}|r_\theta(w)/\|\ell_\theta(w)\| \to 0$ imposed in this theorem can be understood as undersmoothing condition. Unfortunately, to the best of our knowledge, there is no theoretically justified practical procedure in the literature that would lead to a desired level undersmoothing. Some ad hoc suggestions include using cross validation or "plug-in" method to determine the number of series terms that would minimize the asymptotic integrated mean-square error of the series estimator (see Hardle, 1990) and then blow up the estimated number of series terms by some number that grows to infinity as the sample size increases. □

To perform pointwise inference, we consider the t-statistic:

$$t(w) = \frac{\widehat{\theta}(w) - \theta(w)}{\widehat{\sigma}_\theta(w)}.$$

We can carry out standard inference based on this statistic because of the following theorem.

Theorem 5.2 (Pointwise Inference for Linear Functionals)**.** *Assume that the conditions of Theorem 4.2 and Lemma 5.1 are satisfied. In addition, assume that* $\sqrt{n}|r_\theta(w)|/\|\ell_\theta(w)\| \to 0$*. Then*

$$t(w) \to_d N(0, 1).$$

The same comments apply here as those given in Section 4.2 for pointwise results on estimating the function g itself.

5.2. **Uniform Limit Theory for Linear Functionals.** In obtaining uniform rates of convergence and inference results for linear functionals, we will denote

$$\xi_{k,\theta} := \sup_{w\in\mathcal{I}} \|l_\theta(w)\| \ \text{ and } \ \xi^L_{k,\theta} := \sup_{w,w'\in\mathcal{I}:\, w\neq w'} \frac{\|\ell_\theta(w)-\ell_\theta(w')\|}{\|w-w'\|}.$$

The value of $\xi_{k,\theta}$ depends on the choice of the basis for the series estimator and on the linear functional. Newey (1997) and Chen (2007) provides several examples. In the case of splines with $\mathcal{X}=[0,1]^d$, it has been established that $\xi_k \lesssim \sqrt{k}$ and $\sup_{x\in\mathcal{X}}\|\partial^m_{x_j}p(x)\| \lesssim k^{1/2+m}$; see, for example, Newey (1997). With this basis we have for

1. the function g itself: $\theta(x)=g(x)$, $\ell_\theta(x)=p(x)$, and $\xi_{k,\theta}\lesssim\sqrt{k}$;
2. the derivatives: $\theta(x)=\partial_{x_j}g(x)$, $\ell_\theta(x)=\partial_{x_j}p(x)$, $\xi_{k,\theta}\lesssim k^{3/2}$;
3. the average derivatives: $\theta=\int\partial_{x_j}g(x)d\mu(x)$, $\ell_\theta=\int\partial_{x_j}p(x)d\mu(x)$, and $\xi_{k,\theta}\lesssim 1$,

where in the last example it is assumed that $\text{supp}(\mu)\subset\text{int}\mathcal{X}$, x_1 is continuously distributed with the density bounded below from zero on $\text{supp}(\mu)$, and $|\partial_{x_l}\mu(x)|\lesssim 1$ for all $l=1,\dots,k$.

We will impose the following regularity condition on the loadings on the coefficients $\ell_\theta(w)$:

Assumption A.6 (Loadings) *Loadings on the coefficients satisfy (i)* $\sup_{w\in\mathcal{I}} 1/\|\ell_\theta(w)\|\lesssim 1$ *and (ii)* $\log\xi^L_{k,\theta}\lesssim\log k$.

The first part of this condition implies that the linear functional is normalized appropriately. The second part is a very mild restriction on the modulus of continuity (with respect to w) of the linear functional $\theta(w)$.

Under Conditions A.1-A.6, results presented in Lemma 4.2 on uniform linearization can be extended to cover general linear functionals considered here:

Lemma 5.2 (Uniform Linearization for Linear Functionals)**.** *Assume that Conditions A.1-A.6 are satisfied. Then for* $\alpha_\theta(w)=\ell_\theta(w)/\|\ell_\theta(w)\|$,

$$\sqrt{n}\alpha_\theta(w)'(\widehat{\beta}-\beta)=\alpha_\theta(w)'\mathbb{G}_n[p_i(\epsilon_i+r_i)]+R_{1n}(\alpha_\theta(w)),$$

where $R_{1n}(\alpha_\theta(w))$*, summarizing the impact of unknown design, obeys*

$$R_{1n}(\alpha(w))\lesssim_P\sqrt{\frac{\xi_k^2\log n}{n}}(n^{1/m}\sqrt{\log n}+\sqrt{k}\cdot\ell_kc_k)=\bar{R}_{1n}$$

uniformly over $w \in \mathcal{I}$. Moreover,

$$\sqrt{n}\alpha_\theta(w)'(\widehat{\beta} - \beta) = \alpha_\theta(w)'\mathbb{G}_n[p_i\epsilon_i] + R_{1n}(\alpha_\theta(w)) + R_{2n}(\alpha_\theta(w)),$$

where $R_{2n}(\alpha_\theta(w))$, summarizing the impact of approximation error on the sampling error of the estimator, obeys

$$R_{2n}(\alpha_\theta(w)) \lesssim_P \sqrt{\log n} \cdot \ell_k c_k = \bar{R}_{2n}$$

uniformly over $w \in \mathcal{I}$.

From Lemma 5.2, we can derive the following theorem on uniform rate of convergence for linear functionals.

Theorem 5.3 (Uniform Rate of Convergence for Linear Functionals)**.** *Assume that Conditions A.1-A.6 are satisfied. Then*

$$\sup_{w\in\mathcal{I}} \left|\alpha_\theta(w)'\mathbb{G}_n[p_i\epsilon_i]\right| \lesssim_P \sqrt{\log n}. \tag{5.35}$$

If, in addition, we assume that (i) $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log n)^{1/2}$ and (ii) $\sup_{w\in\mathcal{I}} |r_\theta(w)|/\|l_\theta(w)\| = o((\log n/n)^{1/2})$, then

$$\sup_{w\in\mathcal{I}} |\widehat{\theta}(w) - \theta(w)| \lesssim_P \sqrt{\frac{\xi_{k,\theta}^2 \log n}{n}}. \tag{5.36}$$

Next, we consider the problem of uniform inference for linear functionals based on the series estimator. We base our inference on the t-statistic process:

$$\left\{ t(w) = \frac{\widehat{\theta}(w) - \theta(w)}{\widehat{\sigma}_\theta(w)}, \quad w \in \mathcal{I} \right\}. \tag{5.37}$$

We present two results for inference on linear functionals. The first result is an extension of Theorem 4.4 on strong approximations to cover the case of linear functionals. As we discussed in Comment 4.6, in order to perform uniform in $w \in \mathcal{I}$ inference on $\theta(w)$, we would like to approximate the distribution of the *whole* process (5.37). However, one can show that this process typically does not have a limit distribution in $\ell^\infty(\mathcal{I})$. Yet, we can construct a Gaussian process that would be close to the process (5.37) for all $w \in \mathcal{I}$ simultaneously with large probability. Specifically, we will approximate the t-statistic process by the following Gaussian coupling:

$$\left\{ t_n^*(w) = \frac{\ell(w)'\Omega^{1/2}\mathcal{N}_k/\sqrt{n}}{\sigma_\theta(w)}, \quad w \in \mathcal{I} \right\} \tag{5.38}$$

where $\mathcal{N}_k$ denotes a vector of k i.i.d. $N(0,1)$ random variables.

Theorem 5.4 (Strong Approximation by a Gaussian Process for Linear Functionals)**.** *Assume that the conditions of Theorem 4.4 and Condition A.6 are satisfied. In addition, assume that (i)* $\bar{R}_{2n} \lesssim (\log n)^{1/2}$*, (ii)* $\xi_k \log n / n^{1/2-1/m} = o(a_n^{-1})$*, and that (iii)* $\sup_{w \in \mathcal{I}} \sqrt{n}|r_\theta(w)|/\|\ell_\theta(w)\| = o(a_n^{-1})$*. Then*

$$t(w) =_d t^*(w) + o_P(a_n^{-1}) \textit{ in } \ell^\infty(\mathcal{I}).$$

As in the case of inference on the function $g(x)$, we could also consider weighted bootstrap method to obtain a result analogous to that in Theorem 4.5. For brevity of the paper, however, we do not consider weighted bootstrap method here.

The second result on inference for linear functionals is new and concerns with the problem of constructing uniform confidence bands for the linear functional $\theta(w)$. Specifically, we are interested in the confidence bands of the form

$$[\dot{\imath}(w), \ddot{\imath}(w)] = \left[\widehat{\theta}(w) - c_n(1-\alpha)\widehat{\sigma}_\theta(w), \widehat{\theta}(w) + c_n(1-\alpha)\widehat{\sigma}_\theta(w)\right], \ w \in \mathcal{I} \tag{5.39}$$

where $c_n(1-\alpha)$ is chosen so that $\theta(w) \in [\dot{\imath}(w), \ddot{\imath}(w)]$ for all $w \in \mathcal{I}$ with the prescribed probability $1-\alpha$ where $\alpha \in (0,1)$ is a user-specified level. For this purpose, we would like to set $c_n(1-\alpha)$ as the $(1-\alpha)$-quantile of $\sup_{w \in \mathcal{I}} |t(w)|$. However, this choice is infeasible because the exact distribution of $\sup_{w \in \mathcal{I}} |t(w)|$ is unknown. Instead, Theorem 5.4 suggests that we can set $c_n(1-\alpha)$ as the $(1-\alpha)$-quantile of $\sup_{w \in \mathcal{I}} |t^\star(w)|$ or, if Ω is unknown and has to be estimated, that we can set

$$c_n(1-\alpha) := \text{the conditional } (1-\alpha) - \text{quantile of } \sup_{w \in \mathcal{I}} |\widehat{t}^*(w)| \text{ given the data} \tag{5.40}$$

where

$$\widehat{t}_n^\star(w) := \frac{l(w)'\widehat{\Omega}^{1/2}\mathcal{N}_k/\sqrt{n}}{\widehat{\sigma}_\theta(w)}, \ w \in \mathcal{I}$$

and $\mathcal{N}_k \sim N(0, I_k)$. Note that $c_n(1-\alpha)$ defined in (5.40) can be simulated numerically with any precision. Yet, conditions of Theorem 5.4 are rather strong. Fortunately, Chernozhukov et al. (2012) noticed that when we are only interested in the supremum of the process and do not need the process itself, sufficient conditions for the strong approximation can be much weaker. Specifically, we have the following theorem, which is an application of a general result obtained in Chernozhukov et al. (2012):

Theorem 5.5 (Strong Approximation of Suprema for Linear Functionals)**.** *Assume that Conditions A.1-A.6 are satisfied with* $m \geq 4$*. In addition, assume that (i)* $\bar{R}_{1n} + \bar{R}_{2n} \lesssim$

$1/(\log n)^{1/2}$, (ii) $\xi_k \log^2 n / n^{1/2-1/m} \to 0$, (iii) $1 \lesssim \underline{\sigma}^2$, and (iv) $\sup_{w \in \mathcal{I}} \sqrt{n}|r_\theta(w)|/\|\ell_\theta(w)\| = o(1/(\log n)^{1/2})$. Then

$$\sup_{w \in \mathcal{I}} |t(w)| =_d \sup_{t \in \mathcal{I}} |t^*(w)| + o_P\left(\frac{1}{\sqrt{\log n}}\right).$$

Construction of uniform confidence bands also critically relies on the following anti-concentration lemma:

Lemma 5.3 (Anti-concentration for Separable Gaussian Processes)**.** *Let $Y = (Y_t)_{t \in T}$ be a separable Gaussian process indexed by a semimetric space T such that $E[Y_t] = 0$ and $E[Y_t^2] = 1$ for all $t \in T$. Assume that $\sup_{t \in T} X_t < \infty$ a.s. Then $a(|Y|) := E[\sup_{t \in T} |Y_t|] < \infty$ and*

$$\sup_{x \in \mathbb{R}} P\left\{\left|\sup_{t \in T} |Y_t| - x\right| \leq \varepsilon\right\} \leq A\varepsilon a(|Y|)$$

for all $\varepsilon \geq 0$ and some absolute constant A.

The proof of this lemma can be found in Chernozhukov et al. (2012) (Corollary 2.1). From Theorem 5.5 and Lemma 5.3, we can now derive the following result on uniform validity of confidence bands in (5.39):

Theorem 5.6 (Uniform Inference for Linear Functionals)**.** *Assume that the conditions of Theorem 5.5 are satisfied. In addition, assume that $c_n(1-\alpha)$ is defined by (5.40). Then*

$$P\left\{\sup_{w \in \mathcal{I}} |t_n(w)| \leq c_n(1-\alpha)\right\} = 1 - \alpha + o(1). \tag{5.41}$$

As a consequence, the confidence bands defined in (5.39) satisfy

$$P\Big\{\theta(w) \in [i(w), \ddot{\imath}(w)], \text{ for all } w \in \mathcal{I}\Big\} = 1 - \alpha + o(1). \tag{5.42}$$

The width of the confidence bands $2c_n(1-\alpha)\widehat{\sigma}_n(w)$ obeys

$$2c_n(1-\alpha)\widehat{\sigma}_n(w) \lesssim_P \sigma_n(w)\sqrt{\log n} \lesssim \|\ell_\theta(w)\|\sqrt{\frac{\log n}{n}} \lesssim \sqrt{\frac{\xi_{k,\theta}^2 \log n}{n}} \tag{5.43}$$

uniformly over $w \in \mathcal{I}$.

Comment 5.2. (i) This is our fifth (and last) main result in this paper. The theorem shows that the confidence bands constructed above maintain the required level asymptotically and establishes that the uniform width of the bands is of the same order as the uniform rate of convergence. Moreover, confidence intervals are asymptotically similar.

(ii) The proof strategy of Theorem 5.6 is similar to that proposed in Chernozhukov et al. (2009) for inference on the minimum of a function. Since the limit distribution may not exists, the insight was to use distributions provided by couplings. Because the limit distribution does not necessarily exist, it is not immediately clear that the confidence bands are asymptotically similar or at least maintain the right asymptotic level. Nonetheless, we show that the confidence bands are asymptotically similar with the help of anti-concentration lemma stated above.

(iii) Theorem 5.6 only considers two-sided confidence bands. However, both Theorem 5.5 and Lemma 5.3 continue to hold if we replace suprema of absolute values of the processes by suprema of the processes itself, namely if we replace $\sup_{w\in\mathcal{I}}|t_n(w)|$ and $\sup_{w\in\mathcal{I}}|t_n^*|$ in Theorem 5.5 by $\sup_{w\in\mathcal{I}}t_n(w)$ and $\sup_{w\in\mathcal{I}}t_n^*(w)$, respectively, and $\sup_{t\in T}|Y_t|$ in Lemma 5.3 by $\sup_{t\in T}Y_t$. Therefore, we can show that Theorem 5.6 also applies for one-sided confidence bands, namely Theorem 5.6 holds with $c_n(1-\alpha)$ defined as the conditional $(1-\alpha)$-quantile of $\sup_{w\in\mathcal{I}}\widehat{t}_n^*(w)$ given the data and the confidence bands defined by $[i(w), \ddot{\imath}(w)] := (-\infty, \widehat{\theta}(w) + c_n(1-\alpha)\widehat{\sigma}_n(w)]$ for all $w \in \mathcal{I}$. □

6. Tools: Maximal Inequalities for Matrices and Empirical Processes

In this section we collect the main technical tools that our analysis rely upon, namely Khinchin Inequalities for Matrices and a Data Dependent Maximal Inequalities.

6.1. **Khinchin Inequalities for Matrices.**

Consider the Schatten norm S_P on symmetric $k\times k$ matrices Q as

$$\|Q\|_{S_P} = \left(\sum_{j=1}^{k}|\lambda_j(Q)|^p\right)^{1/p}.$$

The case $p=\infty$ recovers the operator norm $\|\cdot\|$ and $p=2$ the Frobenius norm. It is obvious that for any $p\geq 1$

$$\|Q\| \leq \|Q\|_{S_P} \leq k^{1/p}\|Q\|.$$

Therefore, setting $p=\log k$, we get equivalence

$$\|Q\| \leq \|Q\|_{S_{\log k}} \leq e\|Q\|. \tag{6.44}$$

Lemma 6.1 (Khinchin Inequality for Matrices)**.** *For symmetric $k\times k$-matrices Q_i, $i=1,\dots,n$, and $2\leq p<\infty$, and an i.i.d. sequence of Rademacher variables $\varepsilon_1,\dots,\varepsilon_n$ we have*

$$a_P\left\|\left(\mathbb{E}_n[Q_i^2]\right)^{1/2}\right\|_{S_P} \leq \left(E_\varepsilon\left\|\mathbb{G}_n[\varepsilon_i Q_i]\right\|_{S_P}^p\right)^{1/p} \leq b_P\left\|(\mathbb{E}_n[Q_i^2])^{1/2}\right\|_{S_P}$$

where

$$b_P \leq [2^{1/2}\pi/e]^{1/2}\cdot\sqrt{p}.$$

As a consequence of equivalence (6.44) if $k \geq e^2$ we have

$$E_\varepsilon \|\mathbb{G}_n[\varepsilon_i Q_i]\| \lesssim \sqrt{\log k} \|(\mathbb{E}_n[Q_i^2])^{1/2}\|$$

The notable feature of this inequality is the $\sqrt{\log k}$ factor instead of the $\sqrt{k}$ factor expected from the conventional maximal inequalities based on entropy. This inequality due to Lust-Picard and Pisier (1991) generalizes the Khinchin inequality for vectors. A version of this inequality was derived by Guédon and Rudelson (2007) using generalized entropy (majorizing measure) arguments. This is another striking example where the use of generalized entropy yields drastic improvements over the use of entropy. Prior to this Talagrand (1996a) provided ellipsoidal examples where the difference between the two approaches was even more extreme.

6.2. **LLN for Matrices.** The following lemma is a variant of a fundamental result obtained by Rudelson (1999).

Lemma 6.2 (Matrix LLN)**.** *Let $Q_1, \dots, Q_n$ be i.n.i.d. symmetric non-negative $k \times k$-matrices with $k \geq e^2$ such that $Q = \mathbb{E}_n[E[Q_i]]$ and $\|Q_i\| \leq M$ a.s., then for $\widehat{Q} = \mathbb{E}_n[Q_i]$*

$$\Delta := E\|\widehat{Q} - Q\| \lesssim \sqrt{\frac{M(1 + \|Q\|)\log n}{n}}.$$

In particular, if $Q_i = p_i p_i'$, with $\|p_i\| \leq \xi_k$ a.s., then

$$\Delta := E\|\widehat{Q} - Q\| \lesssim \sqrt{\frac{\xi_k^2(1 + \|Q\|)\log n}{n}}.$$

6.3. **Maximal Inequalities.** Consider a function class $\mathcal{F}$ collecting functions mapping some set $\mathcal{Z}$ to $\mathbb{R}$, equipped with an envelope function $F(z) \geq \sup_{f\in\mathcal{F}} |f(z)|$. The *covering number* $N(\mathcal{F}, L^2(Q), \varepsilon)$ is the minimal number of $L^2(Q)$-balls of radius ε needed to cover the function set $\mathcal{F}$. The *covering number* relative to the envelope function is given by

$$N\left(\mathcal{F}, L^2(Q), \varepsilon \|F\|_{Q,2}\right). \tag{6.45}$$

The *entropy* is the logarithm of the covering number.

We rely on the following result.

Proposition 6.1. *Let $(\epsilon_1, X_1), \dots, (\epsilon_n, X_n)$ be i.i.d. random vectors in $\mathbb{R}^{d+1}$ with $E[\epsilon_i|X_i] = 0$ and $\sigma^2 := \sup_x E[\epsilon_i^2|X_i = x] < \infty$. Let $\mathcal{F}$ be a class of functions on $\mathbb{R}^d$ such that*

$E[f(X_1)^2] = 1$ (normalization) and $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. Let $\mathcal{G} := \{(\epsilon, x) \ni \mathbb{R}^{d+1} \mapsto \epsilon f(x) : f \in \mathcal{F}\}$. Suppose that there exist constants $A > e^2$ and $V \geq 2$ such that

$$\sup_Q N(\mathcal{G}, L^2(Q), \epsilon\|G\|_{L^2(Q)}) \leq (A/\epsilon)^V$$

for all $0 < \epsilon \leq 1$ for the envelope $G(\epsilon, x) := |\epsilon| b$. If for some $m > 2$ $E[|\epsilon_1|^m] < \infty$, then

$$E\left[\left\|\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathcal{F}}\right] \leq C\left[(\sigma + \sqrt{E[|\epsilon_1|^m]})\sqrt{nV\log(Ab)} + Vb^{m/(m-2)}\log(Ab)\right],$$

where C is a universal constant.

The proof is based on a truncation argument and maximal inequalities for uniformly bounded classes of functions developed in Giné and Koltchinskii (2006). We recall its version.

Theorem 6.1 (Giné and Koltchinskii (2006))**.** *Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. Let $\mathcal{F}$ be a suitably measurable class of functions on S with envelope F. Let σ^2 be a constant such that $\sup_{f\in\mathcal{F}} \operatorname{var}(f) \leq \sigma^2 \leq \|F\|^2_{L^2(P)}$. Suppose that there exist constants $A > e^2$ and $V \geq 2$ such that $\sup_Q N(\mathcal{F}, L^2(Q), \epsilon\|F\|_{L^2(Q)}) \leq (A/\epsilon)^V$ for all $0 < \epsilon \leq 1$. Then,*

$$E\left[\left\|\sum_{i=1}^n \{f(\xi_i) - E[f(\xi_1)]\}\right\|_{\mathcal{F}}\right] \leq C\left[\sqrt{n\sigma^2 V \log\frac{A\|F\|_{L^2(P)}}{\sigma}} + V\|F\|_\infty \log\frac{A\|F\|_{L^2(P)}}{\sigma}\right],$$

where C is a universal constant.

Appendix A. Proofs

A.1. **Proofs of Sections 2 and 3.**

Proof of Proposition 2.1. For any γ, $\int(\gamma' p_i)^2 dF = \int(\gamma' p_i)^2(dF/d\mu)d\mu$. Therefore, if $dF/d\mu$ is bounded above and away from zero, the result follows since p_i's are orthonormal under $(\mathcal{X}, \mu)$. □

Proof of Proposition 3.1. Fix $f \in \mathcal{G}$. Let

$$\beta_f^\star := \arg\min_n \|f - p'b\|_{F,\infty}.$$

Then

$$\|r_f\|_{F,\infty} = \|f - p'\beta_f\|_{F,\infty} \leq \|f - p'\beta_f^\star\|_{F,\infty} + \|p'\beta_f^\star - p'\beta_f\|_{F,\infty} \leq c_k + \|p'\beta_f^\star - p'\beta_f\|_{F,\infty}.$$

Further, first order conditions imply that $\beta_f = Q^{-1}E[p(x_1)f(x_1)]$, and so for any $x \in \mathcal{X}$,

$$\begin{aligned} p(x)'\beta_f^\star - p(x)'\beta_f &= p(x)'Q^{-1}Q\beta_f^\star - p(x)'Q^{-1}E[p(x_1)f(x_1)] \\ &= p(x)'Q^{-1}E[p(x_1)(p(x_1)'\beta_f^\star - f(x_1))]. \end{aligned}$$

This implies that

$$\|p'\beta_f^\star - p'\beta_f\|_{F,\infty} \le \xi_k \|E[p(x_1)(p(x_1)'\beta_f^\star - f(x_1))]\|.$$

Moreover, since $E[p(x_1)p(x_1)'] = Q = I$, $E[p_j(x_1)(p(x_1)'\beta_f^\star - f(x_1))]$ is the coefficient on $p_j(x_1)$ of the projection of $p(x_1)'\beta_f^\star - f(x_1)$ onto $p(x_1)$,

$$\|E[p(x_1)(p(x_1)'\beta_f^\star - f(x_1))]\| \le \left(E[(p(x_1)'\beta_f^\star - f(x_1))^2]\right)^{1/2} \le c_k.$$

Conclude that

$$\|r_f\|_{F,\infty} \le c_k + \xi_k c_k = c_k(1 + \xi_k),$$

and so there exists $\ell_k \le 1 + \xi_k$ such that Condition A.3 holds with this ℓ_k. This completes the proof of the proposition. □

Proof of Proposition 3.2. Fix $f \in \mathcal{G}$. Define $\beta_f^\star$ by

$$\beta_f^\star := \arg\min_b \|f - p'b\|_{F,\infty}.$$

Using the fact that for all $x \in \mathcal{X}$,

$$p(x)'(\beta_f^\star - \beta_f) = p(x)'\beta_{f - p'\beta_f^\star},$$

we obtain

$$\begin{aligned} \|r_f\|_{F,\infty} &\le \|f - p'\beta_f^\star\|_{F,\infty} + \|p'\beta_f^\star - p'\beta_f\|_{F,\infty} \\ &\le \|f - p'\beta_f^\star\|_{F,\infty} + \widetilde{\ell}_k \|f - p'\beta_f^\star\|_{F,\infty} \le (1 + \widetilde{\ell}_k) \inf_b \|f - p'b\|_{F,\infty}, \end{aligned}$$

so that

$$\|r_f\|_{F,\infty} \le (1 + \widetilde{\ell}_k) \inf_b \|f - p'b\|_{F,\infty}.$$

Next,

$$c_k \ge \sup_{f \in \mathcal{G}} \inf_b \|f - p'b\|_{F,\infty}$$

implies that

$$\|r_f\|_{F,\infty} \le c_k(1 + \widetilde{\ell}_k),$$

and so Condition A.3 holds with $\ell_k = 1 + \widetilde{\ell}_k$. This completes the proof of the proposition. □

Proof of Proposition 3.3. For a function $f(\cdot)=\sum_{j\geq 1}p_j(\cdot)j^{-\alpha}\xi_j\in\mathcal{F}(\alpha)$, let

$$A_f(x):=\sum_{j\geq k+1}p_j(x)j^{-\alpha}\xi_j \text{ and } \bar{v}_k:=d^{1/2}\sqrt{(\alpha-\beta-1/2)\log k}k^{-\alpha+\beta+1/2}.$$

Then, the statement of the lemma is equivalent to

$$P\left(\sup_{x\in\mathcal{X}}|A_f(x)|\lesssim\bar{v}_k\right)=1-o(1).$$

Therefore, it suffices to prove that

$$P\left(\sup_{x\in\mathcal{X}}|A_f(x)|\geq C\bar{v}_k\right)=o(1).$$

for some large absolute constant C.

Consider an ϵ-net $\mathcal{N}_\epsilon$ for $\mathcal{X}$. For some $L\geq 1$, let

$$\mathcal{H}_k:=\{f\in\mathcal{F}:|A_f(x)-A_f(x')|\leq L\|x-x'\| \text{ for all } x,x'\in\mathcal{X}\}.$$

Then

$$\begin{aligned}P(\sup_{x\in\mathcal{X}}|A_f(x)|\geq C\bar{v}_k)&\leq P(f\notin\mathcal{H}_k)+P(f\in\mathcal{H}_k,\sup_{x\in\mathcal{N}_\epsilon}|A_f(x)|\geq C\bar{v}_k-L\epsilon)\\&\leq P(f\notin\mathcal{H}_k)+|\mathcal{N}_\epsilon|\max_{x\in\mathcal{N}_\epsilon}P(|A_f(x)|\geq C\bar{v}_k-L\epsilon).\end{aligned}$$

Note that we can take $|\mathcal{N}_\epsilon|\leq(\mathrm{diam}(\mathcal{X})/\epsilon)^d$ and

$$\begin{aligned}E[A_f(x)^2]=E[(\sum_{j\geq k+1}j^{-\alpha}p_j(x)\xi_j)^2]=E[\sum_{j\geq k+1}j^{-2\alpha}p_j^2(x)\xi_j^2]\lesssim\sum_{j\geq k+1}j^{-2(\alpha-\beta)}\\\lesssim k^{-2(\alpha-\beta)+1}.\end{aligned}$$

Thus, setting $\epsilon=k^{-\alpha+\beta+1/2}/L$, we have $L\epsilon\lesssim\bar{v}_k$. Further, since $A_f(x)\sim N(0,E[A_f(x)^2])$, we have

$$|\mathcal{N}_\epsilon|\max_{x\in\mathcal{N}_\epsilon}P\left(|A_f(x)|\geq C\bar{v}_k-L\epsilon\right)=o(1).$$

Next, to bound $P(f\notin\mathcal{H}_k)$, note that f is L-Lipschitz if

$$Z:=\sup_{x,x'\in\mathcal{X}}\left|\frac{\sum_{j\geq k+1}\{p_j(x)-p_j(x')\}j^{-\alpha}\xi_j}{\|x-x'\|}\right|\leq L.$$

Since $\sup_{x\in\mathcal{X}}\|\nabla p_j(x)\|\leq M_j$, we have that for any $\delta\in(0,1)$,

$$\begin{aligned}P\left(Z>\sum_{j\geq k+1}j^{-\alpha}M_j\sqrt{2\log(2j^2/\delta)}\right)&\leq P(\exists j\geq k+1:|\xi_j|\geq\sqrt{2\log(2j^2/\delta)})\\&\leq\sum_{j\geq k+1}\delta/j^2\leq\delta.\end{aligned}$$

Since $\sum_{j\geq 1} j^{-\alpha} M_j \sqrt{\log j} < \infty$, $\sum_{j\geq k+1} j^{-\alpha} M_j \sqrt{\log j} \to 0$ as $k \to \infty$, and so we can find a sequence $\{\delta_k\}$ such that $\delta_k \to 0$ as $k \to \infty$ and

$$\sum_{j\geq k+1} j^{-\alpha} M_j \sqrt{2\log(2j^2/\delta_k)} \to 0.$$

Conclude that

$$P(f \notin \mathcal{H}_k) \leq P(Z > L) \lesssim P\left(Z > \sum_{j\geq k+1} j^{-\alpha} M_j \sqrt{2\log(2j^2/\delta)}\right) \to 0.$$

The asserted claim follows. □

A.2. **Proofs of Section 4.1.**

Proof of Theorem 4.1. We have that

$$\|\widehat{g} - g\|_{F,2} \leq \|p'\beta - g\|_{F,2} + \|p'\widehat{\beta} - p'\beta\|_{F,2} \leq c_k + \|p'\widehat{\beta} - p'\beta\|_{F,2}$$

where under the normalization $Q = E[p(x_i)p(x_i)'] = I$ we have

$$\|p'\widehat{\beta} - p'\beta\|_{F,2} = \left[\int (\widehat{\beta} - \beta)' p(x) p(x)' (\widehat{\beta} - \beta) dF(x)\right]^{1/2} = \|\widehat{\beta} - \beta\|.$$

To prove (4.7), we need to show $\|\widehat{\beta} - \beta\| \lesssim_P \sqrt{k/n}$. We have

$$\|\widehat{\beta} - \beta\| = \|\widehat{Q}^{-1}\mathbb{E}_n[p_i(\epsilon_i + r_i)]\| \leq \|\widehat{Q}^{-1}\mathbb{E}_n[p_i\epsilon_i]\| + \|\widehat{Q}^{-1}\mathbb{E}_n[p_i r_i]\|.$$

By the Matrix LLN (Lemma 6.2), which is the critical step, we have that

$$\|\widehat{Q} - Q\| \to_P 0 \ \text{ if } \ \frac{\xi_k^2 \log n}{n} \to 0.$$

Therefore, wp → 1, all eigenvalues of $\widehat{Q}$ are bounded away from zero. Indeed, if at least one eigenvalue of $\widehat{Q}$ is strictly smaller than $1/2$, then there exists a vector a with $a'a = 1$ such that $a'\widehat{Q}a < 1/2$, and so

$$\|\widehat{Q} - Q\| \geq |a'(\widehat{Q} - Q)a| = |a'\widehat{Q}a - a'a| = |a'\widehat{Q}a - 1| > 1/2.$$

Hence, wp → 1, all eigenvalues of $\widehat{Q}$ are not smaller than $1/2$. Therefore,

$$\|\widehat{Q}^{-1}\mathbb{E}_n[p_i\epsilon_i]\| \lesssim_P \|\mathbb{E}_n[p_i\epsilon_i]\| \lesssim_P \sqrt{k/n}$$

where the second inequality follows from

$$E\left[\|\mathbb{E}_n[p_i\epsilon_i]\|^2\right] = E[\epsilon_i^2 p_i' p_i/n] = E[\sigma_i^2 p_i' p_i/n] \lesssim E[p_i' p_i/n] = k/n$$

since $\sigma_i^2 \leq \bar{\sigma}^2$ is bounded. Moreover, since $\widehat{r}_i := p_i' \widehat{Q}^{-1} \mathbb{E}_n[p_i r_i]$ is a sample projection of r_i on p_i,

$$\|\widehat{Q}^{-1/2}\mathbb{E}_n[p_i r_i]\|^2 = \mathbb{E}_n[r_i \widehat{r}_i] = \mathbb{E}_n \widehat{r}_i^2 \leq \mathbb{E}_n[r_i^2] \lesssim_P E[r_i^2] \leq c_k^2, \tag{A.46}$$

by Markov's inequality. Therefore, when $c_k \to 0$,

$$\|\widehat{Q}^{-1}\mathbb{E}_n\left[p_i r_i\right]\| \lesssim_P \|\widehat{Q}^{-1/2}\mathbb{E}_n[p_i r_i]\| \lesssim_P c_k^2$$

where the first inequality follows from all eigenvalues of $\widehat{Q}^{1/2}$ being bounded away from zero wp → 1 and the second from (A.46). This completes the proof of (4.7).

Further, note that

$$\begin{aligned} E\left[\|\mathbb{E}_n[p_i r_i]\|\right] &= \frac{1}{n} E\left[\sqrt{\sum_{j=1}^k \left(\sum_{i=1}^n p_j(x_i) r(x_i)\right)^2}\right] \\ &\leq \frac{1}{n}\sqrt{\sum_{j=1}^k E\left[\left(\sum_{i=1}^n p_j(x_i) r(x_i)\right)^2\right]} \leq \ell_k c_k \sqrt{\frac{E[\|p(x_1)\|^2]}{n}} \leq \ell_k c_k \sqrt{\frac{k}{n}} \end{aligned} \tag{A.47}$$

where we have used $E[p_i r_i] = 0$. Alternatively, the first term in (A.47) can be bounded from above by

$$\frac{1}{n}\sqrt{E\left[\sum_{i=1}^n \sum_{j=1}^k p_j(x_i)^2 r(x_i)^2\right]} \leq \frac{1}{n}\sqrt{E\left[\sum_{i=1}^n \xi_k^2 r(x_i)^2\right]} \leq \xi_k c_k/\sqrt{n}.$$

Therefore, when $c_k \not\to 0$,

$$\|\widehat{Q}^{-1}\mathbb{E}_n[p_i r_i]\| \leq \|\widehat{Q}^{-1}\|\|\mathbb{E}_n[p_i r_i]\| \lesssim_P (\ell_k c_k \sqrt{k/n}) \wedge (\xi_k c_k/\sqrt{n}),$$

and so (4.8) follows. This completes the proof of the theorem. □

A.3. **Proofs of Section 4.2.**

Proof of Lemma 4.1. Decompose

$$\sqrt{n}\alpha'(\widehat{\beta} - \beta) = \alpha' \mathbb{G}_n[p_i(\epsilon_i + r_i)] + \alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p_i(\epsilon_i + r_i)].$$

We divide the proof in three steps. Steps 1 and 2 establish (4.11), the bound on $R_{1n}(\alpha)$. Step 3 proves (4.13), the bound on $R_{2n}(\alpha)$.

Step 1. Conditional on $X = [x_1, \ldots, x_n]$, the term

$$\alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p_i \epsilon_i]$$

has mean zero and variance bounded by $\alpha'[\widehat{Q}^{-1}-I]\widehat{Q}\bar{\sigma}^2[\widehat{Q}^{-1}-I]\alpha$. Next, as in the proof of Theorem 4.1, wp → 1, all eigenvalues of $\widehat{Q}$ are bounded away from zero and from above, and so

$$\alpha'[\widehat{Q}^{-1}-I]\widehat{Q}\bar{\sigma}^2[\widehat{Q}^{-1}-I]\alpha \lesssim \bar{\sigma}^2\|\widehat{Q}\|\|\widehat{Q}^{-1}\|^2\|\widehat{Q}-I\|^2 \lesssim_P \frac{\xi_k^2\log n}{n}$$

where the second inequality follows from Matrix LLN (Lemma 6.2) and $\bar{\sigma}^2 \lesssim 1$. We then conclude by Chebyshev's inequality that

$$\alpha'[\widehat{Q}^{-1}-I]\mathbb{G}_n[p_i\epsilon_i] \lesssim_P \sqrt{\frac{\xi_k^2\log n}{n}}.$$

Step 2. By Matrix LLN (Lemma 6.2), $\|\widehat{Q}-I\| \lesssim_P (\xi_k^2\log n/n)^{1/2}$, and so

$$\begin{aligned}|\alpha'(\widehat{Q}^{-1}-I)\mathbb{G}_n[p_ir_i]| &\leq \|\widehat{Q}^{-1}-I\|\cdot\|\mathbb{G}_n[p_ir_i]\| \\ &\leq \|\widehat{Q}^{-1}\|\cdot\|\widehat{Q}-I\|\cdot\|\mathbb{G}_n[p_ir_i]\| \lesssim_P \sqrt{\frac{\xi_k^2\log n}{n}}\ell_kc_k\sqrt{k},\end{aligned}$$

where we used the bound $\|\mathbb{G}_n[p_ir_i]\| \lesssim_P \ell_kc_k\sqrt{k}$ obtained in the proof of Theorem 4.1. Steps 1 and 2 give the linearization result (4.11).

Step 3. Since $E[p_ir_i]=0$, the term

$$R_{2n}(\alpha)=\alpha'\mathbb{G}_n[p_ir_i]$$

has mean zero and variance

$$E[\alpha'p_ir_i]^2 \leq E[\alpha'p_i]^2\ell_k^2c_k^2 \leq \ell_k^2c_k^2.$$

Thus, (4.13) follows from Chebyshev's inequality. This completes the proof of the lemma. □

Proof of Theorem 4.2. Note that for any $\alpha\in S^{k-1}$, $1\lesssim\|\alpha'\Omega^{1/2}\|$ because $1\lesssim\underline{\sigma}^2\leq\sigma_i^2$ and

$$\Omega \geq \Omega_0 \geq \underline{\sigma}^2 Q. \tag{A.48}$$

Therefore, (4.16) and (4.17) follow from (4.15), and so it suffices to prove (4.15).

By condition (iii) of the theorem and Lemma 4.1, $R_{1n}(\alpha)=o_P(1)$ (note that we can apply Lemma 4.1 because $\bar{\sigma}^2\lesssim 1$ follows from condition (i) and $\xi_k^2\log n/n\to 0$ follows from condition (iii) of the theorem). Therefore, we can write

$$\frac{\sqrt{n}\alpha'}{\|\alpha'\Omega^{1/2}\|}(\widehat{\beta}-\beta)=\frac{\alpha'}{\|\alpha'\Omega^{1/2}\|}\mathbb{G}_n[p_i(\epsilon_i+r_i)]+o_P(1)=\sum_{i=1}^n\omega_{ni}(\epsilon_i+r_i)+o_P(1),$$

where

$$\omega_{ni} = \frac{\alpha'}{\|\alpha'\Omega^{1/2}\|}\frac{p_i}{\sqrt{n}}, \quad |\omega_{ni}| \lesssim \frac{\xi_k}{\sqrt{n}}, \quad |\epsilon_i + r_i| \leq |\epsilon_i| + \ell_k c_k.$$

Further, it follows from (A.48) that

$$nE|\omega_{ni}|^2 \leq E[\alpha' p_i]^2/\alpha'\Omega\alpha \leq 1/\underline{\sigma}^2 \lesssim 1. \tag{A.49}$$

Now we verify Lindberg's condition for the CLT. First, by construction we have

$$\text{var}\left(\sum_{i=1}^n \omega_{ni}(\epsilon_i + r_i)\right) = 1.$$

Second, for each $\delta > 0$

$$\sum_{i=1}^n E\left[|\omega_{ni}|^2(\epsilon_i + r_i)^2 1\{|\omega_{ni}(\epsilon_i + r_i) > \delta\}\right] \to 0,$$

since the left hand side is bounded by

$$2nE\left[|\omega_{ni}|^2\epsilon_i^2 1\{|\epsilon_i| + \ell_k c_k > \delta/|\omega_{ni}|\}\right] \quad + \quad 2nE\left[|\omega_{ni}|^2\ell_k^2 c_k^2 1\{|\epsilon_i| + \ell_k c_k > \delta/|\omega_{ni}|\}\right],$$

and both terms go to zero. Indeed, the first term is bounded from above for some $c > 0$ by

$$\begin{aligned} &2nE\left[|\omega_{ni}|^2 E\left[\epsilon_i^2 1\{|\epsilon_i| + \ell_k c_k > c\delta\sqrt{n}/\xi_k\}|x_i\right]\right] \\ &\lesssim nE\left[|\omega_{ni}|^2\right] \cdot \sup_{x\in\mathcal{X}} E\left[\epsilon_i^2 1\{|\epsilon_i| + \ell_k c_k > c\delta\sqrt{n}/\xi_k\}|x_i = x\right] = o(1) \end{aligned}$$

where we used (A.49), the uniform integrability in the condition (i) and $c\delta\sqrt{n}/\xi_k - \ell_k c_k \to \infty$, which follows from the condition (iii); the second term is bounded from above by

$$\begin{aligned} &2nE\left[|\omega_{ni}|^2\ell_k^2 c_k^2 P\left[|\epsilon_i| + \ell_k c_k > c\delta\sqrt{n}/\xi_k|x_i\right]\right] \\ &\lesssim nE\left[|\omega_{ni}|^2\ell_k^2 c_k^2\right] \cdot \sup_{x\in\mathcal{X}} P\left[|\epsilon_i| + \ell_k c_k > c\delta\sqrt{n}/\xi_k|x_i = x\right] \\ &\lesssim \ell_k^2 c_k^2 \cdot \frac{\bar{\sigma}^2}{[c\delta\sqrt{n}/\xi_k - \ell_k c_k]^2} = o(1) \end{aligned}$$

by Chebyshev's inequality where we used (A.49), $c\delta\sqrt{n}/\xi_k - \ell_k c_k \to \infty$, and $\ell_k c_k = o(\delta\sqrt{n}/\xi_k)$. □

A.4. **Proofs of Section 4.3.**

Proof of Lemma 4.2. Decompose

$$\sqrt{n}\alpha(x)'(\widehat{\beta} - \beta) = \alpha(x)'\mathbb{G}_n[p_i(\epsilon_i + r_i)] + \alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p_i(\epsilon_i + r_i)].$$

We divide the proof in three steps. Steps 1 and 2 establish (4.19), the bound on $R_{1n}(\alpha(x))$. Step 3 proves (4.21), the bound on $R_{2n}(\alpha(x))$.

Step 1. Here we show that

$$\sup_{x \in \mathcal{X}} \left| \alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p_i \epsilon_i] \right| \lesssim_P n^{1/m} \sqrt{\frac{\xi_k^2 \log^2 n}{n}}. \tag{A.50}$$

Conditional on the data, let $T := \{t = (t_1, \dots, t_n) \in \mathbb{R}^n : t_i = \alpha(x)'(\widehat{Q}^{-1} - I)p_i \epsilon_i, x \in \mathcal{X}\}$. Define the norm $\| \cdot \|_{n,2}$ on $\mathbb{R}^n$ by $\|t\|_{n,2}^2 = n^{-1} \sum_{i=1}^n t_i^2$. Letting $\eta_1, \dots, \eta_n$ be independent Rademacher random variables ($P(\eta_1 = 1) = P(\eta_1 = -1) = 1/2$) that are independent of the data, we have for $\eta = (\eta_1, \dots, \eta_n)$ by Dudley's inequality (Dudley, 1967)

$$E_\eta \left[\sup_{x \in \mathcal{X}} |\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[\eta_i p_i \epsilon_i]| \right] \leq C \int_0^\theta \sqrt{\log N(\varepsilon, T, \| \cdot \|_{n,2})} d\varepsilon,$$

where $\theta := 2 \sup_{t \in T} \|t\|_{n,2} = 2 \sup_{x \in \mathcal{X}} \|\alpha(x)'(\widehat{Q}^{-1} - I)p_i \epsilon_i\|_{L^2(\mathbb{P}_n)} \leq 2 \max_{1 \leq i \leq n} |\epsilon_i| \|\widehat{Q}^{-1} - I\| \|\widehat{Q}\|^{1/2}$. Since for any $x, \widetilde{x} \in \mathcal{X}$,

$$\begin{aligned} &\|\alpha(x)'(\widehat{Q}^{-1} - I)p_i \epsilon_i - \alpha(\widetilde{x})'(\widehat{Q}^{-1} - I)p_i \epsilon_i\|_{L^2(\mathbb{P}_n)} \\ &\leq \max_{1 \leq i \leq n} |\epsilon_i| \|\alpha(x) - \alpha(\widetilde{x})\| \|\widehat{Q}^{-1} - I\| \|\widehat{Q}\|^{1/2} \\ &\leq \xi_k^L \max_{1 \leq i \leq n} |\epsilon_i| \|\widehat{Q}^{-1} - I\| \|\widehat{Q}\|^{1/2} \|x - \widetilde{x}\|, \end{aligned}$$

we have for some $C > 0$,

$$N(\varepsilon, T, \| \cdot \|_{n,2}) \leq \left(\frac{C \xi_k^L \max_{1 \leq i \leq n} |\epsilon_i| \| \|\widehat{Q}^{-1} - I\| \|\widehat{Q}\|^{1/2}}{\varepsilon} \right)^d.$$

Thus we have

$$\int_0^\theta \sqrt{\log N(\varepsilon, T, \| \cdot \|_{n,2})} d\varepsilon \leq \max_{1 \leq i \leq n} |\epsilon_i| \|\widehat{Q}^{-1} - I\| \|\widehat{Q}\|^{1/2} \int_0^2 \sqrt{d \log(C \xi_k^L / \varepsilon)} d\varepsilon.$$

By A.4, we have $E[\max_{1 \leq i \leq n} |\epsilon_i| \mid X] \lesssim_P n^{1/m}$ where $X = (x_1, \dots, x_n)$. In addition, note that $\xi_k^{2m/(m-2)} \log n/n \lesssim 1$ for $m > 2$ implies that $\xi_k^2 \log n/n \to 0$. Therefore, we have $\|\widehat{Q}^{-1} - I\| \lesssim_P \sqrt{\xi_k^2 \log n/n}$ and $\|\widehat{Q}\| \lesssim_P 1$. Hence, it follows from $\log \xi_k^L \lesssim \log k \lesssim \log n$ that

$$\begin{aligned} E\left[\sup_{x \in \mathcal{X}} |\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p_i \epsilon_i]| \mid X \right] &\leq 2E\left[E_\eta[\sup_{x \in \mathcal{X}} |\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[\eta_i p_i \epsilon_i]|] \mid X \right] \\ &\lesssim_P n^{1/m} \sqrt{\frac{\xi_k^2 \log^2 n}{n}}, \end{aligned}$$

where the first line is due to the symmetrization inequality. Thus, (A.50) follows.

Step 2. Observe that

$$\sup_{x\in\mathcal{X}} |\alpha(x)'(\widehat{Q}^{-1}-I)\mathbb{G}_n[p_ir_i]| \leq \|\widehat{Q}^{-1}-I\|\cdot\|\mathbb{G}_n[p_ir_i]\| \lesssim_P \sqrt{\frac{\xi_k^2\log n}{n}}\ell_k c_k\sqrt{k}$$

where the second inequality was shown in the proof of Lemma 4.1. Now, Steps 1 and 2 give the linearizarion result (4.19).

Step 3. We wish to bound $\sup_{x\in\mathcal{X}}|\alpha(x)'\mathbb{G}_n[p_ir_i]|$. We use Theorem 6.1. Consider the class of functions

$$\mathcal{F} := \{\alpha(x)'p(\cdot)r(\cdot) : x \in \mathcal{X}\}.$$

Then, $|\alpha(x)'p(\cdot)r(\cdot)| \leq \ell_k c_k \xi_k$ and for any $x, \widetilde{x} \in \mathcal{X}$,

$$|\alpha(x)'p(\cdot)r(\cdot) - \alpha(\widetilde{x})'p(\cdot)r(\cdot)| \leq \ell_k c_k \xi_k^L \xi_k \|x-\widetilde{x}\|,$$

so that

$$\sup_Q N(\mathcal{F}, L^2(Q), \varepsilon\ell_k c_k\xi_k) \leq \left(\frac{C\xi_k^L}{\varepsilon}\right)^d.$$

Thus, by Theorem 6.1, we have

$$E\left[\sup_{x\in\mathcal{X}}|\alpha(x)'\mathbb{G}_n[p_ir_i]|\right] \lesssim \ell_k c_k\sqrt{\log n} + \ell_k c_k\frac{\xi_k\log n}{\sqrt{n}} \lesssim \ell_k c_k\sqrt{\log n},$$

where we have used the fact that

$$\frac{\xi_k\log n}{\sqrt{n}} = \sqrt{\log n}\sqrt{\frac{\xi_k^2\log n}{n}} = o(\sqrt{\log n}).$$

Therefore, we have by Markov's inequality

$$\sup_{x\in\mathcal{X}}|\alpha(x)'\mathbb{G}_n[p_ir_i]| \lesssim_P \ell_k c_k\sqrt{\log n}. \tag{A.51}$$

So, the linearization result (4.21) follows. This completes the proof. □

Proof of Theorem 4.3. Note that (4.23) and (4.24) follow from (4.22) and Lemma 4.2. Therefore, it suffices to prove (4.22), and so we wish to bound $\sup_{x\in\mathcal{X}}|\alpha(x)'\mathbb{G}_n[p_i\epsilon_i]|$. To this end, we use Proposition 6.1. Consider the class of functions

$$\mathcal{G} := \{(\epsilon, x) \mapsto \epsilon\alpha(v)'p(x) : v \in \mathcal{X}\}.$$

Then, $|\alpha(v)'p(x_i)| \leq \xi_k$, $\text{var}(\alpha(v)'p(x_i)) = 1$ and for any $v, \widetilde{v} \in \mathcal{X}$,

$$|\epsilon\alpha(v)'p(x) - \epsilon\alpha(\widetilde{v})'p(x)| \leq |\epsilon|\xi_k^L\xi_k\|v-\widetilde{v}\|.$$

Thus, taking $G(\epsilon, x) := |\epsilon|\xi_k$, we have

$$\sup_Q N(\mathcal{G}, L^2(Q), \varepsilon\|G\|_{L^2(Q)}) \leq \left(\frac{C\xi_k^L}{\varepsilon}\right)^d.$$

Therefore, by Proposition 6.1, we have

$$E\left[\sup_{x\in\mathcal{X}} |\alpha(x)'\mathbb{G}_n[p_i\epsilon_i]|\right] \lesssim \sqrt{\log n} + \frac{\xi_k^{m/(m-2)}\log n}{\sqrt{n}} \lesssim \sqrt{\log n}, \qquad \text{(A.52)}$$

where we have used the following inequality

$$\frac{\xi_k^{m/(m-2)}\log n}{\sqrt{n}} = \sqrt{\log n}\cdot\sqrt{\frac{\xi_k^{2m/(m-2)}\log n}{n}} \lesssim \sqrt{\log n}.$$

This completes the proof. □

Proof of Theorem 4.4. The proof follows similarly to that in Chernozhukov et al. (2009). We shall apply Yurinskii's coupling (see Theorem 10 in Pollard (2002)):

Let $\zeta_1, \dots, \zeta_n$ be independent K-vectors with $E[\zeta_i] = 0$ for each i, and $\Delta := \sum_{i=1}^n E\|\zeta_i\|^3$ finite. Let S denote denote a copy of $\zeta_1 + \cdots + \zeta_n$ on a sufficiently rich probability space $(\Omega, \mathcal{A}, P)$. For each $\delta > 0$ there exists a random vector T in this space with a $N(0, \mathrm{var}(S))$ distribution such that

$$P\{\|S - T\| > 3\delta\} \leq C_0 B\left(1 + \frac{|\log(1/B)|}{K}\right) \text{ where } B := \Delta K\delta^{-3},$$

for some universal constant C_0.

In order to apply the coupling, consider a copy of the first order approximation to our estimator on a suitably rich probability space

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_i, \quad \zeta_i = \Omega^{-1/2}p_i(\epsilon_i + r_i).$$

When $\bar{R}_{2n} = o_P(a_n^{-1})$, a similar argument can be used with $\zeta_i = \Omega^{-1/2}p_i(\epsilon_i + r_i)$ replaced by $\zeta_i = \Omega^{-1/2}p_i\epsilon_i$. As in the proof of Theorem 4.2, all eigenvalues of Ω are bounded away from zero. Therefore,

$$\begin{aligned} E\|\zeta_i\|^3 &\lesssim E\|p_i(\epsilon_i + r_i)\|^3 \\ &\lesssim E[\|p_i\|^3(|\epsilon_i|^3 + |r_i|^3)] \\ &\lesssim E[\|p_i\|^3](1 + \ell_k^3 c_k^3) \\ &\lesssim E[\|p_i\|^2]\xi_k(1 + \ell_k^3 c_k^3) \\ &\lesssim k\xi_k(1 + \ell_k^3 c_k^3) \end{aligned}$$

where we used the assumption that $\sup_{x\in\mathcal{X}} E[|\epsilon_i|^3|x_i = x] \lesssim 1$. Therefore, by Yurinskii's coupling, for each $\delta > 0$,

$$P\left\{\left\|\frac{\sum_{i=1}^n \zeta_i}{\sqrt{n}} - \mathcal{N}_k\right\| \geq 3\delta a_n^{-1}\right\} \lesssim \frac{nk^2\xi_k(1+\ell_k^3 c_k^3)}{(\delta a_n^{-1}\sqrt{n})^3}\left(1+\frac{\log(k^2\xi_k(1+\ell_k^3c_k^3))}{k}\right)$$
$$\lesssim \frac{a_n^3 k^2 \xi_k(1+\ell_k^3c_k^3)}{\delta^3 n^{1/2}}\left(1+\frac{\log n}{k}\right) \to 0$$

because $a_n^6 k^4 \xi_k^2 (1+\ell_k^3 c_k^3)^2 \log^2 n/n \to 0$.

Hence, using (4.18) and (4.19), we obtain

$$\|\sqrt{n}\alpha(x)'(\widehat{\beta}-\beta) - \alpha(x)'\Omega^{1/2}\mathcal{N}_k\| \leq \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha(x)'\Omega^{1/2}\zeta_i - \alpha(x)'\Omega^{1/2}\mathcal{N}_k\right\| + \bar{R}_{1n} = o_P(a_n^{-1}).$$

Since $\|\alpha(x)'\Omega^{1/2}\|$ is bounded from below uniformly over $x \in \mathcal{X}$, we conclude that (4.25) holds, and (4.26) is a direct consequence of (4.25).

Further, under the assumption that $\sup_{x\in\mathcal{X}} n^{1/2}|r(x)|/\|s(x)\| = o_P(a_n^{-1})$,

$$\frac{\sqrt{n}p(x)'(\widehat{\beta}-\beta)}{\|s(x)\|} - \frac{\sqrt{n}(\widehat{g}(x)-g(x))}{\|s(x)\|} = o_P(a_n^{-1}),$$

so that (4.27) follows. This completes the proof of the theorem. □

Proof of Theorem 4.5. Note that $\widehat{\beta}^b$ solves the least squares problem for the rescaled data $\{(\sqrt{h_i}y_i, \sqrt{h_i}p_i) : i = 1,\ldots,n\}$. The weight h_i is independent of (y_i, p_i), $E[h_i] = 1$, $E[h_i^2] = 1$, and $\max_{1\leq i\leq n} h_i \lesssim_P \log n$. That allows us to extend all results from $\widehat{\beta}$ to $\widehat{\beta}^b$ replacing ξ_k by $\xi_k^b = \xi_k \log n$ to account for the larger envelope.

We apply Lemma 4.2 to the original problem (2.3) and to the weighted problem (4.30). Then

$$\sqrt{n}\alpha(x)'\left(\widehat{\beta}^b - \widehat{\beta}\right) = \sqrt{n}\alpha(x)'\left(\widehat{\beta}^b - \beta\right) + \sqrt{n}\alpha(x)'\left(\beta - \widehat{\beta}\right)$$
$$= \alpha(x)'\mathbb{G}_n[(h_i-1)p_i(\epsilon_i + r_i)] + R_{1n}^b(\alpha(x))$$

where

$$R_{1n}^b(\alpha(x)) \lesssim_P \sqrt{\frac{\xi_k^2\log^3 n}{n}}(n^{1/m}\sqrt{\log n} + \sqrt{k}\ell_k c_k)$$

uniformly over $x \in \mathcal{X}$, and so (4.31) follows.

Further, (4.32) follows similarly to Theorem 4.4 by applying Yurinskii's coupling for the weighted process with weights $v_i = h_i - 1$ so that $E[v_i^2] = 1$ and $E[|v_i|^3] \lesssim 1$. Thus there is a Gaussian random vector $\mathcal{N}_k \sim N(0, I_k)$ such that

$$\left\| \frac{\Omega^{-1/2}}{\sqrt{n}} \sum_{i=1}^{n} (h_i - 1) p_i (\epsilon_i + r_i) - \mathcal{N}_k \right\| = o_P(a_n^{-1}). \tag{A.53}$$

Combining (A.53) with (4.31) yields (4.32) by the triangle inequality as in the proof of Theorem 4.4, and (4.33) follows from (4.32).

Note also that the results continue to hold in P-probability if we replace P by $P^*(\cdot|D)$, since $B_n \lesssim_P 1$ implies that $B_n \lesssim_{P^*} 1$. Indeed, the first relation means that $P(|B_n| > \ell_n) = o(1)$ for any $\ell_n \to \infty$, while the second means that $P^*(|B_n| > \ell_n) = o_P(1)$ for any $\ell_n \to \infty$. But the second clearly follows from the first by Markov inequality because $E[P^*(|B_n| > \ell_n)] = P(|B_n| > \ell_n) = o(1)$. □

Proof of Theorem 4.6. Note that it follows from $\bar{R}_{2n} \lesssim (\log n)^{1/2}$ that $\ell_k c_k \lesssim 1$ (see the definition of $\bar{R}_{2n}$ in (4.21)). Therefore, $\|\Sigma\| \lesssim (1 + \ell_k c_k)\|Q\| \lesssim 1$. In addition, it follows from Condition A.4 that $v_n \lesssim n^{1/m}$, and so $\bar{R}_{1n} \lesssim (\log n)^{1/2}$ implies that

$$(v_n \vee 1 + \ell_k c_k) \sqrt{\frac{\xi_k^2 \log n}{n}} \to 0.$$

Further, the first result follows from the Markov inequality and Matrix LLN (Lemma 6.2), which shows that $E[\|\widehat{Q} - Q\|] \lesssim (\xi_k^2 \log n/n)^{1/2} \to 0$.

To establish the second result, we note that

$$\widehat{\Sigma} - \Sigma = \mathbb{E}_n[(\widehat{\epsilon}_i^2 - \{\epsilon_i + r_i\}^2) p_i p_i'] + \mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p_i'] - \Sigma. \tag{A.54}$$

The first term on the right hand side of (A.54) satisfies

$$\begin{aligned} \|\mathbb{E}_n[(\widehat{\epsilon}_i^2 - \{\epsilon_i + r_i\}^2) p_i p_i']\| &\le \|\mathbb{E}_n[\{p_i'(\widehat{\beta} - \beta)\}^2 p_i p_i']\| + 2\|\mathbb{E}_n[(\epsilon_i + r_i) p_i'(\widehat{\beta} - \beta) p_i p_i']\| \\ &\le \max_{1 \le i \le n} |p_i'(\widehat{\beta} - \beta)|^2 \|\mathbb{E}_n[p_i p_i']\| + \max_{1 \le i \le n} (|\epsilon_i| + |r_i|) \max_{1 \le i \le n} |p_i'(\widehat{\beta} - \beta)| \|\mathbb{E}_n[p_i p_i']\| \\ &\lesssim_P \|\widehat{Q}\| \frac{\xi_k^2(\sqrt{\log n} + \bar{R}_{1n} + \bar{R}_{2n})^2}{n} + (v_n \vee 1 + \ell_k c_k) \|\widehat{Q}\| \frac{\xi_k(\sqrt{\log n} + \bar{R}_{1n} + \bar{R}_{2n})}{\sqrt{n}} \end{aligned}$$

since $\max_{1 \le i \le p} |p_i'(\widehat{\beta} - \beta)|^2 \lesssim_P \xi_k^2(\sqrt{\log n} + \bar{R}_{1n} + \bar{R}_{2n})^2/n$ by Theorem 4.3, $\max_{1 \le i \le n} |r_i| \le \ell_k c_k$, and $\max_{1 \le i \le n} |\epsilon_i|^2 \lesssim_P v_n^2$ by Markov's inequality. Therefore,

$$\|\mathbb{E}_n[(\widehat{\epsilon}_i^2 - \{\epsilon_i + r_i\}^2) p_i p_i']\| \lesssim_P (v_n \vee 1 + \ell_k c_k) \sqrt{\frac{\xi_k^2 \log n}{n}}$$

because $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log n)^{1/2}$, $\|\widehat{Q}\| \lesssim_{\mathrm{P}} 1$ by the first result, $\xi_k^2 \log n/n \to 0$, and $v_n \vee 1 + \ell_k c_k$ is bounded away from zero.

To control the second term in (A.54), let $\eta_1, \dots, \eta_n$ be a sequence of independent Rademacher random variables ($P(\eta_1 = 1) = P(\eta_1 = -1) = 1/2$) that are independent of the data. Then for $\eta = (\eta_1, \dots, \eta_n)$,

$$\begin{aligned}
& E\left[\|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p_i'] - \Sigma\|\right] \\
& \lesssim E\left[E_\eta\left[\|\mathbb{E}_n[\eta_i\{\epsilon_i + r_i\}^2 p_i p_i']\|\right]\right] \\
& \lesssim \sqrt{\frac{\log n}{n}} E\left[\left(\|\mathbb{E}_n[\{\epsilon_i + r_i\}^4 \|p_i\|^2 p_i p_i']\|\right)^{1/2}\right] \\
& \leq \sqrt{\frac{\xi_k^2 \log n}{n}} E\left[\max_{1\leq i\leq n} |\epsilon_i + r_i| \left(\|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p_i']\|\right)^{1/2}\right] \\
& \leq \sqrt{\frac{\xi_k^2 \log n}{n}} \left(E\left[\max_{1\leq i\leq n} |\epsilon_i + r_i|^2\right]\right)^{1/2} \left(E\left[\|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p_i']\|\right]\right)^{1/2}
\end{aligned}$$

where the first inequality holds by Symmetrization Lemma (Lemma 2.3.6 van der Vaart and Wellner (1996)), the second by Khinchin inequality (Lemma 6.1), the third by $\max_{1\leq i\leq n} \|p_i\| \leq \xi_k$, and the fourth by the Cauchy-Schwarz inequality.

Since for any positive numbers a, b, and R, $a \leq R(a+b)^{1/2}$ implies $a \leq R^2 + R\sqrt{b}$, the expression above using the triangle inequality yields

$$E\left[\|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p_i'] - \Sigma\|\right] \lesssim \frac{\xi_k^2 \log n}{n}(v_n^2 + \ell_k^2 c_k^2) + \left(\frac{\xi_k^2 \log n}{n}\{v_n^2 + \ell_k^2 c_k^2\}\right)^{1/2} \|\Sigma\|^{1/2},$$

and so

$$E\left[\|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p_i'] - \Sigma\|\right] \lesssim (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log n}{n}}$$

because $\|\Sigma\| \lesssim 1$ and $(v_n^2 + \ell_k^2 c_k^2)\xi_k^2 \log n/n \to 0$. Now, the second result follows from Markov's inequality.

Finally, we have

$$\|\widehat{\Omega} - \Omega\| \lesssim \|(\widehat{Q}^{-1} - Q^{-1})\widehat{\Sigma}\widehat{Q}^{-1}\| + \|Q^{-1}(\widehat{\Sigma} - \Sigma)\widehat{Q}^{-1}\| + \|Q^{-1}\Sigma(\widehat{Q}^{-1} - Q^{-1})\| = o_P(1/a_n)$$

whenever $\|\widehat{Q} - Q\| = o_P(1/a_n)$ and $\|\widehat{\Sigma} - \Sigma\| = o_P(1/a_n)$ because eigenvalues of both Q and Σ are bounded away from zero and from above. We can set $a_n = (v_n \vee 1 + \ell_k c_k)(\xi_k^2 \log n/n)^{1/2}$. This gives the third result of the theorem and completes the proof. □

A.5. **Proofs of Section 5.**

Proof of Lemma 5.1. As in the proof of Theorem 4.2, all eigenvalues of Ω are bounded away from zero. Therefore, by the triangle inequality,

$$\left|\frac{\widehat{\sigma}_\theta(w)}{\sigma_\theta(w)} - 1\right| \leq \frac{\|\ell_\theta(w)'(\widehat{\Omega}^{1/2} - \Omega^{1/2})\|}{\|\ell_\theta(w)'\Omega^{1/2}\|} \lesssim_P \|\widehat{\Omega}^{1/2} - \Omega^{1/2}\|. \tag{A.55}$$

To bound $\|\widehat{\Omega}^{1/2} - \Omega^{1/2}\|$, we shall use the following lemma:

Lemma A.1. *Let A and B be $k \times k$ symmetric positive semidefinite matrices. Assume that B is positive definite. Then $\|A^{1/2} - B^{1/2}\| \leq \|A - B\| \|B^{-1}\|^{1/2}$.*

Proof of Lemma A.1. This is exercise 7.2.18 in Horn and Johnson (1990). For completeness, we derive this result here. Let a be an eigenvector of $E = A^{1/2} - B^{1/2}$ with eigenvalue $\lambda = \|A^{1/2} - B^{1/2}\|$. Then

$$\begin{aligned}
\|A - B\| &\geq |a'(A - B)a| \\
&= |a'(A^{1/2}E + EA^{1/2} - E^2)a| \\
&= |\lambda a'(A^{1/2} + A^{1/2} - E)a| \\
&= \lambda |a'(A^{1/2} + B^{1/2})a| \\
&\geq \lambda |\lambda_{\min}(A^{1/2}) + \lambda_{\min}(B^{1/2})|
\end{aligned}$$

where $\lambda_{\min}(P)$ denotes the minimal eigenvalue of P for $P = A^{1/2}$ or $B^{1/2}$. Since A is positive semidefinite, $\lambda_{\min}(A^{1/2}) \geq 0$. Since B is positive definite, $\lambda_{\min}(B^{1/2}) = \|B^{-1}\|^{-1/2}$. Combining these bounds gives the asserted claim. □

Therefore,

$$\|\widehat{\Omega}^{1/2} - \Omega^{1/2}\| \lesssim \|\widehat{\Omega} - \Omega\| \lesssim_P (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log n}{n}} = o(1) \tag{A.56}$$

where the second inequality and the last equality follow from Theorem 4.6. Combining (A.55) and (A.56) gives the asserted claim. □

A.6. **Proofs of Section 5.1.**

Proof of Theorem 5.1. Fix $w \in \mathcal{I}$. Denote $\alpha := \ell_\theta(w)/\|\ell_\theta(w)\|$. Then

$$\begin{aligned}
|\widehat{\theta}(w) - \theta(w)| &\leq |\ell_\theta(w)'(\widehat{\beta} - \beta)| + |r_\theta(w)| \\
&\leq |\ell_\theta(w)'\mathbb{G}_n[p_i\epsilon_i]|/\sqrt{n} + \|\ell_\theta(w)\| \left(|R_{1n}(\alpha)| + |R_{2n}(\alpha)|\right)/\sqrt{n} + o(\|\ell_\theta(w)\|/\sqrt{n})
\end{aligned}$$

where the second line follows from Lemma 4.1 and condition (i). Next, note that by Lemma 4.1,

$$|R_{1n}(\alpha)| + |R_{2n}(\alpha)| \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}}(1 + \sqrt{k}\ell_k c_k) + \ell_k c_k = o(1)$$

where the last conclusion holds from conditions (iii) and (iv). Finally, condition (ii) implies that

$$E[|\ell_\theta(w)'\mathbb{G}_n[p_i\epsilon_i]|^2] \lesssim \|\ell_\theta(w)\|^2\bar{\sigma}^2\|Q\|^2 \lesssim \|\ell_\theta(w)\|^2,$$

and so the result follows by applying Chebyshev's inequality. □

Proof of Theorem 5.2. Under our conditions, all eigenvalues of Ω are bounded away from zero. Therefore,

$$\frac{r_\theta(w)}{\widehat{\sigma}_\theta(w)} \lesssim_P \frac{r_\theta(w)}{\sigma_\theta(w)} \lesssim \frac{\sqrt{n} r_\theta(w)}{\ell_\theta(w)} \to 0$$

where the first inequality follows from Lemma 5.1. In addition, by Theorem 4.2,

$$\frac{\ell_\theta(w)'(\widehat{\beta} - \beta)}{\sigma_\theta(w)} \to_d N(0,1).$$

Hence,

$$t(w) = \frac{\ell_\theta(w)'(\widehat{\beta} - \beta)}{\widehat{\sigma}_\theta(w)} - \frac{r_\theta(w)}{\widehat{\sigma}_\theta(w)} = \frac{\ell_\theta(w)'(\widehat{\beta} - \beta)}{(1 + o_P(1))\sigma_\theta(w)} + o_P(1) \to_d N(0,1)$$

by Slutsky's lemma. This completes the proof of the theorem. □

A.7. **Proofs of Section 5.2.**

Proof of Lemma 5.2. By triangle inequality,

$$\begin{aligned}\left\|\frac{\ell_\theta(w_1)}{\|\ell_\theta(w_1)\|} - \frac{\ell_\theta(w_2)}{\|\ell_\theta(w_2)\|}\right\| &\le \frac{\|\ell_\theta(w_1) - \ell_\theta(w_2)\|}{\|\ell_\theta(w_1)\|} + \|\ell_\theta(w_2)\|\left|\frac{1}{\|\ell_\theta(w_1)\|} - \frac{1}{\|\ell_\theta(w_2)\|}\right| \\ &\le \frac{2\|\ell_\theta(w_1) - \ell_\theta(w_2)\|}{\|\ell_\theta(w_1)\|} \lesssim \xi_{k,\theta}^L\|w_1 - w_2\|\end{aligned}$$

uniformly over $w_1, w_2 \in \mathcal{I}$ where the last inequality follows from the definition of $\xi_{k,\theta}^L$ and the condition that $1/\|l_\theta(w)\| \lesssim 1$ uniformly over $w \in \mathcal{I}$. Therefore, the proof follows from the same arguments as those given for Lemma 4.2. □

Proof of Theorem 5.3. Given discussion in the proof of Lemma 5.2, (5.35) follows from the same arguments as those used for Theorem 4.3, equation (4.22).

Now we prove (5.36). By the triangle inequality,

$$\sup_{w\in\mathcal{I}} |\widehat{\theta}(w) - \theta(w)| \le \sup_{w\in\mathcal{I}} |\ell_\theta(w)'(\widehat{\beta} - \beta)| + \sup_{w\in\mathcal{I}} |r_\theta(w)|. \tag{A.57}$$

Further,

$$\sup_{w\in\mathcal{I}}|r_\theta(w)| \leq \sup_{w\in\mathcal{I}}\frac{|r_n(w)|}{\|\ell_\theta(w)\|}\sup_{w\in\mathcal{I}}\|\ell_\theta(w)\| \lesssim \sqrt{\frac{\xi_{k,\theta}^2\log n}{n}} \tag{A.58}$$

by the condition (ii) and the definition of $\xi_{k,\theta}$. In addition, by Lemma 5.2 and (5.35),

$$\sup_{w\in\mathcal{I}}|\ell_\theta(w)'(\widehat{\beta}-\beta)| \lesssim_P \frac{1}{\sqrt{n}}\left(\left|\sup_{w\in\mathcal{I}}\alpha_\theta(w)'\mathbb{G}_n[p_i\epsilon_i]\right| + \bar{R}_{1n} + \bar{R}_{2n}\right)\sup_{w\in\mathcal{I}}\|\ell_\theta(w)\| \tag{A.59}$$

$$\lesssim_P \sqrt{\frac{\log n}{n}}\sup_{w\in\mathcal{I}}\|\ell_\theta(w)\| \lesssim \sqrt{\frac{\xi_{k,\theta}^2\log n}{n}}. \tag{A.60}$$

Combining (A.57), (A.58), (A.59), and (A.60) gives the asserted claim. □

Proof of Theorem 5.4. Under our conditions,

$$(v_n \vee 1 + \ell_k c_k)\frac{\xi_k \log n}{\sqrt{n}} = o(a_n^{-1}).$$

Further, as in the proof of Theorem 4.4 and using Lemma 5.2, we can find $\mathcal{N}_k \sim N(0, I_k)$ such that

$$\left\|\sqrt{n}\alpha_\theta(w)'(\widehat{\beta}-\beta) - \alpha_\theta(w)'\Omega^{1/2}\mathcal{N}_k\right\| = o_P(a_n^{-1})$$

uniformly over $w \in \mathcal{I}$. Since $\|\alpha_\theta(w)'\Omega^{1/2}\|$ is bounded away from zero uniformly over $w \in \mathcal{I}$,

$$\left\|\sqrt{n}\frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\|\ell_\theta(w)'\Omega^{1/2}\|} - \frac{\ell_\theta(w)'\Omega^{1/2}\mathcal{N}_k}{\|\ell_\theta(w)'\Omega^{1/2}\|}\right\| = o_P(a_n^{-1}),$$

or, equivalently,

$$\left\|\frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\sigma_\theta(w)} - \frac{\ell_\theta(w)'\Omega^{1/2}\mathcal{N}_k/\sqrt{n}}{\sigma_\theta(w)}\right\| = o_P(a_n^{-1}),$$

uniformly over $w \in \mathcal{I}$. Further,

$$\begin{aligned}\left|\frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\sigma_\theta(w)} - \frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\widehat{\sigma}_\theta(w)}\right| &\leq \frac{|\ell_\theta(w)'(\widehat{\beta}-\beta)|}{\sigma_\theta(w)}\left|1-\frac{\sigma_\theta(w)}{\widehat{\sigma}_\theta(w)}\right| \\ &\lesssim \sqrt{n}|\alpha_\theta(w)'(\widehat{\beta}-\beta)|\left|1-\frac{\sigma_\theta(w)}{\widehat{\sigma}_\theta(w)}\right| \\ &\lesssim_P \sqrt{\log n}(v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2\log n}{n}} = o(a_n^{-1})\end{aligned}$$

uniformly over $w \in \mathcal{I}$ where the second line follows from $\|\alpha_\theta(w)'\Omega^{1/2}\|$ being bounded away from zero uniformly over $w \in I$ and the third line follows from Lemmas 5.1 and 5.2 and

Theorem 5.3. Therefore,

$$\left\| \frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\widehat{\sigma}_\theta(w)} - \frac{\ell_\theta(w)'\Omega^{1/2}\mathcal{N}_k/\sqrt{n}}{\sigma_\theta(w)} \right\| = o_P(a_n^{-1}) \tag{A.61}$$

uniformly over $w \in \mathcal{I}$. In addition, $\sup_{w\in\mathcal{I}} |r_\theta(w)|/\sigma_\theta(w) = o_P(a_n^{-1})$ uniformly over $w \in \mathcal{I}$ and Lemma 5.1 imply that $\sup_{w\in\mathcal{I}} |r_\theta(w)|/\widehat{\sigma}_\theta(w) = o_P(a_n^{-1})$, and so it follows from (A.61) that

$$\left\| \frac{\widehat{g}(w)-g(w)}{\widehat{\sigma}_\theta(w)} - \frac{\ell_\theta(w)'\Omega^{1/2}\mathcal{N}_k/\sqrt{n}}{\sigma_\theta(w)} \right\| = o_P(a_n^{-1})$$

uniformly over $w \in \mathcal{I}$. This completes the proof of the theorem. □

Proof of Theorem 5.5. We have

$$\frac{\widehat{\theta}(w)-\theta(w)}{\widehat{\sigma}_\theta(w)} = \frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\widehat{\sigma}_\theta(w)} - \frac{r_\theta(w)}{\widehat{\sigma}_\theta(w)}. \tag{A.62}$$

Under the conditions $\bar{R}_{2n} \lesssim 1/(\log n)^{1/2}$ and $\xi_k \log^2 n/n^{1/2-1/m} \to 0$,

$$\left| \frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\widehat{\sigma}_n(w)} - \frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\sigma_\theta(w)} \right| \lesssim_P \frac{1}{\sqrt{\log n}} \tag{A.63}$$

uniformly over $w \in \mathcal{I}$ by the argument used in the proof of Theorem 5.4 with $a_n = 1/(\log n)^{1/2}$. Further, by Lemma 5.2,

$$\frac{\ell_\theta(w)'(\widehat{\beta}-\beta)}{\sigma_\theta(w)} = \frac{\ell_\theta(w)'\mathbb{G}_n[p_i\epsilon_i]}{\sigma_\theta(w)} + o_P\left(\frac{1}{\sqrt{\log n}}\right) \tag{A.64}$$

uniformly over $w \in \mathcal{I}$ since $\bar{R}_{1n}+\bar{R}_{2n} \lesssim 1/(\log n)^{1/2}$. In addition, as in the proof of Theorem 5.4 with $a_n = 1/(\log n)^{1/2}$,

$$\frac{|r_\theta(w)|}{\widehat{\sigma}_\theta(w)} \lesssim_P \frac{1}{\sqrt{\log n}} \tag{A.65}$$

uniformly over $w \in \mathcal{I}$. Combining (A.62), (A.63), (A.64), and (A.65) yields

$$\frac{\widehat{\theta}(w)-\theta(w)}{\widehat{\sigma}_\theta(w)} = \frac{\ell_\theta(w)'\mathbb{G}_n[p_i\epsilon_i]}{\sigma_\theta(w)} + o_P\left(\frac{1}{\sqrt{\log n}}\right). \tag{A.66}$$

Now, under the condition $\xi_k \log^2 n/n^{1/2-1/m} \to 0$, the asserted claim follows from Proposition 3.3 in Chernozhukov et al. (2012) applied to the first term on the right hand side of (A.66) (note that Proposition 3.3 in Chernozhukov et al. (2012) only considers a special case where $\ell_\theta(w)$, $w \in \mathcal{I}$, is replaced by $p(x)$, $x \in \mathcal{X}$, but the same proof applies for a more general case studied here, with $\ell_\theta(w)$, $w \in \mathcal{I}$). □

Proof of Theorem 5.6. The proof consists of two steps. The asserted claims are proven in Step 1, and Step 2 contains some intermediate calculations.

Step 1. Under our conditions, it follows from Step 2 that there exists a sequence $\{\varepsilon_n\}$ such that $\varepsilon_n = o(1)$ and

$$P\left\{\left|\sup_{w\in\mathcal{I}}|\widehat{t}_n^\star(w)| - \sup_{w\in\mathcal{I}}|t_n^\star(w)|\right| > \varepsilon_n/\sqrt{\log n}\right\} = o(1). \tag{A.67}$$

Let $c_n^0(1-\alpha)$ denote the $(1-\alpha)$-quantile of $\sup_{w\in\mathcal{I}}|t_n^\star(w)|$. Then in view of (A.67), Lemma A.3 implies that there exists a sequence $\{\nu_n\}$ such that $\nu_n = o(1)$ and

$$P\left\{c_n(1-\alpha) < c_n^0(1-\alpha-\nu_n) - \varepsilon_n/\sqrt{\log n}\right\} = o(1), \tag{A.68}$$

$$P\left\{c_n(1-\alpha) > c_n^0(1-\alpha+\nu_n) + \varepsilon_n/\sqrt{\log n}\right\} = o(1). \tag{A.69}$$

Further, it follows from Theorem 5.5 that there exists a sequence $\{\beta_n\}$ of constants and a sequence $\{Z_n\}$ of random variables such that $\beta_n = o(1)$, Z_n equals in distribution to $\|t_n^\star\|_{\mathcal{I}}$, and

$$P\left\{\left|\sup_{w\in\mathcal{I}}|t_n(w)| - Z_n\right| > \beta_n/\sqrt{\log n}\right\} = o(1). \tag{A.70}$$

Hence, for some universal constant A,

$$\begin{aligned}
P(\sup_{w\in\mathcal{I}}|t_n(w)| \le c_n(1-\alpha)) &\le P(Z_n \le c_n(1-\alpha) + \beta_n/\sqrt{\log n}) + o(1)\\
&\le P(Z_n \le c_n^0(1-\alpha+\nu_n) + (\varepsilon_n+\beta_n)/\sqrt{\log n}) + o(1)\\
&\le P(Z_n \le c_n^0(1-\alpha+\nu_n + A(\varepsilon_n+\beta_n))) + o(1)\\
&= 1-\alpha+\nu_n + A(\varepsilon_n+\beta_n) + o(1)\\
&= 1-\alpha + o(1)
\end{aligned}$$

where the first inequality follows from (A.70), the second from (A.69), and the third from Lemma 5.3. This gives one side of the bound in (5.41). The other side of the bound can be proven by a similar argument. Therefore, (5.41) follows. Further, (5.42) is a direct consequence of (5.41).

Finally, we consider (5.43). The second inequality in (5.43) holds because $\sigma_\theta(w) \lesssim \|\ell_\theta(w)\|/n^{1/2}$ since all eigenvalues of Ω are bounded from above. To prove the first inequality, note that by Lemma 5.1, $\widehat{\sigma}_\theta(w)/\sigma_\theta(w) = 1 + o_P(1)$ uniformly over $w \in \mathcal{I}$. In addition, Step 2 shows that

$$c_n(1-\alpha) \lesssim_P \sqrt{\log n}. \tag{A.71}$$

Therefore, $2c_n(1-\alpha)\widehat{\sigma}_n(w) \lesssim_P (\log n)^{1/2}\sigma_\theta(w)$, uniformly over $w \in \mathcal{I}$, which is the first inequality in (5.43). To complete the proof, we provide auxilliary calculations in Step 2.

Step 2. We first prove (A.67). Note that

$$\left|\sup_{w\in\mathcal{I}}|\widehat{t}^\star_n(w)| - \sup_{w\in\mathcal{I}}|t^\star_n(w)|\right| \leq \sup_{w\in\mathcal{I}}\left|\widehat{t}^*_n(w) - t^*_n(w)\right| = \sup_{w\in\mathcal{I}}\left|\left(\frac{\ell_\theta(w)'\widehat{\Omega}^{1/2}}{\sqrt{n}\widehat{\sigma}_n(w)} - \frac{\ell_\theta(w)'\Omega^{1/2}}{\sqrt{n}\sigma_\theta(w)}\right)\mathcal{N}_k\right|.$$

Denote $T_n(w) := \widehat{t}^*_n(w) - t^*_n(w)$. Then, conditionally on the data, $\{T_n(w), w \in \mathcal{I}\}$ is a zero-mean Gaussian process. Further, we have for $E_{\mathcal{N}_k}[\cdot]$ denoting the expectation with respect to the distribution of $\mathcal{N}_k$,

$$\begin{aligned}
E_{\mathcal{N}_k}[T_n(w)^2]^{1/2} &= \left\|\frac{\ell_\theta(w)'\widehat{\Omega}^{1/2}}{\sqrt{n}\widehat{\sigma}_n(w)} - \frac{\ell_\theta(w)'\Omega^{1/2}}{\sqrt{n}\sigma_\theta(w)}\right\| \\
&\leq \frac{\|\ell_\theta(w)\|}{\sqrt{n}\widehat{\sigma}_n(w)}\|\widehat{\Omega}^{1/2} - \Omega^{1/2}\| + \left\|\frac{\ell_\theta(w)'\Omega^{1/2}}{\sqrt{n}\sigma_\theta(w)}\right\|\left|\frac{\sigma_\theta(w)}{\widehat{\sigma}_n(w)} - 1\right| \\
&\lesssim_P \|\widehat{\Omega}^{1/2} - \Omega^{1/2}\| + \left|\frac{\sigma_\theta(w)}{\widehat{\sigma}_n(w)} - 1\right| \\
&\lesssim_P \|\widehat{\Omega} - \Omega\| = o_P\left(\frac{1}{\sqrt{\log n}}\right)
\end{aligned}$$

uniformly over $w \in \mathcal{I}$ where the last line follows from the proof of Lemma 5.1. In addition, uniformly over $w_1, w_2 \in \mathcal{I}$,

$$\begin{aligned}
E_{\mathcal{N}_k}[(T_n(w_1) - T_n(w_2))^2]^{1/2} &\leq \left\|\frac{\ell_\theta(w_1)'\widehat{\Omega}^{1/2}}{\sqrt{n}\widehat{\sigma}_n(w_1)} - \frac{\ell_\theta(w_2)'\widehat{\Omega}^{1/2}}{\sqrt{n}\widehat{\sigma}_n(w_2)}\right\| + \left\|\frac{\ell_\theta(w_1)'\Omega^{1/2}}{\sqrt{n}\sigma_\theta(w_1)} - \frac{\ell_\theta(w_2)'\Omega^{1/2}}{\sqrt{n}\sigma_\theta(w_2)}\right\| \\
&\lesssim_P \left\|\frac{\ell_\theta(w_1)}{\sqrt{n}\widehat{\sigma}_n(w_1)} - \frac{\ell_\theta(w_2)}{\sqrt{n}\widehat{\sigma}_n(w_2)}\right\| + \left\|\frac{\ell_\theta(w_1)}{\sqrt{n}\sigma_\theta(w_1)} - \frac{\ell_\theta(w_2)}{\sqrt{n}\sigma_\theta(w_2)}\right\|.
\end{aligned}$$

Moreover, uniformly over $w_1, w_2 \in \mathcal{I}$,

$$\begin{aligned}
\left\|\frac{\ell_\theta(w_1)}{\sqrt{n}\widehat{\sigma}_n(w_1)} - \frac{\ell_\theta(w_2)}{\sqrt{n}\widehat{\sigma}_n(w_2)}\right\| &\leq \frac{\|\ell_\theta(w_1) - \ell_\theta(w_2)\|}{\sqrt{n}\widehat{\sigma}_n(w_1)} + \frac{\|\ell_\theta(w_2)\|}{\sqrt{n}}\left|\frac{1}{\widehat{\sigma}_n(w_1)} - \frac{1}{\widehat{\sigma}_n(w_2)}\right| \\
&= \frac{\|\ell_\theta(w_1) - \ell_\theta(w_2)\|}{\sqrt{n}\widehat{\sigma}_n(w_1)} + \frac{\|\ell_\theta(w_2)\|}{\sqrt{n}}\frac{|\widehat{\sigma}_n(w_2) - \widehat{\sigma}_n(w_1)|}{\widehat{\sigma}_n(w_1)\widehat{\sigma}_n(w_2)} \\
&\lesssim_P \frac{\|\ell_\theta(w_1) - \ell_\theta(w_2)\|}{\|\ell_\theta(w_1)\|} \lesssim \xi^L_\theta(k,\mathcal{I})\|w_1 - w_2\|
\end{aligned}$$

where the last inequality follows from Condition A.6. A similar argument shows that

$$\left\|\frac{\ell_\theta(w_1)}{\sqrt{n}\sigma_\theta(w_1)} - \frac{\ell_\theta(w_2)}{\sqrt{n}\sigma_\theta(w_2)}\right\| \lesssim_P \xi^L_\theta(k,\mathcal{I})\|w_1 - w_2\|$$

uniformly over $w_1, w_2 \in \mathcal{I}$. Now, (A.67) follows from Dudley's inequality (Dudley (1967)).

Finally, to show (A.71), we note that in view of (A.67), it suffices to prove that

$$c_n^0(1-\alpha) \lesssim \sqrt{\log n}. \tag{A.72}$$

But $\{t_n^*(w),\ w \in \mathcal{I}\}$ is a zero mean Gaussian process satisfying $E[t_n^*(w)^2]^{1/2} = 1$ for all $w \in \mathcal{I}$ and

$$E[(t_n^*(w_1) - t_n^*(w_2))^2]^{1/2} \leq \left\| \frac{\ell_\theta(w_1)'\Omega^{1/2}}{\sqrt{n}\sigma_\theta(w_1)} - \frac{\ell_\theta(w_2)'\Omega^{1/2}}{\sqrt{n}\sigma_\theta(w_2)} \right\| \lesssim \xi_\theta^L(k,\mathcal{I})\|w_1 - w_2\|$$

where the last inequality was shown above. Hence, (A.72) follows from combining Dudley's and Markov's inequalities. □

A.8. **Proofs of Section 6.**

Proof of Lemma 6.2. Using the Symmetrization Lemma (Lemma 2.3.6 van der Vaart and Wellner (1996)) and the Khinchin inequality, bound

$$\Delta := E\|\widehat{Q} - Q\| \leq 2EE_\varepsilon\|\mathbb{E}_n[\varepsilon_i Q_i]\| \leq \sqrt{\frac{\log n}{n}} E\|(\mathbb{E}_n Q_i^2)^{1/2}\|$$

Since

$$E\|(\mathbb{E}_n Q_i^2)^{1/2}\| = E\|(\mathbb{E}_n Q_i^2)\|^{1/2} \leq \Big[ME\|\mathbb{E}_n Q_i\|\Big]^{1/2}$$

and

$$\|\mathbb{E}_n Q_i\| \leq \Delta + \|Q\|,$$

one has

$$\Delta \leq \sqrt{\frac{M\log n}{n}}[\Delta + \|Q\|]^{1/2},$$

solving which for Δ gives the result stated in the lemma. □

Proof of Proposition 6.1. For a $\tau > 0$ specified later, define $\epsilon_i^- := \epsilon_i I(|\epsilon_i| \leq \tau) - E[\epsilon_i I(|\epsilon_i| \leq \tau)|X_i]$ and $\epsilon_i^+ := \epsilon_i I(|\epsilon_i| > \tau) - E[\epsilon_i I(|\epsilon_i| > \tau)|X_i]$. Since $E[\epsilon_i|X_i] = 0$, $\epsilon_i = \epsilon_i^- + \epsilon_i^+$. Invoke the decomposition

$$\sum_{i=1}^n \epsilon_i f(X_i) = \sum_{i=1}^n \epsilon_i^- f(X_i) + \sum_{i=1}^n \epsilon_i^+ f(X_i).$$

We apply Theorem 6.1 to the first term. Noting that $\mathrm{var}(\epsilon_i^- f(X_i)) \leq \sup_x E[(\epsilon_i^-)^2|X_i = x]E[f(X_i)^2] \leq \sup_x E[\epsilon_i^2|X_i = x] = \sigma^2$ and $\epsilon_i^- f(X_i) \leq 2\tau b$, we have

$$E\left[\left\|\sum_{i=1}^n \epsilon_i^- f(X_i)\right\|_{\mathcal{F}}\right] \leq C\left[\sqrt{n\sigma^2 V\log(Ab)} + V\tau b\log(Ab)\right].$$

On the other hand, applying Theorem 2.14.1 of van der Vaart and Wellner (1996) to the second term, we obtain

$$E\left[\left\|\sum_{i=1}^{n}\epsilon_i^{+}f(X_i)\right\|_{\mathcal{F}}\right]\leq C\sqrt{n}b\sqrt{E[|\epsilon_1^{+}|^2]}\int_0^1\sqrt{V\log(A/\epsilon)}d\epsilon. \tag{A.73}$$

By assumption,

$$E[|\epsilon_1^{+}|^2]\leq E[\epsilon_1^2 I(|\epsilon_1|>\tau)]\leq\tau^{-m+2}E[|\epsilon_1|^m],$$

by which we have

$$(A.73)\leq C\sqrt{E[|\epsilon_1|^m]}b\tau^{-m/2+1}\sqrt{nV\log(A)}.$$

Taking $\tau=b^{2/(m-2)}$, we obtain the desired inequality. □

A.9. **Additional technical results.**

Lemma A.2. *Let Z be a random vector in $\mathbb{R}^k$, M be a $k\times k$ matrix and $\Gamma\subset\mathbb{R}^k\setminus\{0\}$. Then we have that*

$$\sup_{\gamma\in\Gamma}E\left[\left|\frac{\gamma'}{\|\gamma\|}MZ\right|^m\right]\leq\|M\|^m\sup_{\|a\|=1}E\left[|a'Z|^m\right].$$

Proof of Lemma A.2. Let $\bar{\gamma}$ achieve the supremum on the left hand side and set $\bar{a}=\bar{\gamma}/\|\bar{\gamma}\|$. Then we have

$$\begin{array}{rcl} E\left[|\bar{a}'MZ|^m\right] & = & E\left[|(M'\bar{a})'Z|^m\right]=\|M'\bar{a}\|^m E\left[|\frac{(M'\bar{a})'}{\|M'\bar{a}\|}Z|^m\right] \\ & \leq & \|M'\|^m\|\bar{a}\|^k E\left[|\frac{(M'\bar{a})'}{\|M'\bar{a}\|}Z|^m\right] \\ & \leq & \|M\|^m\sup_{\|a\|=1}E\left[|a'Z|^m\right] \end{array}$$

since $\|\bar{a}\|=1$ and $M'\bar{a}/\|M'\bar{a}\|=1$. □

Lemma A.3 (Closeness in Probability Implies Closeness of Conditional Quantiles)**.** *Let X_n and Y_n be random variables and $\mathcal{D}_n$ be a random vector. Let $F_{X_n}(x|\mathcal{D}_n)$ and $F_{Y_n}(x|\mathcal{D}_n)$ denote the conditional distribution functions, and $F_{X_n}^{-1}(p|\mathcal{D}_n)$ and $F_{Y_n}^{-1}(p|\mathcal{D}_n)$ denote the corresponding conditional quantile functions. If $|X_n-Y_n|=o_P(\varepsilon)$, then for some $\nu_n\searrow 0$ with probability converging to one*

$$F_{X_n}^{-1}(p|\mathcal{D}_n)\leq F_{Y_n}^{-1}(p+\nu_n|\mathcal{D}_n)+\varepsilon \text{ and } F_{Y_n}^{-1}(p|\mathcal{D}_n)\leq F_{X_n}^{-1}(p+\nu_n|\mathcal{D}_n)+\varepsilon,\forall p\in(\nu_n,1-\nu_n).$$

Proof of Lemma A.3. We have that for some $\nu_n\searrow 0$, $P\{|X_n-Y_n|>\varepsilon\}=o(\nu_n)$. This implies that $P[P\{|X_n-Y_n|>\varepsilon|\mathcal{D}_n\}\leq\nu_n]\to 1$, i.e. there is a set Ω_n such that $P(\Omega_n)\to 1$ and $P\{|X_n-Y_n|>\varepsilon|\mathcal{D}_n\}\leq\nu_n$ for all $\mathcal{D}_n\in\Omega_n$. So, for all $\mathcal{D}_n\in\Omega_n$

$$F_{X_n}(x|\mathcal{D}_n)\geq F_{Y_n+\varepsilon}(x|\mathcal{D}_n)-\nu_n \text{ and } F_{Y_n}(x|\mathcal{D}_n)\geq F_{X_n+\varepsilon}(x|\mathcal{D}_n)-\nu_n,\forall x\in\mathbb{R},$$

which implies the inequality stated in the lemma, by definition of the conditional quantile function and equivariance of quantiles to location shifts. □

References

Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric models. *Econometrica* **59** 307-345.

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* **74** 539-563.

Belloni, A., Chernozhukov, V. and Fernández-Val, I. (2011). Conditional quantile processes based on series or many regressors. *working paper.*

Burman, P., Chen, K.W. (1989). Nonparametric estimation of a regression function. *Annals of Statistics* **17** 1567-1596.

Cattaneo, M. and Farrell, M. (2013). Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics* **174** 127-143.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In: Heckman, J.J., Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6B. Elsevier (Chapter 76).

Chen, X. and Christensen, T. (2013). Optimal uniform convergence rates for sieve nonparametric instrumental variables regression. *Cemmap working paper* **56**.

Chernozhukov, C., Chetverikov, D., and Kato, K. (2012). Gaussian approximation of suprema of empirical processes. *arXiv:1212.6906.*

Chernozhukov, C., Chetverikov, D., and Kato, K. (2012). Anti-concentration and adaptive honest confidence bands. *arXiv:1303.7152.*

Chernozhukov, C., Fernández-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78** 1093-1125.

Chernozhukov, V., Lee, S. and Rosen, A. (2009). Intersection bounds: estimation and inference. arXiv:0907.3503.

Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal* **1** 54-81.

De Boor, C. (2001). *A practical guide to splines (Revised Edition).* Springer.

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

Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, **1**, 290330.

Eastwood, B.J., Gallant, A.R. (1991). Adaptive rules for seminonparametric estimation that achieve asymptotic normality. *Econometric Theory* **7** 307-340.

Gallant, A.R., Souza, G. (1991). On the asymptotic normality of Fourier flexible functional form estimates. *Journal of Econometrics* **50** 329-353.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143-1216.

Guédon O. and Rudelson, M. (2007). L_p-moments of random vectors via majorizing measures. *Advances in Mathematics* **208** 798-823.

Hardle, W. (1990). *Applied nonparametric regression.* Cambridge University Press.

Horn, R. and Johnson, C. (1990). *Matrix Analysis.* Cambridge University Press.

Horowitz, J.L. (2009). *Semiparametric and Nonparametric Methods in Econometrics.* Springer.

Huang, J.Z. (2003a). Asymptotics for polynomial spline regression under weak conditions. *Statist. Probab. Lett.* **65** 207-216.

Huang, J.Z. (2003b). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600-1635.

Lust-Picard L. and Pisier, G. (1991). Non-commutative Khintchine and Paley inequalities. *Arkiv för Matematik* **29** 241-260.

Mallat, S. (2009). *A Wavelet Tour of Signal Processing.* Third Edition. Academic Press.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863-884.

Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147-168.

Newey, W., Powell, J., and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica* **67** 565-603.

Pollard, D. (2002). A User's Guide to Measure Theoretic Probability. Cambridge Series in Statistics and Probabilistic Mathemathics.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis* **164**, 1, 60-72.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.

Stone, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author. *Ann. Statist.* **22** 118-184.

Talagrand, M. (1996a). Majorizing measures: the generic chaining. *Ann. Probab.* **24** 1049–1103.

Talagrand, M. (1996b). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563.

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

Wasserman, L. (2006). *All of nonparametric statistics.* Springer.

van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer.

Zygmund, A. (2002). *Trigonometric Series.* Cambridge Mathematical Library.