The Annals of Statistics
2020, Vol. 48, No. 3, 1718–1741
https://doi.org/10.1214/19-AOS1865

LARGE SAMPLE PROPERTIES OF PARTITIONING-BASED SERIES ESTIMATORS

BY MATIAS D. CATTANEO[1], MAX H. FARRELL[2] AND YINGJIE FENG[3]

[1]*Department of Operations Research and Financial Engineering, Princeton University, cattaneo@princeton.edu*
[2]*Booth School of Business, University of Chicago, max.farrell@chicagobooth.edu*
[3]*Department of Politics, Princeton University, yingjief@princeton.edu*

We present large sample results for partitioning-based least squares nonparametric regression, a popular method for approximating conditional expectation functions in statistics, econometrics and machine learning. First, we obtain a general characterization of their leading asymptotic bias. Second, we establish integrated mean squared error approximations for the point estimator and propose feasible tuning parameter selection. Third, we develop pointwise inference methods based on undersmoothing and robust bias correction. Fourth, employing different coupling approaches, we develop uniform distributional approximations for the undersmoothed and robust bias-corrected t-statistic processes and construct valid confidence bands. In the univariate case, our uniform distributional approximations require seemingly minimal rate restrictions and improve on approximation rates known in the literature. Finally, we apply our general results to three partitioning-based estimators: splines, wavelets and piecewise polynomials. The Supplemental Appendix includes several other general and example-specific technical and methodological results. A companion R package is provided.

1. Introduction. We study the standard nonparametric regression setup, where $\{(y_i, \mathbf{x}_i'), i = 1, \dots n\}$ is a random sample from the model

$$y_i = \mu(\mathbf{x}_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i | \mathbf{x}_i] = 0, \qquad \mathbb{E}[\varepsilon_i^2 | \mathbf{x}_i] = \sigma^2(\mathbf{x}_i), \tag{1.1}$$

for a scalar response y_i and a d-vector of continuously distributed covariates $\mathbf{x}_i = (x_{1,i}, \dots, x_{d,i})'$ with compact support $\mathcal{X}$. The object of interest is the unknown regression function $\mu(\cdot)$ and its derivatives. We focus on *partitioning-based*, or locally-supported, series (linear sieve) least squares regression estimators, which are characterized by two features. First, the support $\mathcal{X}$ is partitioned into nonoverlapping cells, which are then used to form a set of basis functions. Second, the final fit is determined by a least squares regression using these bases. The key distinguishing characteristic is that each basis function is nonzero only on a small, contiguous set of cells of the partition. Popular examples include splines, compactly supported wavelets and piecewise polynomials. For this class of estimators, we develop novel bias approximations, integrated mean squared error (IMSE) expansions useful for tuning parameter selection and pointwise and uniform estimation and inference results, with and without bias correction techniques.

A partitioning-based estimator is made precise by the partition of $\mathcal{X}$ and basis expansion used. Let $\Delta = \{\delta_l \subset \mathcal{X} : 1 \le l \le \bar{\kappa}\}$ be a collection of $\bar{\kappa}$ open and disjoint sets, the closure of whose union is $\mathcal{X}$ (or, more generally, covers $\mathcal{X}$). We restrict δ_l to be polyhedral, which allows for tensor products of (marginally-formed) intervals as well as other popular partitioning shapes. Based on this partition, the dictionary of K basis functions,

Received November 2018; revised May 2019.
MSC2010 subject classifications. Primary 62E20, 62G08, 62G20; secondary 62M99.
Key words and phrases. Nonparametric regression, series methods, sieve methods, robust bias correction, uniform inference, strong approximation, tuning parameter selection.

each of order m (e.g., $m = 4$ for cubic splines) is denoted by $\mathbf{x}_i \mapsto \mathbf{p}(\mathbf{x}_i) := \mathbf{p}(\mathbf{x}_i; \Delta, m) = (p_1(\mathbf{x}_i; \Delta, m), \ldots, p_K(\mathbf{x}_i; \Delta, m))'$. For $\mathbf{x} \in \mathcal{X}$ and $\mathbf{q} = (q_1, \ldots, q_d)' \in \mathbb{Z}_+^d$, the partial derivative $\partial^{\mathbf{q}}\mu(\mathbf{x})$ is estimated by least squares regression

$$\widehat{\partial^{\mathbf{q}}\mu}(\mathbf{x}) = \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\widehat{\boldsymbol{\beta}}, \qquad \widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^K} \sum_{i=1}^{n} (y_i - \mathbf{p}(\mathbf{x}_i)'\mathbf{b})^2, \tag{1.2}$$

where $\partial^{\mathbf{q}}\mu(\mathbf{x}) = \partial^{q_1+\cdots+q_d}\mu(\mathbf{x})/\partial^{q_1}x_1 \cdots \partial^{q_d}x_d$ (for boundary points defined from the interior of $\mathcal{X}$ as usual) and $\mu(\mathbf{x}) := \partial^{\mathbf{0}}\mu(\mathbf{x})$.

The approximation power of this class of estimators comes from two user-specified parameters: the granularity of the partition Δ and the order $m \in \mathbb{Z}_+$ of the basis. The choice m is often fixed in practice, and hence we regard Δ as the tuning parameter. Under our assumptions, $\bar{\kappa} \to \infty$ as the sample size $n \to \infty$, and the volume of each δ_l shrinks proportionally to h^d, where $h = \max\{\operatorname{diam}(\delta) : \delta \in \Delta\}$ serves as a universal measure of the granularity. Thus, as $\bar{\kappa} \to \infty$, h^d vanishes at the same rate, and with each basis being supported only on a finite number of cells, K diverges proportionally as well. Complete, detailed examples of bases and partitioning schemes are discussed in the online supplement for brevity.

Our first contribution, in Section 3, is a general characterization of the bias of partitioning-based estimators, which we then use for both tuning parameter selection and robust bias correction. In the supplement, we specialize our generic bias approximation to splines, wavelets and piecewise polynomials over different partitioning schemes, leading to novel results.

Our second contribution, in Section 4, is a general integrated mean squared error (IMSE) expansion for partitioning-based estimators. These results lead to IMSE-optimal partitioning choices, and hence deliver IMSE-optimal point estimators of the regression function and its derivatives. We show that the IMSE-optimal choice of partition granularity obeys $h_{\mathtt{IMSE}} \asymp n^{-1/(2m+d)}$, which translates to the familiar $K_{\mathtt{IMSE}} \asymp n^{-d/(2m+d)}$, and give a precise characterization of the leading constant. For simple cases on tensor-product partitions, some results exist for splines [1, 43, 44] and piecewise polynomials [13]. In addition to generalizing these results substantially, our characterization for compactly supported wavelets appears to be new.

The IMSE-optimal partitioning scheme, and consistent implementations thereof, cannot be used directly to form valid pointwise or uniform (in $\mathbf{x} \in \mathcal{X}$) inference procedures. Undersmoothing (employing a finer partition than would be IMSE-optimal) is theoretically valid for inference, but difficult to implement in a principled way. Inspired by results proving that undersmoothing is never optimal relative to bias correction for kernel-based nonparametrics [6], we develop three robust bias-corrected inference procedures using our new bias characterizations of partitioning-based estimators. These methods are more involved than their kernel-based counterparts, but are still based on least squares regression using partitioning-based estimation. Specifically, we show that the conventional partitioning-based estimator $\widehat{\partial^{\mathbf{q}}\mu}(\mathbf{x})$ and the three bias-corrected estimators we propose have a common structure, which we exploit to obtain general pointwise and uniform distributional approximations under weak (sometimes minimal) conditions. These robust bias correction results for partitioning-based estimators, both pointwise and uniform in $\mathbf{x}$, are practically useful because they allow for mean squared error minimizing tuning parameter choices, thus offering a data-driven method combining optimal point estimation and valid inference on the same partitioning scheme.

Section 5 establishes pointwise in $\mathbf{x} \in \mathcal{X}$ distributional approximations for both conventional and robust bias-corrected t-statistics based on partitioning-based estimators. These pointwise distributional results are made uniform in Section 6, where we establish a strong approximation for the whole t-statistic processes, indexed by the point $\mathbf{x} \in \mathcal{X}$, covering both conventional and robust bias-corrected inference. To illustrate, Section 6.3 constructs valid confidence bands for (derivatives of) the regression function using our uniform distributional

approximations. When compared to the current literature, we obtain a strong approximation to the *entire* t-statistic process under either weaker or seemingly minimal conditions on the tuning parameter h (i.e., on K or $\bar{\kappa}$), depending on the case under consideration.

Section 7 summarizes Monte Carlo results, Section 8 gives selected proofs, and Section 9 concludes. A Supplemental Appendix (SA hereafter) [15] gives complete proofs, several new technical and methodological results, further Monte Carlo evidence and applies our general results to splines, wavelets and piecewise polynomials. A companion R package [14] is provided.

1.1. *Related literature.* This paper contributes primarily to two literatures, nonparametric regression and strong approximation. There is a vast literature on nonparametric regression, summarized in many textbook treatments (e.g., [25, 26, 34] and references therein). Of particular relevance are treatments of series (linear sieve) methods, which offer some results concerning partitioning-based estimators in particular, many times limited to splines, wavelets or piecewise polynomials considered separately [2, 3, 13, 16, 17, 27, 28, 32, 43]. Piecewise polynomial fits on partitions have a long and ongoing tradition in statistics, dating at least to the regressogram of Tukey [39], continuing through [37] (named local polynomial regression therein) and [13], and up to modern, data-driven partitioning techniques such as regression trees [5], trend filtering [38] and related methods [42]. Partitioning-based methods have also featured as inputs or preprocessing in treatment effects [9, 12], empirical finance [11], "binscatter" analysis [10] and other settings. The bias corrections we develop for series estimation and uniform inference follow recent work on kernel-based nonparametric inference [6–8]. Our coupling and strong approximation results relate to early work discussed in [24], Chapter 22, and the more recent work in [18–22] and [41], as well as with the results for series estimators in [3] and [2]. See also [40] for a review on strong approximation methods.

1.2. *Notation.* For a d-tuple $\mathbf{q} = (q_1, \ldots, q_d) \in \mathbb{Z}_+^d$, define $[\mathbf{q}] = \sum_{j=1}^d q_j$, $\mathbf{x}^{\mathbf{q}} = x_1^{q_1} x_2^{q_2} \cdots x_d^{q_d}$ and $\partial^{\mathbf{q}}\mu(\mathbf{x}) = \partial^{[\mathbf{q}]}\mu(\mathbf{x})/\partial x_1^{q_1} \ldots \partial x_d^{q_d}$. Unless explicitly stated otherwise, whenever $\mathbf{x}$ is a boundary point of some closed set, the partial derivative is understood as the limit with $\mathbf{x}$ ranging within it. Let $\mathbf{0} = (0, \ldots, 0)'$ be the length-d zero vector. We set $\mu(\mathbf{x}) := \partial^{\mathbf{0}}\mu(\mathbf{x})$ and $\widehat{\mu}_j(\mathbf{x}) := \widehat{\partial^{\mathbf{0}}\mu}_j(\mathbf{x})$ for $j = 0, 1, 2, 3$ and collect the covariates as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$. The tensor product or Kronecker product operator is $\otimes$. The smallest integer greater than or equal to u is $\lceil u \rceil$. For two random variables X and Y, $X =_d Y$ denotes that they have the same probability law.

We use several norms. For a vector $\mathbf{v} = (v_1, \ldots, v_M) \in \mathbb{R}^M$, we write $\|\mathbf{v}\| = (\sum_{i=1}^M v_i^2)^{1/2}$ and $\dim(\mathbf{v}) = M$. For a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, $\|\mathbf{A}\| = \max_i \sigma_i(\mathbf{A})$ and $\|\mathbf{A}\|_\infty = \max_{1 \le i \le M} \sum_{j=1}^N |a_{ij}|$ for operator norms induced by L_2 and L_∞ norms, where $\sigma_i(\mathbf{A})$ is the ith singular value of $\mathbf{A}$, and $\lambda_{\min}(\mathbf{A})$ is the minimum eigenvalue of $\mathbf{A}$ when $M = N$.

We use empirical process notation: $\mathbb{E}_n[g(\mathbf{x}_i)] = \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i)$ and $\mathbb{G}_n[g(\mathbf{x}_i)] = \frac{1}{\sqrt{n}} \times \sum_{i=1}^n (g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i)])$. For sequences, $a_n \lesssim b_n$ denotes $\limsup_n |a_n/b_n|$ is finite, $a_n = O_{\mathbb{P}}(b_n)$ denotes $\limsup_{\epsilon \to \infty} \limsup_n \mathbb{P}[|a_n/b_n| \ge \epsilon] = 0$, $a_n = o(b_n)$ denotes $a_n/b_n \to 0$, $a_n = o_{\mathbb{P}}(b_n)$ denotes $a_n/b_n \to_{\mathbb{P}} 0$, where $\to_{\mathbb{P}}$ is convergence in probability, and $a_n \asymp b_n$ denotes $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Limits are taken as $n \to \infty$ ($h \to 0$, $K \to \infty$, as appropriate), unless otherwise stated.

Finally, throughout the paper, $r_n > 0$ denotes a nonvanishing sequence and $\nu > 0$ denotes a fixed constant used to characterize moment bounds.

2. Setup.

We first make precise our setup and assumptions. Our first assumption restricts the data generating process.

ASSUMPTION 1 (Data generating process).

(a) $\{(y_i, \mathbf{x}_i') : 1 \le i \le n\}$ are i.i.d. satisfying (1.1), where $\mathbf{x}_i$ has compact connected support $\mathcal{X} \subset \mathbb{R}^d$ and an absolutely continuous distribution function. The density of $\mathbf{x}_i$, $f(\cdot)$, and the conditional variance of y_i given $\mathbf{x}_i$, $\sigma^2(\cdot)$, are bounded away from zero and continuous.

(b) $\mu(\cdot)$ is S-times continuously differentiable, for $S \ge [\mathbf{q}]$, and all $\partial^{\boldsymbol{\varsigma}}\mu(\cdot)$, $[\boldsymbol{\varsigma}] = S$, are Hölder continuous with exponent $\varrho > 0$.

The next two assumptions specify a set of high-level conditions on the partition and basis: we require that the partition is "quasi-uniform" and the basis is "locally" supported.

ASSUMPTION 2 (Quasi-uniform partition). The ratio of the sizes of inscribed and circumscribed balls of each $\delta \in \Delta$ is bounded away from zero uniformly in $\delta \in \Delta$, and

$$\frac{\max\{\operatorname{diam}(\delta) : \delta \in \Delta\}}{\min\{\operatorname{diam}(\delta) : \delta \in \Delta\}} \lesssim 1,$$

where $\operatorname{diam}(\delta)$ denotes the diameter of δ. Further, for $h = \max\{\operatorname{diam}(\delta) : \delta \in \Delta\}$, assume $h = o(1)$.

This condition implies that the size of each $\delta \in \Delta$ can be well characterized by the diameter of δ, where we use h as a universal measure of mesh sizes of elements in Δ. In the univariate case, it reduces to a bounded mesh ratio. A special case of a quasi-uniform partition is one formed via a tensor product of univariate marginal partitions on each dimension of $\mathbf{x} \in \mathcal{X}$, with appropriately chosen knot positions. The SA (Section SA-3) gives details and discusses this special example of partitioning scheme. If Δ covers only a strict subset of $\mathcal{X}$, then our results hold on that subset.

We focus on nonrandom partitions. Data-dependent partitioning can be accommodated by sample splitting: estimating the partition configuration in one subsample and performing inference in the other. In this way, quite general partitions can be used with our results, including data-driven methods such as regression trees and other modern machine learning techniques. In fact, these modern methods would typically generate nontensor-product partitioning schemes. We defer studying general data-dependent partitioning for future work, but note that a few specific results are available [5, 9, 33].

The second assumption on the partitioning-based estimators employs generalized notions of *stable local basis* [23] and *active basis* [28]. We say a function $p(\cdot)$ on $\mathcal{X}$ is *active* on $\delta \in \Delta$ if it is not identically zero on δ.

ASSUMPTION 3 (Local basis).

(a) For each basis function p_k, $k = 1, \dots, K$, the union of elements of Δ on which p_k is active is a connected set, denoted by $\mathcal{H}_k$. For all $k = 1, \dots, K$, both the number of elements of $\mathcal{H}_k$ and the number of basis functions which are active on $\mathcal{H}_k$ are bounded by a constant.

(b) For any $\mathbf{a} = (a_1, \dots, a_K)' \in \mathbb{R}^K$,

$$\mathbf{a}' \int_{\mathcal{H}_k} \mathbf{p}(\mathbf{x}; \Delta, m)\mathbf{p}(\mathbf{x}; \Delta, m)'\, d\mathbf{x}\mathbf{a} \gtrsim a_k^2 h^d, \quad k = 1, \dots, K.$$

(c) Let $[\mathbf{q}] < m$. For an integer $\varsigma \in [[\mathbf{q}], m)$, for all $\boldsymbol{\varsigma}$, $[\boldsymbol{\varsigma}] \le \varsigma$,

$$h^{-[\boldsymbol{\varsigma}]} \lesssim \inf_{\delta\in\Delta} \inf_{\mathbf{x}\in \operatorname{clo}(\delta)} \left\|\partial^{\boldsymbol{\varsigma}}\mathbf{p}(\mathbf{x}; \Delta, m)\right\| \le \sup_{\delta\in\Delta} \sup_{\mathbf{x}\in \operatorname{clo}(\delta)} \left\|\partial^{\boldsymbol{\varsigma}}\mathbf{p}(\mathbf{x}; \Delta, m)\right\| \lesssim h^{-[\boldsymbol{\varsigma}]},$$

where $\operatorname{clo}(\delta)$ is the closure of δ, and for $[\boldsymbol{\varsigma}] = \varsigma + 1$,

$$\sup_{\delta\in\Delta} \sup_{\mathbf{x}\in \operatorname{clo}(\delta)} \left\|\partial^{\boldsymbol{\varsigma}}\mathbf{p}(\mathbf{x}; \Delta, m)\right\| \lesssim h^{-\varsigma-1}.$$

Assumption 3 imposes conditions ensuring the stability of the L_2 projection operator onto the approximating space. Condition 3(a) requires that each basis function in $\mathbf{p}(\mathbf{x}; \Delta, m)$ be supported by a region consisting of a finite number of cells in Δ. Therefore, as $\bar{\kappa} \to \infty$ (and $h \to 0$), each element of Δ shrinks and all the basis functions are "locally supported" relative to the whole support of the data. Another common assumption in least squares regression is that the regressors are not too colinear: the minimum eigenvalue of $\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']$ is usually assumed to be bounded away from zero. Since the local support condition in Assumption 3(a) implies a banded structure for this matrix, it suffices to require that the basis functions are not too colinear locally, as stated in Assumption 3(b). These two assumptions are very similar to Condition A.2 and Condition A.3 in the Appendix of [28] and, therefore, they could also be used to establish theoretical results analogous to those discussed in that Appendix (such results are not needed herein because our proofs are different). Finally, Assumption 3(c) controls the magnitude of the local basis in a uniform sense.

Assumptions 2 and 3 implicitly relate the number of approximating series terms, the number of knots used and the maximum mesh size: $K \asymp \bar{\kappa} \asymp h^{-d}$. By restricting the growth rate of these tuning parameters, the least squares partitioning-based estimator satisfying the above conditions is well-defined in large samples. We next state a high-level requirement that gives explicit expression of the leading approximation error. For each $\mathbf{x} \in \mathcal{X}$, let $\delta_{\mathbf{x}}$ be the element of Δ whose closure contains $\mathbf{x}$ and $h_{\mathbf{x}}$ be the diameter of this $\delta_{\mathbf{x}}$.

ASSUMPTION 4 (Approximation error). Let $S \geq m$. For all $\boldsymbol{\varsigma}$ satisfying $[\boldsymbol{\varsigma}] \leq \varsigma$ in Assumption 3, there exists $s^* \in \mathcal{S}_{\Delta,m}$, the linear span of $\mathbf{p}(\mathbf{x}; \Delta, m)$, and

$$\mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x}) = -\sum_{\boldsymbol{u} \in \Lambda_m} \partial^{\boldsymbol{u}} \mu(\mathbf{x}) h_{\mathbf{x}}^{m-[\boldsymbol{\varsigma}]} B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x})$$

such that

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| \partial^{\boldsymbol{\varsigma}} \mu(\mathbf{x}) - \partial^{\boldsymbol{\varsigma}} s^*(\mathbf{x}) + \mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x}) \right| \lesssim h^{m+\varrho-[\boldsymbol{\varsigma}]} \tag{2.1}$$

and

$$\sup_{\delta \in \Delta} \sup_{\mathbf{x}_1, \mathbf{x}_2 \in \operatorname{clo}(\delta)} \frac{|B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}_1) - B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}_2)|}{\|\mathbf{x}_1 - \mathbf{x}_2\|} \lesssim h^{-1} \tag{2.2}$$

where $B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\cdot)$ is a known function that is bounded uniformly over n, and Λ_m is a multiindex set, which depends on the basis, with $[\boldsymbol{u}] = m$ for $\boldsymbol{u} \in \Lambda_m$.

The usual rate-only assumption, $\sup_{\mathbf{x} \in \mathcal{X}} |\partial^{\mathbf{q}} \mu(\mathbf{x}) - \partial^{\mathbf{q}} s^*(\mathbf{x})| \lesssim h^{m-[\mathbf{q}]}$, which is implied by Assumptions 4, will not suffice for our bias correction and IMSE expansion results: (2.1) is needed. The terms $B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x})$ in $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x})$ are known functions of $\mathbf{x}$ which depend on the particular partitioning scheme and bases used. The only unknowns are the higher-order derivatives of $\mu(\cdot)$. In the SA (Section SA-6) we verify this (and other assumptions) for splines, wavelets and piecewise polynomials, including explicit formulas for the leading error in (2.1) and precise characterizations of Λ_m. We assume enough smoothness exists to characterize these terms: see [6] for a discussion when smoothness constrains inference.

The function $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\cdot)$ is understood as the approximation error in L_∞ norm, and is not in general the misspecification (or smoothing) bias of a series estimator. In least squares series regression settings, the leading smoothing bias is described by two terms in general: $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\cdot)$ and the accompanying error from the linear projection of $\mathscr{B}_{m,\mathbf{0}}(\cdot)$ onto $\mathcal{S}_{\Delta,m}$. We formalize this result in Lemma 3.1 below. The second bias term is often ignored because in several cases

the leading approximation error $\mathscr{B}_{m,\mathbf{0}}(\cdot)$ is *approximately orthogonal* to $\mathbf{p}(\cdot)$ with respect to the Lebesgue measure, that is,

$$\max_{1\le k\le K}\int_{\mathcal{H}_k} p_k(\mathbf{x};\Delta,m)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x})\,d\mathbf{x} = o\big(h^{m+d}\big), \tag{2.3}$$

under Assumptions 1–4. In some simple cases, (2.3) is automatically satisfied if one constructs the leading error based on a basis representing the orthogonal complement of $\mathcal{S}_{\Delta,m}$. When (2.3) holds, the leading term in L_∞ approximation error coincides with the leading misspecification (or smoothing) bias of a partitioning-based series estimator. When a stronger quasi-uniformity condition holds (i.e., neighboring cells are of the same size asymptotically), a sufficient condition for (2.3) is simply the orthogonality between $B_{\boldsymbol{u},\mathbf{0}}(\cdot)$ and $\mathbf{p}(\cdot)$ in L_2 with respect to the Lebesgue measure, for all $\boldsymbol{u}\in\Lambda_m$.

For general partitioning-based estimators, this orthogonality need not hold. For example, (2.3) is hard to verify when the partitioning employed is sufficiently uneven, as is usually the case when employing machine learning methods. All our main results hold when this orthogonality fails, and importantly, our bias correction methods and IMSE expansion explicitly account for the L_2 projection of $\mathscr{B}_{m,\mathbf{0}}(\cdot)$ onto the approximating space spanned by $\mathbf{p}(\cdot)$.

3. Characterization and correction of bias. We now precisely characterize the bias of $\widehat{\partial^{\mathbf{q}}\mu}(\mathbf{x})$ under Assumptions 1–4, but not assuming (2.3). Then, using this result, we develop valid IMSE expansions and three robust bias-corrected inference procedures. This section focuses on bias correction, and Section 5 presents the associated robust Studentization adjustments for inference, following the ideas in [6] for kernel-based nonparametrics.

Given our assumptions, the estimator $\widehat{\partial^{\mathbf{q}}\mu}(\mathbf{x})$ of (1.2) can be written as

$$\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) := \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_0(\mathbf{x}_i)y_i\big], \tag{3.1}$$

where

$$\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})' := \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\mathbb{E}_n\big[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)'\big]^{-1} \quad\text{and}\quad \boldsymbol{\Pi}_0(\mathbf{x}_i) := \mathbf{p}(\mathbf{x}_i).$$

The subscript "0" differentiates this estimator from the bias-corrected versions below. We give our first result, proven in the SA, Section SA-10.2.

LEMMA 3.1 (Conditional bias). *Let Assumptions 1, 2, 3 and 4 hold. If* $\frac{\log n}{nh^d} = o(1)$, *then*

$$\begin{aligned}
&\mathbb{E}\big[\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})|\mathbf{X}\big] - \partial^{\mathbf{q}}\mu(\mathbf{x}) \\
&\quad= \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_0(\mathbf{x}_i)\mu(\mathbf{x}_i)\big] - \partial^{\mathbf{q}}\mu(\mathbf{x}) && (3.2)\\
&\quad= \mathscr{B}_{m,\mathbf{q}}(\mathbf{x}) - \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_0(\mathbf{x}_i)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)\big] + O_{\mathbb{P}}\big(h^{m+\varrho-[\mathbf{q}]}\big). && (3.3)
\end{aligned}$$

The proof of this lemma generalizes an idea in [43], Theorem 2.2, to handle partitioning-based series estimators beyond the specific example of B-Splines on tensor-product partitions. The first component $\mathscr{B}_{m,\mathbf{q}}(\mathbf{x})$ is the leading term in the asymptotic error expansion and depends on the function space generated by the series employed. The second component comes from the least squares regression, and it can be interpreted as the projection of the leading approximation error onto the space spanned by the basis employed. Because the approximating basis $\mathbf{p}(\mathbf{x})$ is locally supported (Assumption 3), the orthogonality condition in (2.3), when it holds, suffices to guarantee that the projection of the leading error is of smaller order (such as for B-splines on a tensor-product partition). In general, the bias will be $O(h^{m-[\mathbf{q}]})$ and further, in finite samples both terms may be important even if (2.3) holds.

We consider three bias correction methods to remove the leading bias terms of Lemma 3.1. All three methods rely, in one way or another, on a higher-order basis: for some $\tilde{m} > m$, let

$\tilde{\mathbf{p}}(\mathbf{x}) := \tilde{\mathbf{p}}(\mathbf{x}; \tilde{\Delta}, \tilde{m})$ be a basis of order $\tilde{m}$ defined on a partition $\tilde{\Delta}$ which has maximum mesh $\tilde{h}$. Objects accented with a tilde always pertain to this secondary basis and partition for bias correction. In practice, a simple choice is $\tilde{m} = m + 1$ and $\tilde{\Delta} = \Delta$.

The first approach is to use a higher-order basis in place of the original basis. This is thus named *higher-order-basis bias correction* and numbered as approach $j = 1$. In complete parallel to (3.1) define

$$\widehat{\partial^{\mathbf{q}}\mu}_1(\mathbf{x}) := \widehat{\boldsymbol{\gamma}}_{\mathbf{q},1}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_1(\mathbf{x}_i)y_i\big], \tag{3.4}$$

where

$$\widehat{\boldsymbol{\gamma}}_{\mathbf{q},1}(\mathbf{x})' := \partial^{\mathbf{q}}\tilde{\mathbf{p}}(\mathbf{x})'\mathbb{E}_n\big[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)'\big]^{-1} \quad \text{and} \quad \boldsymbol{\Pi}_1(\mathbf{x}_i) := \tilde{\mathbf{p}}(\mathbf{x}_i).$$

This approach can be viewed as a bias correction of the original point estimator because, trivially, $\widehat{\partial^{\mathbf{q}}\mu}_1(\mathbf{x}) = \widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) - (\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) - \widehat{\partial^{\mathbf{q}}\mu}_1(\mathbf{x}))$. Valid inference based on $\widehat{\partial^{\mathbf{q}}\mu}_1(\mathbf{x})$ can be viewed as "undersmoothing" applied to the higher-order point estimator, but is distinct from undersmoothing $\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})$ (i.e., using a finer partition Δ and keeping the order fixed). Huang [28] used this idea to remove the asymptotic bias of splines estimators.

Our second approach makes use of the generic expression of the least squares bias in (3.2). The unknown objects in this expression are μ and $\partial^{\mathbf{q}}\mu$, both of which can be estimated using the higher-order estimator (3.4). By plugging these into (3.2) and subtracting the result from $\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})$, we obtain the *least squares bias correction*, numbered as approach 2:

$$\begin{aligned}\widehat{\partial^{\mathbf{q}}\mu}_2(\mathbf{x}) &:= \widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) - \big(\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_0(\mathbf{x}_i)\widehat{\mu}_1(\mathbf{x}_i)\big] - \widehat{\partial^{\mathbf{q}}\mu}_1(\mathbf{x})\big)\\ &:= \widehat{\boldsymbol{\gamma}}_{\mathbf{q},2}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_2(\mathbf{x}_i)y_i\big],\end{aligned} \tag{3.5}$$

where

$$\begin{aligned}\widehat{\boldsymbol{\gamma}}_{\mathbf{q},2}(\mathbf{x})' &:= \big(\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})', -\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\big[\mathbf{p}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)'\big]\mathbb{E}_n\big[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)'\big]^{-1} + \widehat{\boldsymbol{\gamma}}_{\mathbf{q},1}(\mathbf{x})'\big)\\ &\quad \text{and} \quad \boldsymbol{\Pi}_2(\mathbf{x}_i) := \big(\mathbf{p}(\mathbf{x}_i)', \tilde{\mathbf{p}}(\mathbf{x}_i)'\big)',\end{aligned}$$

which is exactly of the same form as $\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})$ and $\widehat{\partial^{\mathbf{q}}\mu}_1(\mathbf{x})$ (cf., (3.1) and (3.4)), except for the change in $\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})$ and $\boldsymbol{\Pi}_j(\mathbf{x}_i)$.

Finally, approach number 3 targets the leading terms in equation (3.3). We dub this approach *plug-in bias correction*, as it specifically estimates the leading bias terms, in fixed-n form, of $\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})$ according to Assumption 4. To be precise, we employ the explicit plug-in bias estimator

$$\widehat{\mathscr{B}}_{m,\mathbf{q}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m} \big(\widehat{\partial^{\boldsymbol{u}}\mu}_1(\mathbf{x})\big)h_{\mathbf{x}}^{m-[\mathbf{q}]}B_{\boldsymbol{u},\mathbf{q}}(\mathbf{x}),$$

with $[\mathbf{q}] < m$ and Λ_m as in Assumption 4, leading to

$$\begin{aligned}\widehat{\partial^{\mathbf{q}}\mu}_3(\mathbf{x}) &:= \widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) - \big(\widehat{\mathscr{B}}_{m,\mathbf{q}}(\mathbf{x}) - \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_0(\mathbf{x}_i)\widehat{\mathscr{B}}_{m,\mathbf{0}}(\mathbf{x}_i)\big]\big)\\ &:= \widehat{\boldsymbol{\gamma}}_{\mathbf{q},3}(\mathbf{x})'\mathbb{E}_n\big[\boldsymbol{\Pi}_3(\mathbf{x}_i)y_i\big]\end{aligned} \tag{3.6}$$

where

$$\begin{aligned}\widehat{\boldsymbol{\gamma}}_{\mathbf{q},3}(\mathbf{x})' &:= \Bigg(\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})', \sum_{\boldsymbol{u}\in\Lambda_m}\big\{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{u},1}(\mathbf{x})'h_{\mathbf{x}}^{m-[\mathbf{q}]}B_{\boldsymbol{u},\mathbf{q}}(\mathbf{x})\\ &\qquad - \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\big[\mathbf{p}(\mathbf{x}_i)h_{\mathbf{x}_i}^{m}B_{\boldsymbol{u},\mathbf{0}}(\mathbf{x}_i)\widehat{\boldsymbol{\gamma}}_{\boldsymbol{u},1}(\mathbf{x}_i)'\big]\big\}\Bigg),\\ &\quad \text{and} \quad \boldsymbol{\Pi}_3(\mathbf{x}_i) := \big(\mathbf{p}(\mathbf{x}_i)', \tilde{\mathbf{p}}(\mathbf{x}_i)'\big)'.\end{aligned}$$

When the orthogonality condition (2.3) holds, the second correction term in $\widehat{\partial^{\mathbf{q}}\mu}_3(\mathbf{x})$ is asymptotically negligible relative to the first. However, in finite samples both terms can be important, so we consider the general case.

Our results employing bias correction will require the following conditions on the higher-order basis used for bias estimation.

ASSUMPTION 5 (Bias correction). The partition $\tilde{\Delta}$ satisfies Assumption 2, with maximum mesh $\tilde{h}$, and the basis $\tilde{\mathbf{p}}(\mathbf{x};\tilde{\Delta},\tilde{m})$, $\tilde{m} > m$, satisfies Assumptions 3 and 4 with $\tilde{\varsigma} = \tilde{\varsigma}(\tilde{m}) \geq m$ in place of ς. Let $\rho := h/\tilde{h}$, which obeys $\rho \to \rho_0 \in (0,\infty)$. In addition, for $j = 3$, either (i) $\tilde{\mathbf{p}}(\mathbf{x};\tilde{\Delta},\tilde{m})$ spans a space containing the span of $\mathbf{p}(\mathbf{x};\Delta,m)$, and for all $\boldsymbol{u} \in \Lambda_m$, $\partial^{\boldsymbol{u}}\mathbf{p}(\mathbf{x};\Delta,m) = \mathbf{0}$; or (ii) both $\mathbf{p}(\mathbf{x};\Delta,m)$ and $\tilde{\mathbf{p}}(\mathbf{x};\tilde{\Delta},\tilde{m})$ reproduce polynomials of degree $[\mathbf{q}]$.

In addition to removing the leading bias, Assumption 5 requires that the asymptotic variance of bias-corrected estimators is properly bounded from below in a uniform sense, which is critical for inference. Additional conditions are needed for plug-in bias correction ($j = 3$) due to the more complicated covariance between $\widehat{\partial^{\mathbf{q}}\mu}_0$ and the estimated leading bias. Orthogonality properties due to the projection structure of the least squares bias correction ($j = 2$) removes these "covariance" components in the variance of $\widehat{\partial^{\mathbf{q}}\mu}_2$. The natural choice of $\tilde{\Delta} = \Delta$ and $\tilde{m} = m + 1$ will obey this restriction under intuitive conditions. In the SA, Assumption 5 is verified for splines, wavelets and piecewise polynomials (Section SA-6), and we also compare theoretically the alternative bias correction strategies (Section SA-7.2).

4. IMSE and convergence rates. We establish two main results related to the point estimator $\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})$: a valid IMSE expansion for the estimator, which gives as a by-product an estimate of its L_2 convergence rate, and its uniform convergence rate.

4.1. *IMSE-optimal point estimation.* We first give a general IMSE approximation, which then is specialized for tensor-product partitions. These expansions are used to obtain optimal choices of partition size from a point estimation perspective.

Our first result holds for any partition Δ satisfying Assumption 2.

THEOREM 4.1 (IMSE). *Let Assumptions 1, 2, 3 and 4 hold. If* $\frac{\log n}{nh^d} = o(1)$, *then for a weighting function* $w(\mathbf{x})$ *that is continuous and bounded away from zero on* $\mathcal{X}$,

$$\int_{\mathcal{X}} \mathbb{E}\big[\big(\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x})\big)^2 | \mathbf{X}\big] w(\mathbf{x})\,d\mathbf{x}$$
$$= \frac{1}{n}\big(\mathscr{V}_{\Delta,\mathbf{q}} + o_{\mathbb{P}}\big(h^{-d-2[\mathbf{q}]}\big)\big) + \big(\mathscr{B}_{\Delta,\mathbf{q}} + o_{\mathbb{P}}\big(h^{2m-2[\mathbf{q}]}\big)\big),$$

where

$$\mathscr{V}_{\Delta,\mathbf{q}} = \operatorname{trace}\bigg(\boldsymbol{\Sigma}_0 \int_{\mathcal{X}} \boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})\boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})' w(\mathbf{x})\,d\mathbf{x}\bigg) \asymp h^{-d-2[\mathbf{q}]},$$

$$\mathscr{B}_{\Delta,\mathbf{q}} = \int_{\mathcal{X}} \big(\mathscr{B}_{m,\mathbf{q}}(\mathbf{x}) - \boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}\big[\mathbf{p}(\mathbf{x}_i)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)\big]\big)^2 w(\mathbf{x})\,d\mathbf{x} \lesssim h^{2m-2[\mathbf{q}]},$$

$\boldsymbol{\Sigma}_0 := \mathbb{E}[\boldsymbol{\Pi}_0(\mathbf{x}_i)\boldsymbol{\Pi}_0(\mathbf{x}_i)'\sigma^2(\mathbf{x}_i)]$, *and* $\boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})' := \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']^{-1}$.

This theorem, proven in the SA, Section SA-10.5, shows that the leading term in the integrated (and pointwise) variance of $\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})$ is of order $n^{-1}h^{-d-2[\mathbf{q}]}$. For the bias term, on the other hand, the theorem only establishes an upper bound: to bound the bias component

from below, stronger conditions on the regression function would be needed. This rate bound is sharp in general.

The quantities $\mathscr{V}_{\Delta,\mathbf{q}}$ and $\mathscr{B}_{\Delta,\mathbf{q}}$ are nonrandom sequences depending on the partitioning scheme Δ in a complicated way, and need not converge as $h \to 0$. Nevertheless, when the integrated squared bias does not vanish ($\mathscr{B}_{\Delta,\mathbf{q}} \neq 0$), Theorem 4.1 implies that the IMSE-optimal mesh size $h_{\mathtt{IMSE}}$ is proportional to $n^{-1/(2m+d)}$, or equivalently, the IMSE-optimal number of series terms $K_{\mathtt{IMSE}} \asymp n^{d/(2m+d)}$. Furthermore, because the IMSE expansion is obtained for a given partition scheme, the result in Theorem 4.1 can be used to evaluate different partitioning schemes altogether, and to select the "optimal" one in an IMSE sense.

Theorem 4.1 generalizes prior work substantially. Existing results cover only special cases, such as piecewise polynomials [13] or splines [1, 43, 44] on tensor-product partitions only, and often restricting to $d = 1$ or $[\mathbf{q}] = 0$.

We now consider the special case of a tensor-product partition where the "tuning parameter" Δ reduces to the vector of partitioning knots $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_d)'$, where κ_ℓ is the number of subintervals used for the ℓ-th covariate. We assume that Δ and $\mathbf{p}(\cdot)$ obey the following regularity conditions, so that the limiting constants in the IMSE approximation can be characterized.

ASSUMPTION 6 (Regularity for asymptotic IMSE). Suppose that $\mathcal{X} = \bigotimes_{\ell=1}^{d} \mathcal{X}_\ell \subset \mathbb{R}^d$, which is normalized to $[0, 1]^d$ without loss of generality, and Δ is a tensor-product partition. For $\mathbf{x} \in [0, 1]^d$, denote $\delta_{\mathbf{x}} = \{t_{\ell,l_{\mathbf{x}}} \leq x_\ell \leq t_{\ell,l_{\mathbf{x}}+1}, 1 \leq \ell \leq d\}$, where $l_{\mathbf{x}} < \kappa_\ell$ (see SA, Section SA-3 for details). Let $\mathbf{b}_{\mathbf{x}} = (b_{\mathbf{x},1}, \ldots, b_{\mathbf{x},d})$ collect the interval lengths $b_{\mathbf{x},\ell} = |t_{\ell,l_{\mathbf{x}}+1} - t_{\ell,l_{\mathbf{x}}}|$. In addition:

(a) For $\ell = 1, \ldots, d$, $\sup_{\mathbf{x}\in[0,1]^d} |b_{\mathbf{x},\ell} - \kappa_\ell^{-1} g_\ell(\mathbf{x})^{-1}| = o(\kappa_\ell^{-1})$, where $g_\ell(\cdot)$ is bounded away from zero continuous.

(b) For all $\delta \in \Delta$ and $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \Lambda_m$, there exist constants $\eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{q}}$ such that

$$\int_\delta \frac{h_{\mathbf{x}}^{2m-2[\mathbf{q}]}}{\mathbf{b}_{\mathbf{x}}^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}} B_{\boldsymbol{u}_1,\mathbf{q}}(\mathbf{x}) B_{\boldsymbol{u}_2,\mathbf{q}}(\mathbf{x})\, d\mathbf{x} = \eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{q}} \operatorname{vol}(\delta),$$

where $\operatorname{vol}(\delta)$ denotes the volume of δ.

(c) There exists a set of points $\{\boldsymbol{\tau}_k\}_{k=1}^K$ such that $\boldsymbol{\tau}_k \in \operatorname{supp}(p_k(\cdot))$ for each $k = 1, \ldots, K$, and $\{\boldsymbol{\tau}_k\}_{k=1}^K$ can be assigned into $J + \check{J} < \infty$ groups such that $\{\boldsymbol{\tau}_{s,k_s}\}_{k_s=1}^{K_s}$, $s = 1, \ldots, J + \check{J}$, $\sum_{s=1}^{J+\check{J}} K_s = K$, and the following conditions hold: (i) For all $1 \leq s \leq J$, $\{\delta_{\boldsymbol{\tau}_{s,k_s}}\}_{k_s=1}^{K_s}$ are pairwise disjoint and $\operatorname{vol}([0, 1]^d \setminus \bigcup_{k_s=1}^{K_s} \delta_{\boldsymbol{\tau}_{s,k_s}}) = o(1)$; and (ii) for all $J + 1 \leq s \leq J + \check{J}$, $\operatorname{vol}(\bigcup_{k_s=1}^{K_s} \delta_{\boldsymbol{\tau}_{s,k_s}}) = o(1)$.

Part (a) slightly strengthens the quasi-uniform condition imposed in Assumption 2, but allows for quite general transformations of the knot location. Part (b) ensures that the "local" integral of the product between any two $B_{\boldsymbol{u},\mathbf{q}}(\cdot)$ for $\boldsymbol{u} \in \Lambda_m$, which depend on the basis but not $\mu(\mathbf{x})$, is proportional to the volume of the cell. The scaling factor is due to the use of the lengths of intervals on each axis (denoted by $\mathbf{b}_{\mathbf{x}}$) to characterize the approximation error for a tensor-product partition, instead of the more general diameter used in Section 2. Finally, part (c) describes how the supports of the basis functions cover the whole support of data. Specifically, it requires that the approximating basis $\mathbf{p}(\cdot)$ can be divided into $J + \check{J}$ groups. The supports of functions in each of the first J groups constitute "almost" complete covers of $\mathcal{X}$. In contrast, the supports of functions in other groups are negligible in terms of volume. In such a case, we refer to J as the number of complete covers generated by the supports of basis functions. For tensor product B-splines (with simple knots) and wavelets,

each subrectangle in Δ can be associated with one basis function in $\mathbf{p}(\cdot)$ and the supports of the remaining functions are asymptotically negligible in terms of volume. Thus, $J = 1$ in these two examples. For piecewise polynomials of total order m, within each subrectangle the unknown function is approximated by a multivariate polynomial of degree $m-1$, and thus $J = \binom{d+m-1}{m-1}$. This condition is used to ensure that the summation over the number of basis functions converges to a well-defined integral as $K \asymp h^{-d} \to \infty$.

We then have the following result for $\widehat{\mu}_0(\mathbf{x})$, proven in the SA, Section SA-10.7.

THEOREM 4.2 (Asymptotic IMSE). *Suppose that the conditions in Theorem 4.1 and Assumption 6 hold. Then, for* $[\mathbf{q}] = 0$,

$$\mathscr{V}_{\boldsymbol{\kappa},\mathbf{0}} = \left(\prod_{\ell=1}^{d} \kappa_\ell\right)\mathscr{V}_{\mathbf{0}} + o\big(h^{-d}\big), \qquad \mathscr{V}_{\mathbf{0}} = J\int_{[0,1]^d} \frac{\sigma^2(\mathbf{x})}{f(\mathbf{x})}\left(\prod_{\ell=1}^{d} g_\ell(\mathbf{x})\right) w(\mathbf{x})\,d\mathbf{x},$$

and, provided that (2.3) *holds,*

$$\mathscr{B}_{\boldsymbol{\kappa},\mathbf{0}} = \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m} \boldsymbol{\kappa}^{-(\boldsymbol{u}_1+\boldsymbol{u}_2)}\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} + o\big(h^{2m}\big),$$

$$\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} = \eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}}\int_{[0,1]^d} \frac{\partial^{\boldsymbol{u}_1}\mu(\mathbf{x})\partial^{\boldsymbol{u}_2}\mu(\mathbf{x})}{\mathbf{g}(\mathbf{x})^{\boldsymbol{u}_1+\boldsymbol{u}_2}} w(\mathbf{x})\,d\mathbf{x}.$$

The bias approximation requires the approximate orthogonality condition (2.3) which is satisfied by B-splines, wavelets and piecewise polynomials. It appears to be an open question whether $\mathscr{V}_{\boldsymbol{\kappa},\mathbf{q}}$ and $\mathscr{B}_{\boldsymbol{\kappa},\mathbf{q}}$ converge to a well-defined limit when general basis functions are considered. Cattaneo and Farrell [13] showed convergence to well-defined limits for piecewise polynomials, but their result is not easy to extend to cover other bases functions without imposing $\mathbf{q} = \mathbf{0}$ and the approximate orthogonality condition (2.3). The SA (Section SA-3) contains more details and other results.

Theorem 4.2 justifies the IMSE-optimal choice of number of knots:

$$\boldsymbol{\kappa}_{\mathtt{IMSE},\mathbf{0}} = \operatorname*{arg\,min}_{\boldsymbol{\kappa}\in\mathbb{Z}^d_{++}}\left\{\frac{1}{n}\left(\prod_{\ell=1}^{d}\kappa_\ell\right)\mathscr{V}_{\mathbf{0}} + \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m}\boldsymbol{\kappa}^{-(\boldsymbol{u}_1+\boldsymbol{u}_2)}\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}}\right\},$$

and, in particular, when the same number of knots is used in all margins,

$$\kappa_{\mathtt{IMSE},\mathbf{0}} = \left\lceil \left(\frac{2m\sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m}\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}}}{d\mathscr{V}_{\mathbf{0}}}\right)^{\frac{1}{2m+d}} n^{\frac{1}{2m+d}} \right\rceil.$$

Data-driven versions of this IMSE-optimal choice, and extensions to derivative estimation, are discussed in the SA (Section SA-8) and fully implemented in our companion general-purpose R package `lspartition` [14]. While beyond the scope of this paper, it would be of interest to study the theoretical properties of cross-validation methods as an alternative way of constructing IMSE-optimal tuning parameter selectors for partitioning-based estimators.

4.2. *Convergence rates.* Theorem 4.1 immediately delivers the L_2 convergence rate for the point estimator $\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x})$. For completeness, we also establish its uniform convergence rate. Recall that $\nu > 0$.

THEOREM 4.3 (Convergence rates). *Let Assumptions 1, 2 and 3 hold. Assume also that* $\sup_{\mathbf{x}\in\mathcal{X}}|\partial^{\mathbf{q}}\mu(\mathbf{x}) - \partial^{\mathbf{q}}s^*(\mathbf{x})| \lesssim h^{m-[\mathbf{q}]}$ *with* s^* *defined in Assumption 4. Then, if* $\frac{\log n}{nh^d} = o(1)$,

$$\int_{\mathcal{X}}\big(\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x})\big)^2 w(\mathbf{x})\,d\mathbf{x} \lesssim_{\mathbb{P}} \frac{1}{nh^{d+2[\mathbf{q}]}} + h^{2(m-[\mathbf{q}])}.$$

If, in addition:

(i) $\mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ *and* $\frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2\nu}{4+2\nu}}}{nh^d} \lesssim 1$, *or*

(ii) $\mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] < \infty$ *and* $\frac{(\log n)^3}{nh^d} \lesssim 1$,

then

$$\sup_{\mathbf{x}\in\mathcal{X}} \big|\widehat{\partial^{\mathbf{q}}\mu}_0(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x})\big|^2 \lesssim_{\mathbb{P}} \frac{\log n}{nh^{d+2[\mathbf{q}]}} + h^{2(m-[\mathbf{q}])}.$$

This theorem, proven in the SA, Section SA-10.11, shows that the partitioning-based estimators can attain the optimal L_2 and uniform convergence rate [37] by proper choice of partitioning scheme, under our high-level assumptions. (The full force of Assumption 4 is not needed for this result.) Cattaneo and Farrell [13] were the first to show existence of a series estimator (in particular, piecewise polynomials) attaining the optimal uniform convergence rate, a result that was later generalized to other series estimators in [3, 16].

5. Pointwise inference. We give pointwise inference based on classical undersmoothing and all three bias correction methods. All four point estimators take the form $\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) = \widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_j(\mathbf{x}_i)y_i]$, where $j = 0$ corresponds to the conventional partitioning estimator, and $j = 1, 2, 3$ refer to the three distinct bias correction strategies. Infeasible inference would be based on the standardized t-statistics

$$T_j(\mathbf{x}) = \frac{\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x})}{\sqrt{\Omega_j(\mathbf{x})/n}}, \quad \Omega_j(\mathbf{x}) = \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\boldsymbol{\Sigma}_j\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x}),$$

where, for each $j = 0, 1, 2, 3$, $\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})$ are defined as $\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}$ in (3.1), (3.4), (3.5) and (3.6), respectively, but with sample averages and other estimators replaced by their population counterparts, and $\boldsymbol{\Sigma}_j := \mathbb{E}[\boldsymbol{\Pi}_j(\mathbf{x}_i)\boldsymbol{\Pi}_j(\mathbf{x}_i)'\sigma^2(\mathbf{x}_i)]$. These t-statistics are infeasible, but they nonetheless capture the additional variability introduced by the bias correction approach when $j = 1, 2, 3$, the key idea behind robust bias-corrected inference [6, 8]. We also discuss below Studentization, that is, replacing $\Omega_j(\mathbf{x})$ with a consistent estimator.

5.1. *Distributional approximation.* Our first result establishes the limiting distribution of the standardized t-statistics $T_j(\mathbf{x})$.

THEOREM 5.1 (Asymptotic normality). *Let Assumptions 1, 2, 3 and 4 hold. Assume* $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\varepsilon_i^2 \mathbb{1}\{|\varepsilon_i| > M\}|\mathbf{x}_i = \mathbf{x}] \to 0$ *as* $M \to \infty$, *and* $\frac{\log n}{nh^d} = o(1)$. *Furthermore, for* $j = 0$, *assume* $nh^{2m+d} = o(1)$; *and for* $j = 1, 2, 3$, *assume Assumption 5 holds and* $nh^{2m+d} \lesssim 1$.

Then, for each $j = 0, 1, 2, 3$ *and* $\mathbf{x} \in \mathcal{X}$, $\sup_{u\in\mathbb{R}} |\mathbb{P}[T_j(\mathbf{x}) \le u] - \Phi(u)| = o(1)$, *where* $\Phi(u)$ *denotes the cumulative distribution function of* $\mathsf{N}(0, 1)$.

This theorem, proven in the SA, Section SA-10.9, gives a valid Gaussian approximation for the t-statistics $T_j(\mathbf{x})$, pointwise in $\mathbf{x} \in \mathcal{X}$. The regularity conditions imposed are extremely mild, and in perfect quantitative agreement with those used in [3] for $j = 0$ (undersmoothing). For $j = 1, 2, 3$ (robust bias correction), the result is new and the restrictions are in qualitative agreement with those obtained for kernel-based nonparametrics.

5.2. *Implementation.* To make the results in Theorem 5.1 feasible, we replace $\Omega_j(\mathbf{x})$ with a consistent estimator. Specifically, we consider the four feasible t-statistics, $j = 0, 1, 2, 3$,

$$\widehat{T}_j(\mathbf{x}) = \frac{\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x})}{\sqrt{\widehat{\Omega}_j(\mathbf{x})/n}}, \quad \widehat{\Omega}_j(\mathbf{x}) = \widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})'\widehat{\boldsymbol{\Sigma}}_j\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x}),$$

$$\widehat{\boldsymbol{\Sigma}}_j = \mathbb{E}_n\big[\boldsymbol{\Pi}_j(\mathbf{x}_i)\boldsymbol{\Pi}_j(\mathbf{x}_i)'\widehat{\varepsilon}_{i,j}^2\big], \quad \widehat{\varepsilon}_{i,j} = y_i - \widehat{\mu}_j(\mathbf{x}_i). \tag{5.1}$$

Once the basis functions and partitioning schemes are chosen, the statistic $\widehat{T}_j(\mathbf{x})$ is readily implementable. The following theorem gives sufficient conditions for valid pointwise inference.

THEOREM 5.2 (Variance consistency). *Let Assumptions 1, 2, 3 and 4 hold. If $j = 1, 2, 3$, also let Assumption 5 hold. In addition, assume one of the following holds*:

(i) $\mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ *and* $\frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2\nu}{4+2\nu}}}{nh^d} = o(1)$, *or*

(ii) $\mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] < \infty$ *and* $\frac{(\log n)^3}{nh^d} = o(1)$.

Then, for each $j = 0, 1, 2, 3$, $|\widehat{\Omega}_j(\mathbf{x}) - \Omega_j(\mathbf{x})| = o_{\mathbb{P}}(h^{-d-2[\mathbf{q}]})$.

This result, proven in the SA, Section SA-10.12, together with Theorem 5.1, delivers feasible inference. Valid $100(1-\alpha)\%$, $\alpha \in (0, 1)$, confidence intervals for $\partial^{\mathbf{q}}\mu(\mathbf{x})$ are formed in the usual way:

$$\big[\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) \pm \Phi^{-1}(1-\alpha/2) \cdot \sqrt{\widehat{\Omega}_j(\mathbf{x})/n}\big], \quad j = 0, 1, 2, 3.$$

For $j = 1, 2, 3$, the IMSE-optimal partitioning scheme choice derived in Section 4 can be used directly, while for $j = 0$ the partitioning has to be undersmoothed (i.e., made finer than the IMSE-optimal choice) in order to obtain valid confidence intervals. Our results generalize, under weaker conditions, prior work on univariate regression splines [28, 43, 44].

6. Uniform inference. We next give a valid distributional approximation for the *whole* process $\{\widehat{T}_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, for each $j = 0, 1, 2, 3$. We establish this approximation using two distinct coupling strategies. We then propose a simulation-based feasible implementation of the result. We close by applying our results to construct valid confidence bands for $\partial^{\mathbf{q}}\mu(\cdot)$.

6.1. *Strong approximations.* The stochastic processes $\{\widehat{T}_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ are not asymptotically tight and, therefore, do not converge weakly in $\mathcal{L}^\infty(\mathcal{X})$, where $\mathcal{L}^\infty(\mathcal{X})$ denotes the set of all (uniformly) bounded real functions on $\mathcal{X}$ equipped with uniform norm. Nevertheless, their finite sample distribution can be approximated by carefully constructed Gaussian processes (in a possibly enlarged probability space).

We first employ the following lemma to simplify the problem. Recall that r_n is some nonvanishing positive sequence and $\nu > 0$.

LEMMA 6.1 (Hats Off). *Let Assumptions 1, 2, 3 and 4 hold. Assume one of the following holds*:

(i) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}|\mathbf{x}_i = \mathbf{x}] < \infty$ *and* $\frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2+2\nu}{2+\nu}}}{nh^d} = o(r_n^{-2})$; *or*

(ii) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)|\mathbf{x}_i = \mathbf{x}] < \infty$ *and* $\frac{(\log n)^4}{nh^d} = o(r_n^{-2})$.

Furthermore, if $j = 0$, assume $nh^{d+2m} = o(r_n^{-2})$; and, if $j = 1, 2, 3$, assume Assumption 5 holds and $nh^{d+2m+2\varrho} = o(r_n^{-2})$. Then

$$\sup_{\mathbf{x}\in\mathcal{X}} \big|\widehat{T}_j(\mathbf{x}) - t_j(\mathbf{x})\big| = o_{\mathbb{P}}(r_n^{-1}), \quad t_j(\mathbf{x}) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}} \mathbb{G}_n\big[\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i\big].$$

Lemma 6.1 requires that the estimation and sampling uncertainty of $\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})$ and $\widehat{\Omega}_j(\mathbf{x})$, as well as the smoothing bias of $\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x})$, be negligible uniformly over $\mathbf{x} \in \mathcal{X}$. The proof is

given in Section 8.2 and relies on a new technical lemma stated in Section 8.1. This technical approximation step allows us to focus on developing a distributional approximation for the infeasible stochastic processes $\{t_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, $j = 0, 1, 2, 3$. We make precise our uniform distributional approximation in the following definition.

DEFINITION 6.1 (Strong approximation). For each $j = 0, 1, 2, 3$, the law of the stochastic process $\{t_j(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ is approximated by that of a Gaussian process $\{Z_j(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ in $\mathcal{L}^\infty(\mathcal{X})$ if the following condition holds: in a sufficiently rich probability space, there exists a copy $t_j'(\cdot)$ of $t_j(\cdot)$ and a standard Normal random vector $\mathbf{N}_{K_j} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{K_j})$ with $K_j = \dim(\mathbf{\Pi}_j(\mathbf{x}))$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} |t_j'(\mathbf{x}) - Z_j(\mathbf{x})| = o_{\mathbb{P}}(r_n^{-1}), \quad Z_j(\mathbf{x}) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})' \mathbf{\Sigma}_j^{1/2}}{\sqrt{\Omega_j(\mathbf{x})}} \mathbf{N}_{K_j}.$$

This approximation is denoted by $t_j(\cdot) =_d Z_j(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$.

This definition gives the precise meaning of uniform distributional approximation of $t_j(\cdot)$ by a Gaussian process $Z_j(\cdot)$, and also provides the explicit characterization of such Gaussian process. We establish this strong approximation in two distinct ways. For $d = 1$, we develop a novel two-step coupling approach based on the classical Komlós–Major–Tusnády (KMT) construction [30, 31]. For $d > 1$, we apply an improved version of the classical Yurinskii construction [29].

6.1.1. *Unidimensional regressor.* Let $d = 1$. The following theorem gives a valid distributional approximation for $\{\widehat{T}_j(x) : x \in \mathcal{X}\}$ using the Gaussian process $\{Z_j(x) : x \in \mathcal{X}\}$, for $j = 0, 1, 2, 3$, in the sense of Definition 6.1.

THEOREM 6.1 (Strong approximation: KMT). *Let the assumptions and conditions of Lemma 6.1 hold with $d = 1$. If $j = 2, 3$, also assume $\frac{(\log n)^{3/2}}{\sqrt{nh}} = o(r_n^{-2})$. Then, for each $j = 0, 1, 2, 3$, $t_j(\cdot) =_d Z_j(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$, where $Z_j(\cdot)$ is given in Definition 6.1.*

The proof of this result, in Section 8.3, employs a two-step coupling approach:

Step 1. On a sufficiently rich probability space, there exists a copy $t_j'(\cdot)$ of $t_j(\cdot)$, and an i.i.d. sequence $\{\zeta_i : 1 \le i \le n\}$ of standard Normal random variables, such that

$$\sup_{x\in\mathcal{X}} |t_j'(x) - z_j(x)| = o_{\mathbb{P}}(r_n^{-1}), \quad z_j(x) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(x)'}{\sqrt{\Omega_j(x)}} \mathbb{G}_n\big[\mathbf{\Pi}_j(x_i)\sigma(x_i)\zeta_i\big].$$

Step 2. On a sufficiently rich probability space, there exists a copy $z_j'(\cdot)$ of $z_j(\cdot)$, and the standard Normal random vector $\mathbf{N}_{K_j}$ from Definition 6.1 such that $z_j'(\cdot) =_d \bar{Z}_j(\cdot)$ conditional on $\mathbf{X}$, where

$$\bar{Z}_j(x) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(x)' \bar{\mathbf{\Sigma}}_j^{1/2}}{\sqrt{\Omega_j(x)}} \mathbf{N}_{K_j}, \quad \bar{\mathbf{\Sigma}}_j := \mathbb{E}_n\big[\mathbf{\Pi}_j(x_i)\mathbf{\Pi}_j(x_i)'\sigma^2(x_i)\big],$$

and

$$\sup_{x\in\mathcal{X}} |\bar{Z}_j(x) - Z_j(x)| = o_{\mathbb{P}}(r_n^{-1}).$$

These two steps summarize our strategy for constructing the unconditionally Gaussian process $\{Z_j(x), x \in \mathcal{X}\}$ approximating the distribution of the whole t-statistic processes $\{t_j(x) : x \in \mathcal{X}\}$: we first couple $t_j(\cdot)$ to the process $z_j(\cdot)$, which is Gaussian only conditionally on $\mathbf{X}$ but not unconditionally (Step 1), and we then show that the unconditionally Gaussian process $Z_j(\cdot)$ approximates the distribution of $z_j(\cdot)$ (Step 2).

To complete the first coupling step, we employ a version of the classical KMT inequalities that applies to independent but nonidentically distributed random variables [35, 36]. We do this because the processes $\{t_j(x) : x \in \mathcal{X}\}$ are characterized by a sum of independent but not identically distributed random variables conditional on $\mathbf{X}$. This part of our proof is inspired by, but is distinct from, the one given in [24], Chapter 22, where a conditional strong approximation for smoothing splines is established. Our proof relies instead on the new general coupling Lemma 8.2 in Section 8.3.

The intermediate coupling result in Step 1 has the obvious drawback that the process $\{z_j(x) : x \in \mathcal{X}\}$ is Gaussian only conditionally on $\mathbf{X}$ but not unconditionally. Step 2 addresses this shortcoming by establishing an unconditional coupling, that is, approximating the distribution of the stochastic process $z_j(\cdot)$ by that of the (unconditional) Gaussian process $Z_j(\cdot)$. As shown in Section 8.3, verifying the second coupling step boils down to controlling the supremum of a Gaussian random vector of increasing dimension, and in particular the crux is to prove precise (rate) control on $\|\bar{\mathbf{\Sigma}}_j^{1/2} - \mathbf{\Sigma}_j^{1/2}\|$, $j = 0, 1, 2, 3$. Both $\bar{\mathbf{\Sigma}}_j$ and $\mathbf{\Sigma}_j$ are symmetric and positive semidefinite. Further, for $j = 0, 1$, $\lambda_{\min}(\mathbf{\Sigma}_j) \gtrsim h^d$ for generic partitioning-based estimators under our assumptions and, therefore, we use the bound

$$\|\mathbf{A}_1^{1/2} - \mathbf{A}_2^{1/2}\| \leq \lambda_{\min}(\mathbf{A}_2)^{-1/2}\|\mathbf{A}_1 - \mathbf{A}_2\|, \tag{6.1}$$

which holds for symmetric positive semidefinite $\mathbf{A}_1$ and symmetric positive definite $\mathbf{A}_2$ [4], Theorem X.3.8. Using this bound, we obtain unconditional coupling from conditional coupling without additional rate restrictions.

However, for $j = 2, 3$ the bound (6.1) cannot be used in general because $\mathbf{p}$ and $\tilde{\mathbf{p}}$ are typically not linearly independent, and hence $\mathbf{\Sigma}_j$ will be singular. To circumvent this problem, we employ the weaker bound [4], Theorem X.1.1: if $\mathbf{A}_1$ and $\mathbf{A}_2$ are symmetric positive semidefinite matrices, then

$$\|\mathbf{A}_1^{1/2} - \mathbf{A}_2^{1/2}\| \leq \|\mathbf{A}_1 - \mathbf{A}_2\|^{1/2}. \tag{6.2}$$

This bound can be used for any partitioning-based estimator, with or without bias correction, at the cost of slowing the approximation error rate r_n when constructing the unconditional coupling, and hence leading to the stronger side rate condition as shown in the Theorem 6.1 below. When $r_n = 1$, there is no rate penalty, while the penalty is only in terms of $\log n$ terms when $r_n = \sqrt{\log n}$ (as in Theorem 6.4 further below). Furthermore, for certain partitioning-based series estimators it is still possible to use (6.1) even when $j = 2, 3$, as the following remark discusses.

REMARK 6.1 (Square-root convergence and improved rates). The additional restriction imposed in Theorem 6.1 for $j = 2, 3$, that $(\log n)^{3/2}/\sqrt{nh} = o(r_n^{-2})$, can be dropped in some special cases. For some bases, it is possible to find a transformation matrix $\mathbf{\Upsilon}$, with $\|\mathbf{\Upsilon}\|_\infty \lesssim 1$, and a basis $\check{\mathbf{p}}$, which obeys Assumption 3, such that $(\mathbf{p}(\cdot)', \tilde{\mathbf{p}}(\cdot)')' = \mathbf{\Upsilon}\check{\mathbf{p}}(\cdot)$. In other words, the two bases $\mathbf{p}$ and $\tilde{\mathbf{p}}$ can be expressed in terms of another basis $\check{\mathbf{p}}$ without linear dependence. Then a positive lower bound holds for $\lambda_{\min}(\mathbf{\Sigma}_j)$, $j = 2, 3$, implying that the bound (6.1) can be used instead of (6.2). For example, for piecewise polynomials and B-splines with equal knot placements for $\mathbf{p}$ and $\tilde{\mathbf{p}}$, a natural choice of $\check{\mathbf{p}}$ is simply a higher-order polynomial basis on the same partition. Since each function in $\mathbf{p}$ and $\check{\mathbf{p}}$ is a polynomial on each $\delta \in \Delta$ and nonzero on a fixed number of cells, the "local representation" condition $\|\mathbf{\Upsilon}\|_\infty \lesssim 1$ automatically holds. See the SA (Section SA-6) for more details.

An alternative unconditional strong approximation for general series estimators was obtained by [3] for the case of undersmoothing inference ($j = 0$). Their proof employs the classical Yurinskii's coupling inequality that controls the convergence rate of partial sums in terms of Euclidean norm, leading to the rate restriction $r_n^6 K^5/n \to 0$, up to $\log n$ terms, which does not depend on $\nu > 0$. In contrast, Theorem 6.1 employs a (conditional) KMT-type coupling and then a second (unconditional) coupling approximation, and make use of the banded structure of the Gram matrix formed by local bases, to obtain weaker restrictions. Under bounded polynomial moments, we require only $r_n^6 K^3/n^{3\nu/(2+\nu)} \to 0$, up to $\log n$ terms. For example, when $\nu = 2$ and $r_n = \sqrt{\log n}$, this translates to $K^2/n \to 0$, up to $\log n$ terms, which is weaker than previous results in the literature. Under the subexponential conditional moment restriction, the rate condition can be relaxed all the way to $K/n \to 0$, up to $\log n$ terms, which appears to be a minimal condition. This is for the entire t-statistic process. In addition, Theorem 6.1 gives novel strong approximation results for robust bias-corrected t-statistic processes.

REMARK 6.2 (Strong approximation: KMT for Haar basis). Our two-step coupling approach builds on the new coupling Lemma 8.2, which appears to be hard to extend to $d > 1$, except for the important special case the undersmoothed ($j = 0$) t-statistic process $\{\widehat{T}_0(x) : x \in \mathcal{X}\}$ constructed using Haar basis, which is a spline, wavelet, and piecewise polynomial with $m = 1$. In the SA, Section SA-5.1, we establish $t_0(\cdot) =_d Z_0(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$ for any $d \geq 1$ under the same conditions of Lemma 6.1.

6.1.2. *Multidimensional regressors.* Let $d \geq 1$. The method of proof employed to establish Theorem 6.1 does not extend easily to multivariate regressors ($d > 1$) in general. Therefore, we present an alternative strong approximation result based on an improved version of the classical Yurinskii's coupling inequality, recently developed by [2].

THEOREM 6.2 (Strong approximation: Yurinskii). *Let the assumptions and conditions of Lemma 6.1 hold. Furthermore, assume $\nu \geq 1$ and $\frac{(\log n)^4}{nh^{3d}} = o(r_n^{-6})$. Then, for each $j = 0, 1, 2, 3$, $t_j(\cdot) =_d Z_j(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$, where $Z_j(\cdot)$ is given in Definition 6.1.*

This strong approximation result, proven in Section 8.4, does not have optimal (i.e., minimal) restrictions, but nonetheless improves on previous results by exploiting the specific structure of the partitioning-based estimators, while also allowing for any $d \geq 1$. Specifically, the result sets $\nu = 1$ and requires $r_n^6 K^3/n \to 0$, up to $\log n$ terms. While not optimal when $\nu > 3$ (see Remark 6.2 for a counterexample), it improves on the condition $r_n^6 K^5/n \to 0$, up to $\log n$ terms, mentioned previously. In addition, Theorem 6.2 gives novel strong approximation results for robust bias-corrected t-statistic processes.

6.2. *Implementation.* We present a simple plug-in approach that gives a (feasible) approximation to the infeasible standardized Gaussian processes $\{Z_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, in order to conduct inference using the results in Theorem 6.1 or Theorem 6.2. In the SA (Section SA-5.2), we also give another plug-in approach and one based on the wild bootstrap. The following definition gives a precise description of how the approximation works.

DEFINITION 6.2 (Simulation-based strong approximation). Let $\mathbb{P}^*[\cdot] = \mathbb{P}[\cdot|\mathbf{y}, \mathbf{X}]$ denote the probability operator conditional on the data. For each $j = 0, 1, 2, 3$, the law of the Gaussian process $\{Z_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ is approximated by a (feasible) Gaussian process $\{\widehat{Z}_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, with known distribution conditional on the data $(\mathbf{y}, \mathbf{X})$, in $\mathcal{L}^\infty(\mathcal{X})$, if the

following condition holds: on a sufficiently rich probability space there exists a copy $\widehat{Z}'_j(\cdot)$ of $\widehat{Z}_j(\cdot)$ such that $\widehat{Z}'_j(\cdot) =_d Z_j(\cdot)$ conditional on the data, and

$$\mathbb{P}^*\Big[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{Z}'_j(\mathbf{x}) - Z_j(\mathbf{x})| \geq \eta r_n^{-1}\Big] = o_{\mathbb{P}}(1) \quad \forall \eta > 0,$$

where, for a $\mathbf{N}_{K_j} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{K_j})$ with $K_j = \dim(\mathbf{\Pi}_j(\mathbf{x}))$,

$$\widehat{Z}_j(\mathbf{x}) = \frac{\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})'\widehat{\mathbf{\Sigma}}_j^{1/2}}{\sqrt{\widehat{\Omega}_j(\mathbf{x})}}\mathbf{N}_{K_j}, \quad \mathbf{x}\in\mathcal{X},\ j = 0,1,2,3.$$

This approximation is denoted by $\widehat{Z}_j(\cdot) =_{d^*} Z_j(\cdot) + o_{\mathbb{P}^*}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$.

From a practical perspective, Definition 6.2 implies that sampling from $\widehat{Z}_j(\cdot)$, conditional on the data, is possible and provides a valid distributional approximation of $Z_j(\cdot)$, for each $j = 0, 1, 2, 3$. The feasible process $\widehat{Z}_j(\cdot)$ relies on a direct plug-in approach, where all the unknown quantities are replaced by consistent estimators already used in forming $\widehat{T}_j(\mathbf{x})$. Resampling is from a multivariate standard Gaussian of dimension K_j, not n.

THEOREM 6.3 (Plug-in approximation). *Let the assumptions and conditions of Lemma 6.1 hold. Furthermore, for $j = 2, 3$:*

(i) *when* $\sup_{\mathbf{x}\in\mathcal{X}}\mathbb{E}[|\varepsilon_i|^{2+\nu}|\mathbf{x}_i = \mathbf{x}] < \infty$, *assume* $\dfrac{n^{\frac{1}{2+\nu}}(\log n)^{\frac{4+3\nu}{4+2\nu}}}{\sqrt{nh^d}} = o(r_n^{-2})$; *or*

(ii) *when* $\sup_{\mathbf{x}\in\mathcal{X}}\mathbb{E}[|\varepsilon_i|^3\exp(|\varepsilon_i|)|\mathbf{x}_i = \mathbf{x}] < \infty$, *assume* $\dfrac{(\log n)^{5/2}}{\sqrt{nh^d}} = o(r_n^{-2})$.

Then, for each $j = 0, 1, 2, 3$, $\widehat{Z}_j(\cdot) =_{d^*} Z_j(\cdot) + o_{\mathbb{P}^*}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$, where $\widehat{Z}_j(\cdot)$ is given in Definition 6.2.

This result, proven in Section 8.5, strengthens the rate condition for $j = 2, 3$ compared to Theorems 6.1 ($d = 1$) and 6.2 ($d \geq 1$) only by logarithmic factors when $r_n = \sqrt{\log n}$. Moreover, if the structure discussed in Remark 6.1 holds, then this additional condition can be dropped.

6.3. *Application*: *Confidence bands*. A natural application of Theorems 6.1, 6.2 and 6.3 is to construct confidence bands for the regression function or its derivatives. Specifically, for $j = 0, 1, 2, 3$ and $\alpha \in (0, 1)$, we seek a quantile $q_j(\alpha)$ such that

$$\mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| \leq q_j(\alpha)\Big] = 1 - \alpha + o(1),$$

which then can be used to construct uniform $100(1-\alpha)$-percent confidence bands for $\partial^{\mathbf{q}}\mu(\mathbf{x})$ of the form

$$\big[\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) \pm q_j(\alpha)\sqrt{\widehat{\Omega}_j(\mathbf{x})/n} : \mathbf{x}\in\mathcal{X}\big].$$

The following theorem, proven in Section 8.6, establishes a valid distributional approximation for the suprema of the t-statistic processes $\{\widehat{T}_j(\mathbf{x}) : \mathbf{x}\in\mathcal{X}\}$ using [19], Lemma 2.4, to convert our strong approximation results into convergence of distribution functions in terms of Kolmogorov distance.

THEOREM 6.4 (Confidence bands). *Let the conditions of Theorem 6.1 or Theorem 6.2 hold with $r_n = \sqrt{\log n}$. If the corresponding conditions of Theorem 6.3 hold for each $j = 0, 1, 2, 3$, then*

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| \leq u\Big] - \mathbb{P}^*\Big[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{Z}_j(\mathbf{x})| \leq u\Big]\Big| = o_{\mathbb{P}}(1).$$

Chernozhukov et al. [18, 19] recently showed that if one is only interested in the supremum of an empirical process rather than the *whole* process, then the sufficient conditions for distributional approximation could be weakened compared to earlier literature. Their result applied Stein's method for Normal approximation to show that suprema of general empirical processes can be approximated by a sequence of suprema of Gaussian processes, under the usual undersmoothing conditions (i.e., $j = 0$). They illustrate their general results by considering t-statistic processes for both kernel-based and series-based nonparametric regression: [19], Remark 3.5, establishes a result analogous to Theorem 6.4 under the side rate condition $K/n^{1-2/(2+\nu)} = o(1)$, up to $\log n$ terms (with $q = 2 + \nu$ in their notation). In comparison, our result for $j = 0$ and $d = 1$ in Theorem 6.4, under the same moment conditions, requires exactly the same side condition, up to $\log n$ terms. Theorems 6.1 and 6.4 show that the *whole* t-statistic process for partitioning-based series estimators, and not just the suprema thereof, can be approximated under the same weak conditions when $d = 1$. The same result holds for sub-exponential moments, where the rate condition becomes minimal: $K/n = o(1)$, up to $\log n$ factors. In addition, Theorem 6.4 gives new inference results for bias-corrected estimators ($j = 1, 2, 3$).

For the case of special univariate regression splines, [43] constructs conservative confidence bands under the assumption of normal errors and the rate restriction $K^2/n = o(1)$. In comparison, the confidence band constructed using Theorem 6.4 has asymptotically exact coverage rate, and requires substantially weaker tuning parameter rate restrictions.

7. Simulations. We conducted a Monte Carlo investigation of the finite sample performance of our methods. Only a summary is given here, while the SA (Section SA-9) contains complete results and details.

We considered three univariate ($d = 1$), two bivariate ($d = 2$), and two trivariate ($d = 3$) data generating processes. We shall summarize one univariate design here for brevity. We set $\mu(x) = \sin(\pi x - \pi/2)/(1 + 2(2x - 1)^2(\text{sign}(2x - 1) + 1))$, with $\text{sign}(\cdot)$ denoting the sign function. We generate samples $\{(y_i, x_i) : i = 1, \ldots, n\}$ from $y_i = \mu(x_i) + \varepsilon_i$, where $x_i \sim \mathsf{U}[0, 1]$ and $\varepsilon_i \sim \mathsf{N}(0, 1)$, independent of each other. We consider 5000 simulated datasets with $n = 1000$ each. Results based on splines and wavelets are presented. We use linear splines or Daubechies (father) wavelets of order 2 ($m = 2$) to form the point estimator $\widehat{\mu}_0(x)$, and quadratic splines or Daubechies wavelets of order 3 ($\tilde{m} = 3$) for bias correction, on the same evenly spaced partitioning for point estimation and bias correction ($\Delta = \tilde{\Delta}$).

Table 1 reports (simulated) root mean squared error for point estimators (column "RMSE"), coverage rate and average interval length of pointwise 95% nominal confidence intervals at $x = 0.5$ (columns "CR" and "AL"), and analogous uniform inference results (columns under "Uniform"). For B-splines, κ is set using either an infeasible IMSE-optimal choice ($\kappa_{\mathtt{IMSE}}$), a rule-of-thumb estimate ($\hat{\kappa}_{\mathtt{ROT}}$), or a direct plug-in estimate ($\hat{\kappa}_{\mathtt{DPI}}$). For wavelets, the tuning parameter is instead the resolution level (resp., $s_{\mathtt{IMSE}}$, $\hat{s}_{\mathtt{ROT}}$ or $\hat{s}_{\mathtt{DPI}}$), which is the logarithm of the number of subintervals (to base 2). Finally, Table 1 reports all four (estimation and) inference methods discussed in this paper ($j = 0, 1, 2, 3$), except for wavelets with plug-in bias correction ($j = 3$) because of the lack of smoothness of low-order wavelet bases.

Numerical findings are consistent with our theoretical results. Robust bias correction seems to perform quite well, always delivering close-to-correct coverage, both pointwise and uniformly. In addition, our rule-of-thumb (ROT) and direct plug-in (DPI) knot selection procedures for tensor-product partitions exhibited good numerical performance.

TABLE 1
Simulation Evidence

B-Splines ($m = 2$, $\tilde{m} = 3$, $\Delta = \tilde{\Delta}$, Evenly Spaced Partition)

			Pointwise		Uniform	
	κ	RMSE	CR	AL	UCR	AW
$j = 0$						
$\kappa_{\mathtt{IMSE}}$	3.0	0.046	91.6	0.328	79.9	0.384
$\hat{\kappa}_{\mathtt{ROT}}$	4.1	0.002	94.9	0.254	90.1	0.433
$\hat{\kappa}_{\mathtt{DPI}}$	4.7	0.008	93.8	0.311	91.6	0.460
$j = 1$						
$\kappa_{\mathtt{IMSE}}$	3.0	0.003	94.8	0.226	93.8	0.426
$\hat{\kappa}_{\mathtt{ROT}}$	4.1	0.008	94.8	0.297	93.5	0.473
$\hat{\kappa}_{\mathtt{DPI}}$	4.7	0.007	94.9	0.294	93.2	0.497
$j = 2$						
$\kappa_{\mathtt{IMSE}}$	3.0	0.004	94.7	0.268	94.1	0.443
$\hat{\kappa}_{\mathtt{ROT}}$	4.1	0.008	94.8	0.312	93.4	0.497
$\hat{\kappa}_{\mathtt{DPI}}$	4.7	0.004	94.8	0.330	93.6	0.526
$j = 3$						
$\kappa_{\mathtt{IMSE}}$	3.0	0.016	90.0	0.320	89.0	0.413
$\hat{\kappa}_{\mathtt{ROT}}$	4.1	0.007	94.2	0.275	93.0	0.463
$\hat{\kappa}_{\mathtt{DPI}}$	4.7	0.005	94.2	0.322	93.2	0.490

Wavelets ($m = 2$, $\tilde{m} = 3$, $\Delta = \tilde{\Delta}$, Evenly Spaced Partition)

			Pointwise		Uniform	
	s	RMSE	CR	AL	UCR	AW
$j = 0$						
$s_{\mathtt{IMSE}}$	3.0	0.002	94.2	0.476	91.1	0.509
$\hat{s}_{\mathtt{ROT}}$	2.0	0.002	94.2	0.476	91.1	0.509
$\hat{s}_{\mathtt{DPI}}$	2.8	0.002	94.2	0.476	91.1	0.509
$j = 1$						
$s_{\mathtt{IMSE}}$	3.0	0.036	93.6	0.449	89.9	0.504
$\hat{s}_{\mathtt{ROT}}$	2.0	0.036	93.6	0.449	89.9	0.504
$\hat{s}_{\mathtt{DPI}}$	2.8	0.036	93.6	0.449	89.9	0.504
$j = 2$						
$s_{\mathtt{IMSE}}$	3.0	0.009	94.2	0.523	91.4	0.576
$\hat{s}_{\mathtt{ROT}}$	2.0	0.009	94.2	0.523	91.4	0.576
$\hat{s}_{\mathtt{DPI}}$	2.8	0.009	94.2	0.523	91.4	0.576

Notes: (i) Pointwise = pointwise inference at $x = 0.5$, Uniform = uniform inference.
(ii) RMSE = root MSE of point estimator, CR = coverage rate of 95% nominal confidence intervals, AL = average interval length of 95% nominal confidence intervals.
(iii) UCR = uniform coverage rate of 95% nominal confidence band, AW = average width of 95% nominal confidence band.
(iv) $\kappa_{\mathtt{IMSE}}$ and $s_{\mathtt{IMSE}}$ = infeasible IMSE-optimal number of partitions, $\hat{\kappa}_{\mathtt{ROT}}$ and $\hat{s}_{\mathtt{ROT}}$ = feasible rule-of-thumb (ROT) implementation of $\kappa_{\mathtt{IMSE}}$, $\hat{\kappa}_{\mathtt{DPI}}$ and $\hat{s}_{\mathtt{DPI}}$ = feasible direct plug-in (DPI) implementation of $\kappa_{\mathtt{IMSE}}$. See Section SA-8 and Section SA-9 in the Supplemental Appendix for more details.

8. Main technical lemmas and proofs.

8.1. *Technical lemma.* Let $\widehat{\mathbf{Q}}_m = \mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']$, $\widehat{\mathbf{Q}}_{\tilde{m}} = \mathbb{E}_n[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)']$, $\mathbf{Q}_m = \mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']$, and $\mathbf{Q}_{\tilde{m}} = \mathbb{E}[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)']$.

LEMMA 8.1. *Let Assumptions 1, 2, 3 and 5 hold. If* $\frac{\log n}{nh^d} = o(1)$, *then*: (i) $\|\widehat{\mathbf{Q}}_m - \mathbf{Q}_m\| \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log n}{nh^d}}$, $\|\widehat{\mathbf{Q}}_m - \mathbf{Q}_m\|_\infty \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log n}{nh^d}}$; (ii) $\|\widehat{\mathbf{Q}}_m\| \lesssim_{\mathbb{P}} h^d$, $\|\widehat{\mathbf{Q}}_m^{-1}\|_\infty \lesssim_{\mathbb{P}} h^{-d}$; (iii) *for each* $j = 0, 1, 2, 3$, $\sup_{\mathbf{x}\in\mathcal{X}} \|\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\|_\infty \lesssim h^{-d-[\mathbf{q}]}$, $\sup_{\mathbf{x}\in\mathcal{X}} \|\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})' - \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\|_\infty \lesssim h^{-d-[\mathbf{q}]}\sqrt{\frac{\log n}{nh^d}}$, $\inf_{\mathbf{x}\in\mathcal{X}} \|\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\| \gtrsim h^{-d-[\mathbf{q}]}$; *and* (iv) *for* $j = 0, 1, 2, 3$, $\sup_{\mathbf{x}\in\mathcal{X}} \Omega_j(\mathbf{x}) \lesssim h^{-d-2[\mathbf{q}]}$ *and* $\inf_{\mathbf{x}\in\mathcal{X}} \Omega_j(\mathbf{x}) \gtrsim h^{-d-2[\mathbf{q}]}$.

PROOF. SA, Section SA-10. □

These results for $\widehat{\mathbf{Q}}_m$ and $\mathbf{Q}_m$ also hold for $\widehat{\mathbf{Q}}_{\tilde{m}}$ and $\mathbf{Q}_{\tilde{m}}$ under Assumption 5. See the SA (Section SA-2) for details and other related results.

8.2. *Proof of Lemma 6.1.* First, suppose condition (i) holds. Theorem SA-4.2 of the SA shows $\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\Omega}_0(\mathbf{x}) - \Omega_0(\mathbf{x})| \lesssim_{\mathbb{P}} n^{-\frac{1}{2}}h^{-\frac{3d}{2}-2[\mathbf{q}]}[(\log n)^{\frac{1}{2}} + n^{\frac{1}{2+\nu}}(\log n)^{\frac{\nu}{4+2\nu}} + \sqrt{n}h^{\frac{d}{2}+m}]$ and, for $j = 1, 2, 3$, $\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\Omega}_j(\mathbf{x}) - \Omega_j(\mathbf{x})| \lesssim_{\mathbb{P}} n^{-\frac{1}{2}}h^{-\frac{3d}{2}-2[\mathbf{q}]}[(\log n)^{\frac{1}{2}} + n^{\frac{1}{2+\nu}}(\log n)^{\frac{\nu}{4+2\nu}} + \sqrt{n}h^{\frac{d}{2}+m+\varrho}]$. Then, for $j = 0, 1, 2, 3$, $\sup_{\mathbf{x}\in\mathcal{X}} |(\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x}))/\sqrt{\Omega_j(\mathbf{x})/n} - (\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x}))/\sqrt{\widehat{\Omega}_j(\mathbf{x})/n}| \lesssim_{\mathbb{P}} \sqrt{n}h^{3d/2+3[\mathbf{q}]} \sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\partial^{\mathbf{q}}\mu}_j(\mathbf{x}) - \partial^{\mathbf{q}}\mu(\mathbf{x})| \sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\Omega}_j(\mathbf{x}) - \Omega_j(\mathbf{x})| = o_{\mathbb{P}}(r_n^{-1})$, where the result follows from Lemma 8.1, Theorem SA-4.1, the uniform convergence rate of $\widehat{\Omega}_j(\mathbf{x})$, and the rate conditions imposed.

The result under the conditions in (ii) follows similarly.

8.3. *Proof of Theorem 6.1.* We first prove the following general lemma. Let $TV_{\mathcal{X}}(g(\cdot))$ denote the total variation of $g(\cdot)$ on $\mathcal{X} \subseteq \mathbb{R}$.

LEMMA 8.2 (Kernel-based KMT coupling). *Suppose* $\{(x_i, \varepsilon_i) : 1 \le i \le n\}$ *are i.i.d., with* $x_i \in \mathcal{X} \subseteq \mathbb{R}$ *and* $\sigma_i^2 := \sigma^2(x_i) = \mathbb{E}[\varepsilon_i^2|x_i]$. *Let* $\{A(x) := \mathbb{G}_n[\mathscr{K}(x, x_i)\varepsilon_i], x \in \mathcal{X}\}$ *be a stochastic process with* $\mathscr{K}(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}$ *an n-varying kernel function possibly depending on* $\mathbf{X}$. *Assume one of the following holds*:

(i) $\sup_{x\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}|x_i = x] < \infty$, *for some* $\nu > 0$, *and*

$$\sup_{x\in\mathcal{X}} \max_{1\le i\le n} |\mathscr{K}(x, x_i)| = o_{\mathbb{P}}\big(r_n^{-1}n^{-\frac{1}{2+\nu}+\frac{1}{2}}\big),$$

$$\sup_{x\in\mathcal{X}} TV_{\mathcal{X}}\big(\mathscr{K}(x,\cdot)\big) = o\big(r_n^{-1}n^{-\frac{1}{2+\nu}+\frac{1}{2}}\big); \quad \textit{or}$$

(ii) $\sup_{x\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)|x_i = x] < \infty$ *and*

$$\sup_{x\in\mathcal{X}} \max_{1\le i\le n} |\mathscr{K}(x, x_i)| = o_{\mathbb{P}}\big(r_n^{-1}(\log n)^{-1}\sqrt{n}\big),$$

$$\sup_{x\in\mathcal{X}} TV_{\mathcal{X}}\big(\mathscr{K}(x,\cdot)\big) = o\big(r_n^{-1}(\log n)^{-1}\sqrt{n}\big).$$

Then, on a sufficiently rich probability space, there exists a copy $A'(\cdot)$ *of* $A(\cdot)$, *and an i.i.d. sequence* $\{\zeta_i : 1 \le i \le n\}$ *of standard Normal random variables such that* $A(x) =_d \mathbb{G}_n[\mathscr{K}(x, x_i)\sigma_i\zeta_i] + o_{\mathbb{P}}(r_n^{-1})$ *in* $\mathcal{L}^\infty(\mathcal{X})$.

PROOF. Suppose condition (i) holds. Let $\{x_{(i)} : 1 \le i \le n\}$ be the order statistics of $\{x_i : 1 \le i \le n\}$ such that $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, which also induces the concomitants $\{\varepsilon_{[i]} : 1 \le i \le n\}$ and $\{\sigma^2_{[i]} = \sigma^2(x_{(i)}) : 1 \le i \le n\}$. Conditional on $\mathbf{X}$, $\{\varepsilon_{[i]} : 1 \le i \le n\}$ is an independent mean zero sequence with $\mathbb{V}[\varepsilon_{[i]}|\mathbf{X}] = \sigma^2_{[i]}$. By [36], Corollary 5, there exists a sequence of i.i.d. standard normal random variables $\{\zeta_{[i]} : 1 \le i \le n\}$ such that $\max_{1\le l\le n} |S_{l,n}| \lesssim_{\mathbb{P}} n^{\frac{1}{2+\nu}}$, where $S_{l,n} := \sum_{i=1}^{l}(\varepsilon_{[i]} - \sigma_{[i]}\zeta_{[i]})$. Then, using summation by parts,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^{n}\mathscr{K}(x,x_{(i)})(\varepsilon_{[i]}-\sigma_{[i]}\zeta_{[i]})\right| \\
&\quad= \sup_{x\in\mathcal{X}}\left|\mathscr{K}(x,x_{(n)})S_{n,n} - \sum_{i=1}^{n-1} S_{i,n}\big(\mathscr{K}(x,x_{(i+1)}) - \mathscr{K}(x,x_{(i)})\big)\right| \\
&\quad\le \left(\sup_{x\in\mathcal{X}}\max_{1\le i\le n}\big|\mathscr{K}(x,x_i)\big| + \sup_{x\in\mathcal{X}}\sum_{i=1}^{n-1}\big|\mathscr{K}(x,x_{(i+1)}) - \mathscr{K}(x,x_{(i)})\big|\right)\max_{1\le l\le n}|S_{l,n}|.
\end{aligned}$$

Since $\sum_{i=1}^{n-1}|\mathscr{K}(x,x_{(i+1)}) - \mathscr{K}(x,x_{(i)})| \le TV_{\mathcal{X}}(\mathscr{K}(x,\cdot))$, we have $A(x) =_d \mathbb{G}_n[\mathscr{K}(x, x_i)\sigma_i\zeta_i] + o_{\mathbb{P}}(r_n^{-1})$.

When (ii) holds, the proof is the same except that under the stronger moment restriction, $\max_{1\le l\le n}|S_{l,n}| \lesssim_{\mathbb{P}} \log n$ by [35], Theorem 1. □

To prove Theorem 6.1, for each $j = 0, 1, 2, 3$, let $\mathscr{K}(x,u) = \boldsymbol{\gamma}_{\mathbf{q},j}(x)'\boldsymbol{\Pi}_j(u)/\sqrt{\Omega_j(x)}$ and observe that $\sup_{x\in\mathcal{X}}\sup_{u\in\mathcal{X}}|\mathscr{K}(x,u)| \lesssim h^{-d/2}$. By Lemma 8.1, the uniform bound on the total variation of $\mathscr{K}(x,u)$ can be verified easily. Alternatively, simply note that $|\sum_{i=1}^{n-1} S_{i,n}(\mathscr{K}(x,x_{(i+1)}) - \mathscr{K}(x,x_{(i)}))| \le \|\boldsymbol{\gamma}_{\mathbf{q},j}(x)'/\sqrt{\Omega_j(x)}\|_\infty \|\sum_{i=1}^{n-1} S_{i,n}(\boldsymbol{\Pi}_j(x_{(i+1)}) - \boldsymbol{\Pi}_j(x_{(i)}))\|_\infty$. By Assumption 3 and Lemma 8.1, $\sup_{x\in\mathcal{X}}\|\boldsymbol{\gamma}_{\mathbf{q},j}(x)'/\sqrt{\Omega_j(x)}\|_\infty \lesssim h^{-d/2}$. Denote the lth element of $\boldsymbol{\Pi}_j(\cdot)$ by $\pi_{j,l}(\cdot)$. Then $\max_{1\le l\le K_j}|\sum_{i=1}^{n-1}(\pi_{j,l}(x_{(i+1)}) - \pi_{j,l}(x_{(i)})) \times S_{l,n}| \le \max_{1\le l\le K_j}\sum_{i=1}^{n-1}|\pi_{j,l}(x_{(i+1)}) - \pi_{j,l}(x_{(i)})| \max_{1\le \ell\le n}|S_{\ell,n}|$. By Assumptions 3 and 5, $\max_{1\le l\le K_j}\sum_{i=1}^{n-1}|\pi_{j,l}(x_{(i+1)}) - \pi_{j,l}(x_{(i)})| \lesssim 1$. Thus, using Lemma 8.2, under the corresponding moment conditions and rate restrictions, there exists an independent standard normal sequence $\{\zeta_i : 1 \le i \le n\}$ such that $\mathbb{G}_n[\mathscr{K}(x,x_i)\varepsilon_i] =_d z_j(x) + o_{\mathbb{P}}(r_n^{-1})$.

Next, note that

$$z_j(\mathbf{x}) =_{d|\mathbf{X}} \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\boldsymbol{\Sigma}_j^{1/2}\mathbf{N}_{K_j}/\sqrt{\Omega_j(\mathbf{x})} + \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'(\bar{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2})\mathbf{N}_{K_j}/\sqrt{\Omega_j(\mathbf{x})},$$

where $\mathbf{N}_{K_j}$ is a K_j-dimensional standard normal vector (independent of $\mathbf{X}$) and "$=_{d|\mathbf{X}}$" denotes that two processes have the same conditional distribution given $\mathbf{X}$. For the second term, by the Gaussian maximal inequality [22], Lemma 13, $\mathbb{E}[\|(\bar{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2})\mathbf{N}_{K_j}\|_\infty|\mathbf{X}] \lesssim \sqrt{\log n}\|\bar{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2}\|$. By the same argument as that in the proof of Lemma SA-2.1 in Section SA-10.1, $\|\bar{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| \lesssim_{\mathbb{P}} h^d(\log n/(nh^d))^{1/2}$. Then by [4], Theorem X.1.1, $\|\bar{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2}\| \lesssim_{\mathbb{P}} h^{d/2}(\log n/(nh^d))^{1/4}$. For $j = 0, 1$, a sharper bound is available: by [4], Theorem X.3.8, and Lemma 8.1, $\|\bar{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2}\| \le \lambda_{\min}(\boldsymbol{\Sigma}_j)^{-1/2}\|\bar{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| \lesssim_{\mathbb{P}} h^{d/2}\sqrt{\frac{\log n}{nh^d}}$. Thus, $\mathbb{E}[\sup_{\mathbf{x}\in\mathcal{X}}|\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'(\bar{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2})\mathbf{N}_{K_j}/\sqrt{\Omega_j(\mathbf{x})}||\mathbf{X}] \lesssim_{\mathbb{P}} h^{-\frac{d}{2}}\sqrt{\log n}\|\bar{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2}\| = o_{\mathbb{P}}(r_n^{-1})$. The results now follow from Markov inequality and dominated convergence.

8.4. *Proof of Theorem 6.2.* It suffices to verify the conditions in Lemma 39 of [2]. For $j = 0, 1, 2, 3$, let $\boldsymbol{\xi}_i = \frac{1}{\sqrt{n}} \boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i$. $\{\boldsymbol{\xi}_i : 1 \le i \le n\}$ is an i.i.d. sequence of K_j-dimensional random vectors, and

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\big[\|\boldsymbol{\xi}_i\|^2 \|\boldsymbol{\xi}_i\|_\infty\big] &= \mathbb{E}\big[\big\|\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i\big\|^2 \big\|\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i\big\|_\infty\big]/\sqrt{n} \\ &\lesssim \mathbb{E}\big[\boldsymbol{\Pi}_j(\mathbf{x}_i)'\boldsymbol{\Pi}_j(\mathbf{x}_i)|\varepsilon_i|^3\big]/\sqrt{n} \lesssim n^{-1/2}\end{aligned}$$

by Assumption 3, the moment condition imposed, and Lemma 8.1. On the other hand, let $\{\mathbf{g}_i : 1 \le i \le n\}$ be a sequence of independent Gaussian vectors with mean zero and variance $\boldsymbol{\Sigma}_j/n$. By properties of Gaussian random variables and Lemma 8.1, $(\mathbb{E}[\|\mathbf{g}_i\|_\infty^2])^{1/2} \lesssim \sqrt{\frac{\log n}{n}}$,

$$\sum_{i=1}^{n} \big(\mathbb{E}\big[\|\mathbf{g}_i\|^4\big]\big)^{1/2} \lesssim \operatorname{trace}\left(\sum_{i=1}^{n} \mathbb{E}\big[\boldsymbol{\xi}_i \boldsymbol{\xi}_i'\big]\right) \lesssim 1,$$

and thus $L_n := \sum_{i=1}^n \mathbb{E}[\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_\infty] + \sum_{i=1}^n \mathbb{E}[\|\mathbf{g}_i\|^2\|\mathbf{g}_i\|_\infty] \lesssim \sqrt{\frac{\log n}{n}}$. Then there exists a K_j-dimensional normal vector $\mathbf{N}_{K_j}$ with variance $\boldsymbol{\Sigma}_j$ such that for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left\|\sum_{i=1}^{n} \boldsymbol{\xi}_i - \mathbf{N}_{K_j}\right\|_\infty > 3r_n^{-1}h^{\frac{d}{2}}t\right) \\ &\quad\le \min_{\tau \ge 0}\big(2\mathbb{P}(\|\mathbf{Z}\|_\infty > \tau) + h^{-\frac{3d}{2}}t^{-3}r_n^3 L_n \tau^2\big) \lesssim t^{-3}r_n^3(\log n)^{\frac{3}{2}}/\sqrt{nh^{3d}},\end{aligned}$$

where $\mathbf{Z}$ is a K_j-dimensional standard Gaussian vector, and the second inequality follows by setting $\tau = C\sqrt{\log n}$ for a sufficiently large $C > 0$. Using $\sup_{x\in\mathcal{X}} \|\boldsymbol{\gamma}_{\mathbf{q},j}(x)'/\sqrt{\Omega_j(x)}\|_\infty \lesssim h^{-d/2}$ again, the result follows.

8.5. *Proof of Theorem 6.3.* For each $j = 0, 1, 2, 3$,

$$\widehat{Z}_j(\mathbf{x}) - Z_j(\mathbf{x}) = \left(\frac{\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})}{\widehat{\Omega}_j^{1/2}(\mathbf{x})} - \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\Omega_j^{1/2}(\mathbf{x})}\right)\widehat{\boldsymbol{\Sigma}}_j^{1/2}\mathbf{N}_{K_j} + \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}}[\widehat{\boldsymbol{\Sigma}}_j^{1/2} - \boldsymbol{\Sigma}_j^{1/2}]\mathbf{N}_{K_j}.$$

Each term on the RHS is a mean-zero Gaussian process conditional on the data. The desired results then follow by Lemma 8.1, Theorem SA-4.2 in Section SA-4, and applying the Gaussian maximal inequality to each term as in Section 8.3.

8.6. *Proof of Theorem 6.4.* By Theorem 6.1 or Theorem 6.2, there exists a sequence of constants η_n such that $\eta_n = o(1)$ and $\mathbb{P}(|\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| - \sup_{\mathbf{x}\in\mathcal{X}}|Z_j(\mathbf{x})|| > \eta_n/r_n) = o(1)$. Therefore, for any $u \in \mathbb{R}$,

$$\begin{aligned}&\mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| \le u\Big] \\ &\quad\le \mathbb{P}\Big[\Big\{\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| \le u\Big\} \cap \Big\{\Big|\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| - \sup_{\mathbf{x}\in\mathcal{X}}|Z_j(\mathbf{x})|\Big| \le \eta_n/r_n\Big\}\Big] \\ &\qquad + \mathbb{P}\Big[\Big\{\Big|\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| - \sup_{\mathbf{x}\in\mathcal{X}}|Z_j(\mathbf{x})|\Big| > \eta_n/r_n\Big\}\Big] \\ &\quad\le \mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|Z_j(\mathbf{x})| \le u + \eta_n/r_n\Big] + o(1) \\ &\quad\le \mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|Z_j(\mathbf{x})| \le u\Big] + Cr_n^{-1}\eta_n\mathbb{E}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|Z_j(\mathbf{x})|\Big] + o(1)\end{aligned}$$

for some constant $C > 0$, where the last inequality holds by the anticoncentration inequality due to [19]. By the Gaussian maximal inequality, $\mathbb{E}[\sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})|] \lesssim \sqrt{\log n}$. Since $r_n = \sqrt{\log n}$, the second term on the right of the last inequality is $o(1)$. The reverse of the inequality follows similarly, and thus

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})| \le u\Big] - \mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}|Z_j(\mathbf{x})| \le u\Big]\right| = o(1).$$

In addition, by Theorem 6.3, $\widehat{Z}_j(\cdot)$ is approximated by the same Gaussian process conditional on the data. The result then follows by the same argument.

9. Conclusion. We presented new asymptotic results for partitioning-based least squares regression estimators. The first main contribution gave a general IMSE expansion for the point estimators. The second set of contributions were pointwise and uniform distributional approximations, with and without robust bias correction, for t-statistic processes indexed by $\mathbf{x} \in \mathcal{X}$, with improvements in rate restrictions and convergence rates. For the case of $d = 1$, our uniform approximation results rely on a new coupling approach, which delivered seemingly minimal rate restrictions. Furthermore, we apply our general results to three popular special cases: B-splines, compactly supported wavelets and piecewise polynomials. Finally, we provide a general purpose R package `lspartition` [14].

Acknowledgments. We thank Victor Chernozhukov, Denis Chetverikov, Michael Jansson, Xinwei Ma and Whitney Newey for useful discussions. We also thank the co-editors, Edward George and Ming Yuan, an associate editor and the reviewers for thoughtful comments that significantly improved this paper. See [15] for supplementary materials.

The first author was support by National Science Foundation (SES 1459931).

SUPPLEMENTARY MATERIAL

Additional technical results, omitted proofs, implementation details and further simulation results (DOI: 10.1214/19-AOS1865SUPP; .pdf). The SA gives omitted proofs and additional technical results that may be of independent interest, including pointwise and uniform stochastic linearization useful in semiparametric settings (Section SA-4; see, in particular, Remark SA-4.1), theoretical comparisons between bias correction approaches and a discussion of the relationship between B-Splines and polynomials. Details on implementation, specific examples, and further simulation evidence are also reported.

REFERENCES

[1] AGARWAL, G. G. and STUDDEN, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8** 1307–1325. MR0594647

[2] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and FERNANDEZ-VAL, I. (2019). Conditional quantile processes based on series or many regressors. *J. Econometrics*. To appear.

[3] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

[4] BHATIA, R. (2013). *Matrix Analysis*. Springer, New York. MR1477662 https://doi.org/10.1007/978-1-4612-0653-8

[5] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[6] CALONICO, S., CATTANEO, M. D. and FARRELL, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *J. Amer. Statist. Assoc.* **113** 767–779. MR3832225 https://doi.org/10.1080/01621459.2017.1285776

[7] CALONICO, S., CATTANEO, M. D. and FARRELL, M. H. (2019). Coverage error optimal confidence intervals for local polynomial regression. arXiv:1808.01398.
[8] CALONICO, S., CATTANEO, M. D. and TITIUNIK, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82** 2295–2326. MR3301169 https://doi.org/10.3982/ECTA11757
[9] CALONICO, S., CATTANEO, M. D. and TITIUNIK, R. (2015). Optimal data-driven regression discontinuity plots. *J. Amer. Statist. Assoc.* **110** 1753–1769. MR3449071 https://doi.org/10.1080/01621459.2015.1017578
[10] CATTANEO, M. D., CRUMP, R. K., FARRELL, M. H. and FENG, Y. (2019). On binscatter. arXiv:1902.09608.
[11] CATTANEO, M. D., CRUMP, R. K., FARRELL, M. H. and SCHAUMBURG, E. (2019). Characteristic-sorted portfolios: Estimation and inference. *Rev. Econ. Stat.* To appear.
[12] CATTANEO, M. D. and FARRELL, M. H. (2011). Efficient estimation of the dose-response function under ignorability using subclassification on the covariates. In *Missing Data Methods: Cross-Sectional Methods and Applications*. *Adv. Econom.* **27** 93–127. Emerald Group Publ, Ltd., Bingley. MR3495703
[13] CATTANEO, M. D. and FARRELL, M. H. (2013). Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators. *J. Econometrics* **174** 127–143. MR3045024 https://doi.org/10.1016/j.jeconom.2013.02.002
[14] CATTANEO, M. D., FARRELL, M. H. and FENG, Y. (2019). `lspartition`: Partitioning-based least squares regression. *R J.* To appear.
[15] CATTANEO, M. D., FARRELL, M. H. and FENG, Y. (2020). Supplement to "Large sample properties of partitioning-based series estimators." https://doi.org/10.1214/19-AOS1865SUPP.
[16] CHEN, X. and CHRISTENSEN, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *J. Econometrics* **188** 447–465. MR3383220 https://doi.org/10.1016/j.jeconom.2015.03.010
[17] CHEN, X. and CHRISTENSEN, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quant. Econ.* **9** 39–84. MR3789729 https://doi.org/10.3982/QE722
[18] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461 https://doi.org/10.1214/14-AOS1230
[19] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468 https://doi.org/10.1214/14-AOS1235
[20] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields* **162** 47–70. MR3350040 https://doi.org/10.1007/s00440-014-0565-9
[21] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stochastic Process. Appl.* **126** 3632–3651. MR3565470 https://doi.org/10.1016/j.spa.2016.04.009
[22] CHERNOZHUKOV, V., LEE, S. and ROSEN, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica* **81** 667–737. MR3043345 https://doi.org/10.3982/ECTA8718
[23] DAVYDOV, O. (2001). Stable local bases for multivariate spline spaces. *J. Approx. Theory* **111** 267–297. MR1849550 https://doi.org/10.1006/jath.2001.3577
[24] EGGERMONT, P. P. B. and LARICCIA, V. N. (2009). *Maximum Penalized Likelihood Estimation. Volume II: Regression*. *Springer Series in Statistics*. Springer, Dordrecht. MR2817245 https://doi.org/10.1007/b12285
[25] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587
[26] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848
[27] HUANG, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26** 242–272. MR1611780 https://doi.org/10.1214/aos/1030563984
[28] HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600–1635. MR2012827 https://doi.org/10.1214/aos/1065705120
[29] JURINSKIĬ, V. V. (1978). On the error of the Gaussian approximation for convolutions. *Theory Probab. Appl.* **22** 236–247. MR0517490
[30] KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131. MR0375412 https://doi.org/10.1007/BF00533093

[31] KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrsch. Verw. Gebiete* **34** 33–58. MR0402883 https://doi.org/10.1007/BF00532688

[32] NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147–168. MR1457700 https://doi.org/10.1016/S0304-4076(97)00011-0

[33] NOBEL, A. (1996). Histogram regression estimation using data-dependent partitions. *Ann. Statist.* **24** 1084–1105. MR1401839 https://doi.org/10.1214/aos/1032526958

[34] RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2009). *Semiparametric Regression*. Cambridge Univ. Press, New York. MR1998720 https://doi.org/10.1017/CBO9780511755453

[35] SAKHANENKO, A. I. (1985). Convergence rate in the invariance principle for non-identically distributed variables with exponential moments. In *Advances in Probability Theory*: *Limit Theorems for Sums of Random Variables* 2–73.

[36] SAKHANENKO, A. I. (1991). On the accuracy of normal approximation in the invariance principle [translation of Trudy Inst. Mat. (Novosibirsk) **13** (1989), Asimptot. Analiz Raspred. Sluch. Protsess., 40–66; MR1037248 (91d:60082)] **1** 58–91. Siberian Adv. Math., 4, Siberian Advances in Mathematics. MR1138005

[37] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[38] TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

[39] TUKEY, J. W. (1961). Curves as parameters, and touch estimation. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 681–694. Univ. California Press, Berkeley, CA. MR0132677

[40] ZAĬTSEV, A. YU. (2013). The accuracy of strong Gaussian approximation for sums of independent random vectors. *Russian Math. Surveys* **68** 721–761. MR3154816 https://doi.org/10.1070/rm2013v068n04abeh004851

[41] ZHAI, A. (2018). A high-dimensional CLT in W2 distance with near optimal convergence rate. *Probab. Theory Relat. Fields* **3-4** 821–845.

[42] ZHANG, H. and SINGER, B. H. (2010). *Recursive Partitioning and Applications*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2674991 https://doi.org/10.1007/978-1-4419-6824-1

[43] ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277 https://doi.org/10.1214/aos/1024691356

[44] ZHOU, S. and WOLFE, D. A. (2000). On derivative estimation in spline regression. *Statist. Sinica* **10** 93–108. MR1742102