LARGE SAMPLE PROPERTIES OF PARTITIONING-BASED SERIES ESTIMATORS

BY MATIAS D. CATTANEO*, MAX H. FARRELL AND YINGJIE FENG

Princeton University, University of Chicago, and Princeton University

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 develop 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 popular 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 \( \{(y_i, x_i'), i = 1, \ldots, n\} \) is a random sample from the model

\[
y_i = \mu(x_i) + \varepsilon_i, \quad E[\varepsilon_i|x_i] = 0, \quad E[\varepsilon_i^2|x_i] = \sigma^2(x_i),
\]

for a scalar response \( y_i \) and a \( d \)-vector of continuously distributed covariates \( x_i = (x_{i1}, \ldots, x_{id})' \) 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 non-overlapping cells and these are 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

*Financial support from the National Science Foundation (SES 1459931) is gratefully acknowledged.

MSC 2010 subject classifications: Primary 62H10, 62M99, 57R12; secondary 62M99

Keywords and phrases: nonparametric regression, series methods, sieve methods, robust bias correction, uniform inference, strong approximation, tuning parameter selection
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 $\Delta = \{\delta_l \subset \mathcal{X} : 1 \leq l \leq \bar{k}\}$ be a collection of $\bar{k}$ open and disjoint sets, the closure of whose union is $\mathcal{X}$ (or, more generally, covers $\mathcal{X}$). We restrict $\delta_l$ 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, each of order $m$ (e.g., $m = 4$ for cubic splines) is denoted by $x_i \mapsto \mathbf{p}(x_i) := \mathbf{p}(x_i; \Delta, m) = (p_1(x_i; \Delta, m), \ldots, p_K(x_i; \Delta, m))'$. For $x \in \mathcal{X}$ and $q = (q_1, \ldots, q_d)' \in \mathbb{Z}_+^d$, the partial derivative $\partial^q \mu(x)$ is estimated by least squares regression

\begin{equation}
\begin{aligned}
\hat{\partial}^q \mu(x) &= \partial^q \mathbf{p}(x)' \hat{\beta}, \\
\hat{\beta} &\in \arg \min_{\mathbf{b} \in \mathbb{R}^K} \sum_{i=1}^n (y_i - \mathbf{p}(x_i)' \mathbf{b})^2,
\end{aligned}
\end{equation}

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

The approximation power of this class of estimators comes from two user-specified parameters: the granularity of the partition $\Delta$ and the order $m \in \mathbb{Z}_+$ of the basis. The choice $m$ is often fixed in practice, and hence we regard $\Delta$ as the tuning parameter. Under our assumptions, $\bar{k} \to \infty$ as the sample size $n \to \infty$, and the volume of each $\delta_l$ shrinks proportionally to $h^d$, where $h = \max \{\text{diam} (\delta) : \delta \in \Delta\}$ serves as a universal measure of the granularity. Thus, as $\bar{k} \to \infty$, $h^d$ 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 $h_{\text{IMSE}} \asymp n^{-1/(2m+d)}$, where

which translates to the familiar \( K_{\text{IMSE}} \propto n^{-d/(2m+d)} \), and give a precise characterization of the leading constant. For simple cases on tensor-product partitions, some results exist for splines \([1, 41, 42]\) 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 \( x \in \mathcal{X} \)) inference procedures. Undersmoothing (employing a finer partition 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 \( \hat{\partial} \mu(x) \) 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 \( x \), 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 \( x \in \mathcal{X} \) 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 \( x \in \mathcal{X} \), 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 \( \hat{\kappa} \)), 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) gives complete proofs, several new technical and methodological results, further Monte Carlo evidence, and applies our general results for splines, wavelets, and piecewise polynomials. A companion \( \text{R} \) 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., 23, 24, 31, and references therein]. Of particular relevance are treatments of series (linear sieve) methods in general, and while some results concerning partitioning-based estimators exist, they are mainly limited to splines, wavelets, or piecewise polynomials, considered separately [29, 25, 41, 26, 13, 15, 2]. Piecewise polynomial fits on partitions have a long and ongoing tradition in statistics, dating at least to the regressogram of Tukey [36], continuing through [34] (named local polynomial regression therein) and [13], and up to modern, data-driven partitioning techniques such as regression trees [5], trend filtering [35], and related methods [40]. Partitioning-based methods have also featured as inputs or preprocessing such as regression trees [5], trend filtering [35], and related methods [40].

Literature on nonparametric regression, summarized in many textbook treatments, nonparametric regression and strong approximation. There is a vast literature on nonparametric regression, summarized in many textbook treatments [e.g., 23, 24, 31, and references therein]. Of particular relevance are treatments of series (linear sieve) methods in general, and while some results concerning partitioning-based estimators exist, they are mainly limited to splines, wavelets, or piecewise polynomials, considered separately [29, 25, 41, 26, 13, 15, 2]. Piecewise polynomial fits on partitions have a long and ongoing tradition in statistics, dating at least to the regressogram of Tukey [36], continuing through [34] (named local polynomial regression therein) and [13], and up to modern, data-driven partitioning techniques such as regression trees [5], trend filtering [35], and related methods [40]. Partitioning-based methods have also featured as inputs or preprocessing such as regression trees [5], trend filtering [35], and related methods [40].

Therein) and [13], and up to modern, data-driven partitioning techniques such as regression trees [5], trend filtering [35], and related methods [40]. Partitioning-based methods have also featured as inputs or preprocessing such as regression trees [5], trend filtering [35], and related methods [40].

Therein) and [13], and up to modern, data-driven partitioning techniques such as regression trees [5], trend filtering [35], and related methods [40]. Partitioning-based methods have also featured as inputs or preprocessing such as regression trees [5], trend filtering [35], and related methods [40].

1.2. Notation. For a $d$-tuple $q = (q_1, \ldots, q_d) \in \mathbb{Z}^d$, define $[q] = \sum_{j=1}^d q_j$, $x^q = x_1^{q_1} x_2^{q_2} \cdots x_d^{q_d}$ and $\partial^q \mu(x) = \partial^{q_1} \mu(x)/\partial x_1^{q_1} \cdots \partial x_d^{q_d}$. Unless explicitly stated otherwise, whenever $x$ is a boundary point of some closed set, the partial derivative is understood as the limit with $x$ ranging within it. Let $0 = (0, \cdots, 0)'$ be the length-$d$ zero vector. We set $\mu(x) := \partial^0 \mu(x)$ and $\mu_j(x) := \partial^0 \mu_j(x)$ for $j = 0, 1, 2, 3$ and collect the covariates as $X = [x_1, \ldots, x_n]'$. 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 $v = (v_1, \ldots, v_M) \in \mathbb{R}^M$, we write $\|v\| = (\sum_{i=1}^M v_i^2)^{1/2}$ and $\dim(v) = M$. For a matrix $A \in \mathbb{R}^{M \times N}$, $\|A\| = \max_i \sigma_i(A)$ and $\|A\|_\infty = \max_{1 \leq i \leq M} \sum_{j=1}^N |a_{ij}|$ for operator norms induced by $L_2$ and $L_\infty$ norms, where $\sigma_i(A)$ is the $i$-th singular value of $A$, and $\lambda_{\min}(A)$ is the minimum eigenvalue of $A$.

We use empirical process notation: $E_n[g(x_i)] = \frac{1}{n} \sum_{i=1}^n g(x_i)$ and $G_n[g(x_i)] = \frac{1}{\sqrt{n}} \sum_{i=1}^n (g(x_i) - E[g(x_i)])$. For sequences we let $a_n \lesssim b_n$ denote that $\limsup_n |a_n/b_n|$ is finite, $a_n = O_P(b_n)$ denote $\limsup_{n \to \infty} \limsup_n P[|a_n/b_n| \geq \epsilon] = 0$, $a_n = o_P(b_n)$ denote $a_n/b_n \to 0$, $a_n = o_P(b_n)$ denote $a_n/b_n \to 0$, where $\rightarrow$ is convergence in probability, and $a_n \lesssim b_n$ denote $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Limits are...
taken as \( n \to \infty \) \((h \to 0, K \to \infty, \text{as appropriate})\), unless otherwise stated.

Finally, throughout the paper, \( r_n > 0 \) denotes a non-vanishing 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) \( \{ (y_i, x'_i) : 1 \leq i \leq n \} \) are i.i.d. satisfying (1.1), where \( x_i \) has compact connected support \( \mathcal{X} \subset \mathbb{R}^d \) and an absolutely continuous distribution function. The density of \( x_i \), \( f(\cdot) \), and the conditional variance of \( y_i \) given \( x_i \), \( \sigma^2(\cdot) \), are bounded away from zero and continuous.

(b) \( \mu(\cdot) \) is \( S \)-times continuously differentiable, for \( S > \lceil q \rceil \), and all \( \partial^s \mu(\cdot), [s] = S \), 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

\[
\frac{\max \{ \text{diam}(\delta) : \delta \in \Delta \} }{ \min \{ \text{diam}(\delta) : \delta \in \Delta \} } \lesssim 1,
\]

where \( \text{diam}(\delta) \) denotes the diameter of \( \delta \).

This condition implies that the size of each \( \delta \in \Delta \) can be well characterized by the diameter of \( \delta \) and that we can use \( h = \max \{ \text{diam}(\delta) : \delta \in \Delta \} \) as a universal measure of mesh sizes of elements in \( \Delta \). 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 \( x \in \mathcal{X} \), with appropriately chosen knot positions. The SA (§SA-3) gives details and discusses this special example of partitioning scheme. If \( \Delta \) covers only 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 non-tensor-product partitioning schemes. We defer studying general data-dependent partitioning for future work, but note that a few specific results are available [5, 30, 9].

The second assumption on the partitioning-based estimators employs generalized notions of stable local basis [21] and active basis [26]. We say a function \( p(\cdot) \) on \( X \) is active on \( \delta \in \Delta \) if it is not identically zero on \( \delta \).

**Assumption 3 (Local Basis).**

(a) For each basis function \( p_k, k = 1, \ldots, K \), the union of elements of \( \Delta \) on which \( p_k \) is active is a connected set, denoted by \( \mathcal{H}_k \). For all \( k = 1, \ldots, K \), 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 \( a = (a_1, \cdots, a_K) \in \mathbb{R}^K \),

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

(c) For an integer \( \varsigma \in ([q], m) \), for all \( \varsigma, [\varsigma] \leq \varsigma \),

\[
  h^{-[\varsigma]} \lesssim \inf_{\delta \in \Delta} \inf_{x \in \text{clo}(\delta)} \| \partial^{[\varsigma]} p(x; \Delta, m) \| \leq \sup_{\delta \in \Delta} \sup_{x \in \text{clo}(\delta)} \| \partial^{[\varsigma]} p(x; \Delta, m) \| \lesssim h^{-[\varsigma]}
\]

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

\[
  \sup_{\delta \in \Delta} \sup_{x \in \text{clo}(\delta)} \| \partial^{[\varsigma]} p(x; \Delta, m) \| \lesssim h^{-[\varsigma]-1}.
\]

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 \( p(x; \Delta, m) \) be supported by a region consisting of a finite number of cells in \( \Delta \). Therefore, as \( \bar{\kappa} \to \infty \) (and \( h \to 0 \)), each element of \( \Delta \) 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 co-linear: the minimum eigenvalue of \( \mathbb{E}[p(x_i)p(x_i)'] \) 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 co-linear locally, as stated in Assumption 3(b). These two assumptions are very similar to Conditions A.2 and Conditions A.3 in the Appendix of [26], 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: $K \asymp \bar{K} \asymp h^{-d}$. 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 $x \in \mathcal{X}$, let $\delta_x$ be the element of $\Delta$ whose closure contains $x$ and $h_x$ for the diameter of this $\delta_x$.

**Assumption 4 (Approximation Error).** For all $\varsigma$ satisfying $[\varsigma] \leq \varsigma$ in Assumption 3, there exists $s^* \in S_{\Delta, m}$, the linear span of $p(x; \Delta, m)$, and
\[
\mathcal{B}_{m, \varsigma}(x) = - \sum_{u \in \Lambda_m} \partial^u \mu(x) h_x^{m-[\varsigma]} B_{u, \varsigma}(x)
\]
such that
\[
(2.1) \quad \sup_{x \in \mathcal{X}} |\partial^q \mu(x) - \partial^q s^*(x) + \mathcal{B}_{m, \varsigma}(x)| \lesssim h^{m+q-[\varsigma]}
\]
and
\[
(2.2) \quad \sup_{\delta \in \Delta \mathcal{X}_1, x_2 \in \text{clo}(\delta)} \frac{|B_{u, \varsigma}(x_1) - B_{u, \varsigma}(x_2)|}{\|x_1 - x_2\|} \lesssim \frac{1}{h^{-1}}
\]
where $B_{u, \varsigma}(\cdot)$ is a known function that is bounded uniformly over $n$, and $\Lambda_m$ is a multi-index set, which depends on the basis, with $[u] = m$ for $u \in \Lambda_m$.

The usual rate-only assumption, $\sup_{x \in \mathcal{X}} |\partial^q \mu(x) - \partial^q s^*(x)| \lesssim h^{m-[q]}$, which is implied by Assumptions 4, will not suffice for our bias correction and IMSE expansion results: (2.1) is needed. The terms $B_{u, \varsigma}(x)$ in $\mathcal{B}_{m, \varsigma}(x)$ are known functions of $x$ which depend on the particular partitioning scheme and bases used. The only unknowns are the higher-order derivatives of $\mu(\cdot)$. In the SA (§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 $\Lambda_m$. We assume enough smoothness exists to characterize these terms: see [6] for a discussion when smoothness constrains inference.

The function $\mathcal{B}_{m, \varsigma}$ is understood as the approximation error in $L_\infty$ 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: $\mathcal{B}_{m, \varsigma}$ and the accompanying error from the linear projection of $\mathcal{B}_{m, 0}$ onto $S_{\Delta, m}$. We formalize this result in Lemma 3.1 below. The second bias term is often ignored because in several
cases the leading approximation error $\mathcal{B}_{m,0}$ is *approximately orthogonal* to $p$ with respect to the Lebesgue measure, that is,

\begin{equation}
\max_{1 \leq k \leq K} \int_{\mathcal{H}_k} p_k(x; \Delta, m) \mathcal{B}_{m,0}(x) \, dx = o(h^{m+d}),
\end{equation}

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 $S_{\Delta,m}$. When (2.3) holds, the leading term in $L_\infty$ 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 $B_{u,0}$ and $p$ in $L_2$ with respect to the Lebesgue measure, for all $u \in \Lambda_m$.

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 $\mathcal{B}_{m,0}$ onto the approximating space spanned by $p$.

3. Characterization and Correction of Bias. We now precisely characterize the bias of $\hat{\partial_p \mu}(x)$ 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 $\hat{\partial_p \mu}(x)$ of (1.2) can be written as

\begin{equation}
\hat{\partial_p \mu}_0(x) := \tilde{\gamma}_{q,0}(x)'E_n[\Pi_0(x_i)|y_i],
\end{equation}

where

\[ \tilde{\gamma}_{q,0}(x)' := \partial^q p(x)'E_n[p(x_i)p(x_i)]^{-1} \quad \text{and} \quad \Pi_0(x_i) := p(x_i). \]

The subscript “0” differentiates this estimator from the bias-corrected versions below. We give our first result, proven in the SA, §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{equation}
E[\hat{\partial_p \mu}_0(x)|X] - \partial^q \mu(x)
\end{equation}

\[ = \tilde{\gamma}_{q,0}(x)'E_n[\Pi_0(x_i)|y_i] - \partial^q \mu(x)
\]

\begin{equation}
= \mathcal{B}_{m,q}(x) - \tilde{\gamma}_{q,0}(x)'E_n[\Pi_0(x_i)|\mathcal{B}_{m,0}(x_i)] + O_p(h^{m+\varepsilon}[q]),
\end{equation}
The proof of this lemma generalizes an idea in [41, Theorem 2.2] to handle partitioning-based series estimators beyond the specific example of $B$-Splines on tensor-product partitions. The first component $B_{m,q}(x)$ 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 $p(x)$ is locally supported (Assumption 3), the orthogonality condition in (2.3), when it holds, suffices to guarantee that the projection of leading error is of smaller order (such as for $B$-splines on a tensor-product partition). In general the bias will be $O(h^{m-[q]})$ 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 $\tilde{p}(x) := \tilde{p}(x; \tilde{\Delta}, \tilde{m})$ 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

$$\hat{\partial q}_{1}(x) := \hat{\gamma}_{q,1}(x) / \mathbb{E}_{n}[\Pi_{1}(x_{i})y_{i}],$$

where

$$\hat{\gamma}_{q,1}(x) := \partial q \tilde{p}(x)^{\prime} / \mathbb{E}_{n}[\tilde{p}(x_{i})\tilde{p}(x_{i})]^{-1} \quad \text{and} \quad \Pi_{1}(x_{i}) := \tilde{p}(x_{i}).$$

This approach can be viewed as a bias correction of the original point estimator because, trivially, $\hat{\partial q}_{1}(x) = \hat{\partial q}_{0}(x) - (\hat{\partial q}_{0}(x) - \hat{\partial q}_{1}(x))$. Valid inference based on $\hat{\partial q}_{1}(x)$ can be viewed as “undersmoothing” applied to the higher-order point estimator, but is distinct from undersmoothing $\hat{\partial q}_{0}(x)$ (i.e., using a finer partition $\Delta$ and keeping the order fixed). [26] 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 $\mu$ and $\partial q \mu$, both of which can be estimated using the higher-order estimator (3.4).

By plugging these into (3.2) and subtracting the result from $\hat{\partial q}_{0}(x)$, we obtain the least squares bias correction, numbered as approach 2:

$$\hat{\partial q}_{2}(x) := \hat{\partial q}_{0}(x) - \left( \hat{\gamma}_{q,0}(x) / \mathbb{E}_{n}[\Pi_{0}(x_{i})\hat{\mu}_{1}(x_{i})] - \hat{\partial q}_{1}(x) \right)$$

$$:= \hat{\gamma}_{q,2}(x) / \mathbb{E}_{n}[\Pi_{2}(x_{i})y_{i}]$$
where
\[
\tilde{\gamma}_{q,2}(x)' := \left( \tilde{\gamma}_{q,0}(x)', -\tilde{\gamma}_{q,0}(x)' \mathbb{E}_n[p(x_i)\tilde{p}(x_i)']\mathbb{E}_n[\tilde{p}(x_i)\tilde{p}(x_i)']^{-1} + \tilde{\gamma}_{q,1}(x)' \right)
\]
and
\[
\Pi_2(x_i) := (p(x_i)', \tilde{p}(x_i)')',
\]
which is exactly of the same form as \( \partial^q \mu_0(x) \) and \( \partial^q \mu_1(x) \) (cf., (3.1) and (3.4)), except for the change in \( \tilde{\gamma}_{q,j}(x) \) and \( \Pi_j(x_i) \).

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 \( \partial^q \mu_0(x) \) according to Assumption 4.

To be precise, we employ the explicit plug-in bias estimator
\[
\hat{\mu}_3(x) = - \sum_{u \in \Lambda_m} \left( \partial^u \mu_1(x) \right) h_x^{m-q} B_{u,q}(x),
\]
with \([q] < m\) and \( \Lambda_m \) as in Assumption 4, leading to
\[
\hat{\partial^q \mu_3}(x) := \hat{\partial^q \mu_0}(x) - \left( \hat{\partial^q \mu_3}(x) - \tilde{\gamma}_{q,0}(x)' \mathbb{E}_n[\Pi_0(x_i)\hat{\partial^q \mu_3}(x_i)] \right)
\]
\[
:= \tilde{\gamma}_{q,3}(x)' \mathbb{E}_n[\Pi_3(x_i)y_i]
\]
where
\[
\tilde{\gamma}_{q,3}(x)' := \left( \tilde{\gamma}_{q,0}(x)', \sum_{u \in \Lambda_m} \left\{ \tilde{\gamma}_{u,1}(x)' h_x^{m-[q]} B_{u,q}(x) \right. \right.
\]
\[
- \left. \tilde{\gamma}_{q,0}(x)' \mathbb{E}_n[p(x_i)h_x^{m}B_{u,0}(x_i)\tilde{\gamma}_{u,1}(x_i)'] \right\},
\]
and
\[
\Pi_3(x_i) := (p(x_i)', \tilde{p}(x_i)')'.
\]

When the orthogonality condition (2.3) holds, the second correction term in \( \hat{\partial^q \mu_3}(x) \) 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 \( \hat{\Delta} \) satisfies Assumption 2, with maximum mesh \( \hat{h} \) and the basis \( \tilde{p}(x; \hat{\Delta}, \hat{m}) \) satisfies Assumptions 3 and 4 with \( \zeta = \hat{\zeta}(\hat{m}) \geq m \) in place of \( \zeta \). Let \( \rho := h/\hat{h} \), which obeys \( \rho \to \rho_0 \in (0, \infty) \). In addition, for \( j = 3 \), either (i) \( \tilde{p}(x; \hat{\Delta}, \hat{m}) \) spans a space containing the span of \( p(x; \Delta, m) \), and for all \( u \in \Lambda_m \), \( \partial^u p(x; \Delta, m) = 0 \); or (ii) both \( p(x; \Delta, m) \) and \( \tilde{p}(x; \hat{\Delta}, \hat{m}) \) reproduce polynomials of degree \([q] \).
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 \(\hat{\partial}q\mu_0\) 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 \(\hat{\partial}q\mu_2\). The natural choice of \(\hat{\Delta} = \Delta\) and \(\hat{m} = m + 1\) will obey this restriction under intuitive conditions. In the SA, Assumption 5 is verified for splines, wavelets, and piecewise polynomials (§SA-6), and we also compare theoretically the alternative bias correction strategies (§SA-7.2).

4. IMSE and Convergence Rates. We establish two main results related to the point estimator \(\hat{\partial}q\mu_0(x)\): 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 \(\Delta\) satisfying Assumption 2.

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

\[
\int_{\mathcal{X}} \mathbb{E}[(\hat{\partial}q\mu_0(x) - \partial q\mu(x))^2 | X] w(x) \, dx
\]

\[
= \frac{1}{n} \left( \mathcal{V}_{\Delta,q} + o_p(h^{-d-2[q]}) \right) + \left( \mathcal{B}_{\Delta,q} + o_p(h^{2m-2[q]}) \right)
\]

where

\(\mathcal{V}_{\Delta,q} = \text{trace} \left( \Sigma_0 \int_{\mathcal{X}} \gamma_{q,0}(x) \gamma_{q,0}(x)^\prime w(x) \, dx \right) \asymp h^{-d-2[q]}\),

\(\mathcal{B}_{\Delta,q} = \int_{\mathcal{X}} \left( \mathcal{B}_{m,q}(x) - \gamma_{q,0}(x)^\prime \mathbb{E}[p(x_i) \mathcal{B}_{m,0}(x_i)] \right)^2 w(x) \, dx \lesssim h^{2m-2[q]}\),

\(\Sigma_0 := \mathbb{E}[\Pi_0(x_i) \Pi_0(x_i)^\prime \sigma^2(x_i)]\), and \(\gamma_{q,0}(x)^\prime := \partial q p(x)^\prime \mathbb{E}[p(x_i)p(x_i)^\prime]^{-1}\).

This theorem, proven in the SA, §SA-10.5, shows that the leading term in the integrated (and pointwise) variance of \(\hat{\partial}q\mu_0(x)\) is of order \(n^{-1}h^{-d-2[q]}\).
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 $\mathcal{Y}_{\Delta,q}$ and $\mathcal{B}_{\Delta,q}$ are nonrandom sequences depending on the partitioning scheme $\Delta$ in a complicated way, and need not converge as $h \to 0$. Nevertheless, when the integrated squared bias does not vanish ($\mathcal{B}_{\Delta,q} \neq 0$), Theorem 4.1 implies that the IMSE-optimal mesh size $h_{\text{IMSE}}$ is proportional to $n^{-1/(2m+d)}$, or equivalently, the IMSE-optimal number of series terms $K_{\text{IMSE}} \propto n^{d/(2m+d)}$. 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, 41, 42] on tensor-product partitions only, and often restricting to $d = 1$ or $|q| = 0$.

We now consider the special case of a tensor-product partition where the “tuning parameter” $\Delta$ reduces to the vector of partitioning knots $\kappa = (\kappa_1, \ldots, \kappa_d)'$, where $\kappa_\ell$ is the number of subintervals used for the $\ell$-th covariate. We assume that $\Delta$ and $p(\cdot)$ 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 $X = \otimes_{\ell=1}^d X_\ell \subset \mathbb{R}^d$, which is normalized to $[0,1]^d$ without loss of generality, and $\Delta$ is a tensor-product partition. For $x \in [0,1]^d$, denote $\delta_x = \{t_{\ell,x_\ell} \leq x_\ell \leq t_{\ell,x_\ell+1}, 1 \leq \ell \leq d\}$, where $b_x < \kappa_\ell$ (see SA, §8A-3 for details). Let $b_x = (b_{x,1}, \ldots, b_{x,d})$ collect the interval lengths $b_{x,\ell} = |t_{\ell,x_{\ell+1}} - t_{\ell,x_\ell}|$. In addition:

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

(b) For all $\delta \in \Delta$ and $u_1, u_2 \in \Lambda_m$, there exist constants $\eta_{u_1,u_2,q}$ such that

$$\int_{\delta} \frac{h_X^{2m-2|q|}}{b_x^{u_1+u_2-2q}} B_{u_1,q}(x)B_{u_2,q}(x) \, dx = \eta_{u_1,u_2,q} \, \text{vol}(\delta)$$

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

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

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 \( B_{u,q}(\cdot) \) for \( u \in \Lambda_m \), which depend on the basis but not \( \mu(x) \), 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 \( b_x \)) 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 \( p \) can be divided into \( J + \tilde{J} \) 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 \( \Delta \) can be associated with one basis function in \( p \) 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 \( J = \left( \frac{d + m - 1}{m - 1} \right) \). This condition is used to ensure that the summation over the number of basis functions converges to a well-defined integral as \( K \approx h^{-d} \to \infty \).

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

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

\[
\mathcal{V}_{\kappa,0} = \left( \prod_{\ell = 1}^{d} \kappa_{\ell} \right) \mathcal{V}_0 + o(h^{-d}), \quad \mathcal{V}_0 = J \int_{[0,1]^d} \sigma^2(x) \left( \prod_{\ell = 1}^{d} g_{\ell}(x) \right) w(x) \, dx,
\]

and, provided that \( (2.3) \) holds,

\[
\mathcal{B}_{\kappa,0} = \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1 + u_2)} \mathcal{B}_{u_1, u_2,0} + o(h^{2m}),
\]

\[
\mathcal{B}_{u_1, u_2,0} = \eta_{u_1, u_2,0} \int_{[0,1]^d} \frac{\partial^{u_1} \mu(x) \partial^{u_2} \mu(x)}{g(x)^{u_1 + u_2}} w(x) \, dx.
\]

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 \( \mathcal{V}_{\kappa,q} \) and \( \mathcal{B}_{\kappa,q} \) converge to a
well-defined limit when general basis functions are considered. [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 $q = 0$ and the approximate orthogonality condition (2.3). The SA ($\S$SA-3) contains more details and other results.

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

$$
\kappa_{\text{IMSE},0} = \arg\min_{\kappa \in \mathbb{Z}_{+}^d} \left\{ \frac{1}{n} \left( \prod_{\ell=1}^{d} \kappa_{\ell} \right) \mathcal{Y}_0 + \sum_{u_1,u_2 \in \Lambda_m} \kappa_{-(u_1+u_2)} \mathcal{B}_{u_1,u_2,0} \right\},
$$

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

$$
\kappa_{\text{IMSE},0} = \left\lceil \left( \frac{2m}{d} \sum_{u_1,u_2 \in \Lambda_m} \mathcal{B}_{u_1,u_2,0} \right)^{\frac{1}{m+d}} \frac{1}{n^{\frac{1}{2m+d}}} \right\rceil,
$$

Data-driven versions of this IMSE-optimal choice, and extensions to derivative estimation, are discussed in the SA ($\S$SA-8) and fully implemented in our companion general-purpose R package lsplit [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 $\hat{\partial^q \mu}_0(x)$. 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 $\sup_{x \in \mathcal{X}} |\partial^q \mu(x) - \partial^q s^*(x)| \lesssim h^{m-|q|}$ with $s^*$ defined in Assumption 4. Then, if $\frac{\log n}{nh^n} = o(1)$,

$$
\int_{\mathcal{X}} \left( \frac{\hat{\partial^q \mu}_0(x) - \partial^q \mu(x)}{w(x)} \right)^2 w(x) dx \lesssim \frac{1}{nh^{d+2|q|}} + h^{2(m-|q|)}
$$

If, in addition,

(i) $\sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] < \infty$ and $\frac{(\log n)^3}{nh^n} \lesssim 1$, or

(ii) $\sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ and $\frac{n^{\frac{2\nu}{\nu+2}} (\log n)^{\frac{2\nu}{\nu+2}}}{nh^n} \lesssim 1$,

then

$$
\sup_{x \in \mathcal{X}} \left| \frac{\hat{\partial^q \mu}_0(x) - \partial^q \mu(x)}{w(x)} \right|^2 \lesssim \frac{\log n}{nh^{d+2|q|}} + h^{2(m-|q|)}.
$$
This theorem, proven in the SA, §SA-10.11, shows that the partitioning-based estimators can attain the optimal \( L_2 \) and uniform convergence rate [34] by proper choice of partitioning scheme, under our high-level assumptions. (The full force of Assumption 4 is not needed for this result.) [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, 15].

5. Pointwise Inference. We give pointwise inference based on classical undersmoothing and all three bias correction methods. All four point estimators take the form

\[
\hat{\partial^q \mu_j(x)} = \hat{\gamma}_{q,j}(x)'E_n[\Pi_j(x_i)y_i],
\]

where, for each \( j = 0, 1, 2, 3 \), \( \gamma_{q,j}(x) \) are defined as \( \hat{\gamma}_{q,j} \) in (3.1), (3.4), (3.5), and (3.6), respectively, but with sample averages and other estimators replaced by their population counterparts, and \( \Sigma_j := E[\Pi_j(x_i)\Pi_j(x_i)']\sigma^2(x_i) \). 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 [8, 6]. We also discuss below Studentization, that is, replacing \( \Omega_j(x) \) with a consistent estimator.

5.1. Distributional Approximation. Our first result establishes the limiting distribution of the standardized t-statistics \( T_j(x) \).

**Theorem 5.1 (Asymptotic Normality).** Let Assumptions 1, 2, 3, and 4 hold. Assume \( \sup_{x \in \mathcal{X}} E[\varepsilon_i^2] \{ |\varepsilon_i| > M \} |x_i = x| \to 0 \) as \( M \to \infty \), and \( \frac{\log n}{nh} = o(1) \). 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 \( x \in \mathcal{X} \), \( \sup_{u} \left| P(T_j(x) \leq u) - \Phi(u) \right| = o(1) \), where \( \Phi(u) \) denotes the cumulative distribution function of \( N(0,1) \).

This theorem, proven in the SA, §SA-10.9, gives a valid Gaussian approximation for the t-statistics \( T_j(x) \), pointwise in \( x \in \mathcal{X} \). 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 \( \Omega_j(x) \) with a consistent estimator. Specifically, we consider the four feasible \( t \)-statistics, \( j = 0, 1, 2, 3 \),

\[
\hat{T}_j(x) = \frac{\hat{\partial} q \mu_j(x) - \partial q \mu(x)}{\sqrt{\hat{\Omega}_j(x)/n}}, \quad \hat{\Omega}_j(x) = \hat{\gamma}_{q,j}(x)' \hat{\Sigma}_j \hat{\gamma}_{q,j}(x),
\]

(5.1)
\[
\hat{\Sigma}_j = \mathbb{E}_n[\Pi_j(x_i) \Pi_j(x_i) / \hat{\varepsilon}_{i,j}^2], \quad \hat{\varepsilon}_{i,j} = y_i - \hat{\mu}_j(x_i),
\]

Once the basis functions and partitioning schemes are chosen, the statistic \( \hat{T}_j(x) \) 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) \( \sup_{x \in X} \mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty \) and \( \frac{n^{2+\nu} (\log n)^{2+\nu}}{nh^d} = o(1) \), or

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

Then, for each \( j = 0, 1, 2, 3 \),

\[
|\hat{\Omega}_j(x) - \Omega_j(x)| = o_P(h^{-d-2}[q]).
\]

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

\[
\left[ \hat{\partial} q \mu_j(x) \pm \Phi^{-1}(1-\alpha/2) \cdot \sqrt{\hat{\Omega}_j(x)/n} \right], \quad j = 0, 1, 2, 3.
\]

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 [41, 42, 26].

6. Uniform Inference. We next give a valid distributional approximation for the whole process \( \{\hat{T}_j(x) : x \in X\} \), 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 \( \partial q \mu(\cdot) \).
6.1. **Strong Approximations.** The stochastic processes \( \{\hat{T}_j(x) : x \in \mathcal{X}\} \) 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 non-vanishing 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) \( \sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^2 + \nu |x = x] < \infty \) and \( \frac{n^2 \log(n)}{n h^d} \cdot \frac{2 + \nu}{2 + 2 \nu} = o(r_n^{-2}) \); or

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

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

\[
\sup_{x \in \mathcal{X}} \left| \hat{T}_j(x) - t_j(x) \right| = o_P(r_n^{-1}), \quad t_j(x) = \frac{\gamma_{q,j}(x)'}{\sqrt{\Omega_j(x)}} \mathcal{G}_n[\Pi_j(x_i)\varepsilon_i].
\]

Lemma 6.1 requires that the estimation and sampling uncertainty of \( \hat{\gamma}_{q,j} \) and \( \hat{\Omega}_j(x) \), as well as the smoothing bias of \( \hat{\partial} \mu_j(x) \), be negligible uniformly over \( x \in \mathcal{X} \). The proof is given in §8.2 and relies on a new technical lemma stated in §8.1. This technical approximation step allows us to focus on developing a distributional approximation for the infeasible stochastic processes \( \{t_j(x) : x \in \mathcal{X}\} \), \( j = 0, 1, 2, 3 \). 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 \( \{t_j(x), x \in \mathcal{X}\} \) is approximated by that of a Gaussian process \( \{Z_j(x), x \in \mathcal{X}\} \) 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 \( \mathcal{N}_{K_j} \sim \mathcal{N}(0, \mathbf{I}_{K_j}) \) with \( K_j = \text{dim}(\Pi_j(x)) \) such that

\[
\sup_{x \in \mathcal{X}} \left| t'_j(x) - Z_j(x) \right| = o_P(r_n^{-1}), \quad Z_j(x) = \frac{\gamma_{q,j}(x) \Sigma_j^{1/2}}{\sqrt{\Omega_j(x)}} \mathcal{N}_{K_j}.
\]

This approximation is denoted by \( t_j(\cdot) =_d Z_j(\cdot) + o_P(r_n^{-1}) \) 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 [27, 28]. For \( d > 1 \), we apply an improved version of the classical Yurinskii construction [37].

6.1.1. Unidimensional Regressor. Let \( d = 1 \). The following theorem gives a valid distributional approximation for \( \{ \hat{T}_j(x) : x \in \mathcal{X} \} \) 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 \( \frac{\log n}{\sqrt{nh}} = o(r_n^{-2}) \). Then, for each \( j = 0, 1, 2, 3 \), \( t_j(\cdot) =_d Z_j(\cdot) + o_P(r_n^{-1}) \) in \( L^\infty(\mathcal{X}) \), where \( Z_j(\cdot) \) is given in Definition 6.1.

The proof of this result, in §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 \( \{ \zeta_i : 1 \leq i \leq n \} \) of standard Normal random variables, such that

\[
\sup_{x \in \mathcal{X}} \left| t'_j(x) - z_j(x) \right| = o_P(r_n^{-1}), \quad z_j(x) = \frac{\gamma q_j(x)'}{\sqrt{\Omega_j(x)}} G_n[\Pi_j(x_i) \sigma(x_i) \zeta_i].
\]

**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 \( N_{K_j} \) from Definition 6.1 such that \( z'_j(\cdot) =_d Z_j(\cdot) \) conditional on \( X \), where

\[
Z_j(x) = \frac{\gamma q_j(x)' \Sigma_j^{1/2}}{\sqrt{\Omega_j(x)}} N_{K_j}, \quad \Sigma_j := \mathbb{E}_n[\Pi_j(x_i) \Pi_j(x_i)' \sigma^2(x_i)],
\]

and

\[
\sup_{x \in \mathcal{X}} \left| \tilde{Z}_j(x) - Z_j(x) \right| = o_P(r_n^{-1}).
\]

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 \( X \) 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 non-identically distributed random variables [32, 33]. 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 [22, 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 §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 §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 \( \|\tilde{\Sigma}_j^{1/2} - \Sigma_j^{1/2}\| \), \( j = 0, 1, 2, 3 \). Both \( \tilde{\Sigma}_j \) and \( \Sigma_j \) are symmetric and positive semi-definite. Furthermore, for \( j = 0, 1 \), \( \lambda_{\min}(\Sigma_j) \gtrsim h^d \) for generic partitioning-based estimators under our assumptions, and therefore we use the bound

\[
\|A_1^{1/2} - A_2^{1/2}\| \leq \lambda_{\min}(A_2)^{-1/2}\|A_1 - A_2\|,
\]

which holds for symmetric positive semi-definite \( A_1 \) and symmetric positive definite \( A_2 \) [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 \( \tilde{\Sigma}_j \) will be singular. To circumvent this problem, we employ the weaker bound [4, Theorem X.1.1]: if \( A_1 \) and \( A_2 \) are symmetric positive semi-definite matrices, then

\[
\|A_1^{1/2} - A_2^{1/2}\| \leq \|A_1 - A_2\|^{1/2}.
\]

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 \((\log n)^{3/2}/\sqrt{n}h = o(r_n^{-2})\), can be dropped in some special cases. For some bases it is possible to find a transformation matrix \(\Upsilon\), with \(\|\Upsilon\|_\infty \lesssim 1\), and a basis \(\tilde{p}\), which obeys Assumption 3, such that \((p(\cdot)', \tilde{p}(\cdot)')' = \Upsilon\tilde{p}(\cdot)\). In other words, the two bases \(p\) and \(\tilde{p}\) can be expressed in terms of another basis \(\tilde{p}\) without linear dependence. Then, a positive lower bound holds for \(\lambda_{\min}(\Sigma_j)\), 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 \(p\) and \(\tilde{p}\), a natural choice of \(\tilde{p}\) is simply a higher-order polynomial basis on the same partition. Since each function in \(p\) and \(\tilde{p}\) is a polynomial on each \(\delta \in \Delta\) and nonzero on a fixed number of cells, the “local representation” condition \(\|\Upsilon\|_\infty \lesssim 1\) automatically holds. See the SA (§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 \(r_n^6K^5/n \to 0\), 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 \(r_n^6K^3/n^{3\nu/(2+\nu)} \to 0\), 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 sub-exponential conditional moment restriction, the restriction 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 \(\{\hat{T}_0(x) : x \in \mathcal{X}\}\) constructed using Haar basis, which is a spline, wavelet, and piecewise polynomial with \(m = 1\). In the SA, §SA-5.1, we establish \(t_0(\cdot) = d Z_0(\cdot) + o_P(r_n^{-1})\) in \(L^\infty(\mathcal{X})\) 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 \( \frac{(\log n)^4}{nh^3} = o(r_n^{-6}) \). Then, for each \( j = 0, 1, 2, 3 \), \( t_j(\cdot) = d \ Z_j(\cdot) + o_P(r_n^{-1}) \) in \( L_\infty(\mathcal{X}) \), where \( Z_j(\cdot) \) is given in Definition 6.1.*

This strong approximation result, proven in §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 \( r_n K^3/n \to 0 \), up to log \( n \) terms. While not optimal when \( \nu > 3 \) (see Remark 6.2 for a counterexample), it improves on the condition \( r_n K^5/n \to 0 \), 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 \( \{Z_j(x) : x \in \mathcal{X}\} \), in order to conduct inference using the results in Theorem 6.1 or Theorem 6.2. In the SA (§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 \( P^*[:]=P[:\mid y, X] \) denote the probability operator conditional on the data. For each \( j = 0, 1, 2, 3 \), the law of the Gaussian process \( \{Z_j(x) : x \in \mathcal{X}\} \) is approximated by a (feasible) Gaussian process \( \{\tilde{Z}_j(x) : x \in \mathcal{X}\} \), with known distribution conditional on the data \( (y, X) \), in \( L_\infty(\mathcal{X}) \), if the following condition holds: on a sufficiently rich probability space there exists a copy \( \tilde{Z}_j'(\cdot) \) of \( \tilde{Z}_j(\cdot) \) such that \( \tilde{Z}_j'(\cdot) = d \ Z_j(\cdot) \) conditional on the data, and

\[
P^* \left[ \sup_{x \in \mathcal{X}} |\tilde{Z}_j'(x) - Z_j(x)| \geq \eta r_n^{-1} \right] = o_P(1), \quad \forall \eta > 0,
\]

where, for a \( N_{K_j} \sim N(0, I_{K_j}) \) with \( K_j = \dim(\Pi_j(x)) \),

\[
\tilde{Z}_j(x) = \frac{\tilde{\gamma}_{q,j}(x) \Sigma_j^{1/2}}{\sqrt{\Omega_j(x)}} N_{K_j}, \quad x \in \mathcal{X}, \quad j = 0, 1, 2, 3.
\]
This approximation is denoted by $\hat{Z}_j(\cdot) = d^* Z_j(\cdot) + o_P(r_n^{-1})$ in $L^\infty(\mathcal{X})$. ■

From a practical perspective, Definition 6.2 implies that sampling from $\hat{Z}_j(\cdot)$, 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 $\hat{Z}_j(\cdot)$ relies on a direct plug-in approach, where all the unknown quantities are replaced by consistent estimators already used in forming $\hat{T}_j(x)$. 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 $\sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}|x_i = x] < \infty$, assume $n^{-\frac{d}{2}} (\log(n))^{\frac{d+2\nu}{2}} = o(r_n^{-2})$; 
(ii) when $\sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^{3} \exp(|\varepsilon_i|)|x_i = x] < \infty$, assume $\frac{\log(n)^{\nu/2}}{\sqrt{nh^d}} = o(r_n^{-2})$.

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

This result, proven in §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

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

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

$$\left[ \partial^q \mu_j(x) \pm q_j(\alpha) \sqrt{\hat{\Omega}_j(x)/n} : x \in \mathcal{X} \right].$$

The following theorem, proven in §8.6, establishes a valid distributional approximation for the suprema of the $t$-statistic processes $\{\hat{T}_j(x) : x \in \mathcal{X}\}$ using [17, 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

\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left[ \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| \leq u \right] - \mathbb{P}^* \left[ \sup_{x \in \mathcal{X}} |\hat{Z}_j(x)| \leq u \right] \right| = o_p(1).
\]

[16, 17] 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: [17, 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, [41] 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 (§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 \( x_i \sim \mathcal{U}[0, 1] \) and \( \varepsilon_i \sim \mathcal{N}(0, 1) \), independent of each other. We consider 5,000 simulated
datasets with \( n = 1,000 \) 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 \( \hat{\mu}_0(x) \), 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, \( \kappa \) is set using either an infeasible IMSE-optimal choice \( (\kappa_{\text{IMSE}}) \), a rule-of-thumb estimate \( (\kappa_{\text{ROT}}) \), or a direct plug-in estimate \( (\kappa_{\text{DPI}}) \). For wavelets, the tuning parameter is instead the resolution level \( (\text{resp.}, s_{\text{IMSE}}, \hat{s}_{\text{ROT}}, \text{or} \hat{s}_{\text{DPI}}) \), 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) correction seems to perform quite well, always delivering close-to-correct coverage.


8.1. Technical Lemma. Let \( \hat{Q}_m = \mathbb{E}_n[p(x_i)p(x_i')] \), \( \hat{Q}_{\tilde{m}} = \mathbb{E}_n[\hat{p}(x_i)\hat{p}(x_i')] \), \( Q_m = \mathbb{E}[p(x_i)p(x_i')] \), and \( Q_{\tilde{m}} = \mathbb{E}[\hat{p}(x_i)\hat{p}(x_i')] \).

**Lemma 8.1.** Let Assumptions 1, 2, 3, and 5 hold. If \( \frac{\log n}{nh^{d+2}} = o(1) \), then:

(i) \( \|\hat{Q}_m - Q_m\| \preceq P h^d \sqrt{\frac{\log n}{nh^{d+2}}} \), \( \|\hat{Q}_m - Q_m\|_\infty \preceq P h^d \sqrt{\frac{\log n}{nh^{d+2}}} \); (ii) \( \|\hat{Q}_{\tilde{m}} - Q_{\tilde{m}}\| \preceq P h^d \), \( \|\hat{Q}_{\tilde{m}} - Q_{\tilde{m}}\|_\infty \preceq P h^d \); (iii) for each \( j = 0, 1, 2, 3 \), \( \sup_{x \in \mathcal{X}} \|\gamma_{q,j}(x)'\|_\infty \preceq h^{-d-[q]} \), \( \sup_{x \in \mathcal{X}} \|\gamma_{q,j}(x)\|_\infty \preceq h^{-d-[q]} \); (iv) for \( j = 1, 2, 3 \), \( \sup_{x \in \mathcal{X}} \Omega_j(x) \preceq P h^{-d-2[q]} \).

**Proof:** SA, Section SA-10.

These results for \( \hat{Q}_m \) and \( Q_m \) also hold for \( \hat{Q}_{\tilde{m}} \) and \( Q_{\tilde{m}} \) under Assumption 5. See the SA (§SA-2) for details and other related results.

8.2. Proof of Lemma 6.1. First, suppose that condition (i) holds. Theorem SA-4.2 of the SA shows that \( \sup_{x \in \mathcal{X}} |\hat{\Omega}_0(x) - \Omega_0(x)| \preceq P n^{-\frac{1}{2}} h^{-\frac{d}{2}} - 2q |(\log n)^{\frac{1}{2}} +
\[ n^{\frac{1}{2+\nu}}(\log n)^{\frac{\nu}{4+\nu}} + \sqrt{n} h \mathbf{1} \] and, for \( j = 1, 2, 3 \), \( \sup_{x \in X} \left| \hat{\Omega}_j(x) - \Omega_j(x) \right| \leq \epsilon \).

Then, on a sufficiently rich probability space, there exists a copy \( \Omega(x) \) such that

\[ A \leq 1 \]

\( \Omega(x) \) converges rate of \( \hat{\Omega}_j(x), \Omega_j(x) \) = \( \mathcal{O}(r^{-1}_n) \), and the rate conditions imposed.

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

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

**Lemma 8.2 (Kernel-Based KMT Coupling).** Suppose \{\( (x_i, \varepsilon_i) : 1 \leq i \leq n \)\} are i.i.d., with \( x_i \in X \subseteq \mathbb{R} \) and \( \sigma_i^2 := \sigma^2(x_i) = \mathbb{E}[\varepsilon_i^2|x_i] \). Let \( \{A(x) := G_n(x, x_i) : x \in X \} \) be a stochastic process with \( X(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R} \) an \( n \)-varying kernel function possibly depending on \( X \). Assume one of the following holds:

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

\[
\sup_{x \in X} \max_{1 \leq i \leq n} |X(x, x_i)| = o_P(r^{-1}_n n^{-\frac{1}{2+\nu} + \frac{1}{2}}),
\]

\[
TV_X(X(x, \cdot)) = o(r^{-1}_n n^{-\frac{1}{2+\nu} + \frac{1}{2}}); \quad \text{or}
\]

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

\[
\sup_{x \in X} \max_{1 \leq i \leq n} |X(x, x_i)| = o_P(r^{-1}_n (\log n)^{-1} \sqrt{n}),
\]

\[
TV_X(X(x, \cdot)) = o(r^{-1}_n (\log n)^{-1} \sqrt{n}).
\]

Then, on a sufficiently rich probability space, there exists a copy \( A(\cdot) \) of \( A(\cdot) \), and an i.i.d. sequence \{\( \zeta_i : 1 \leq i \leq n \)\} of standard Normal random variables such that \( A(x) = d G_n \left[ \mathcal{X}(x, x_i) \sigma_i \zeta_i \right] + o_P(r^{-1}_n) \) in \( \mathcal{L}_{\infty}(X) \).

**Proof.** Suppose condition (i) holds. Let \{\( x(i) : 1 \leq i \leq n \)\} be the order statistics of \( \{x_i : 1 \leq i \leq n \} \) such that \( x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \), which also induces the concomitants \{\( \varepsilon_{(i)} : 1 \leq i \leq n \)\} and \( \sigma_{(i)}^2 = \sigma^2(x_{(i)}) : 1 \leq i \leq n \). Conditional on \( X \), \( \{\varepsilon_{(i)} : 1 \leq i \leq n \} \) is an independent mean zero sequence with \( \forall \varepsilon_{(i)}|X| = \sigma_{(i)}^2 \).

By [33, Corollary 5], there exists a sequence of i.i.d standard normal random variables \{\( \zeta_{(i)} : 1 \leq i \leq n \)\} such that \( \max_{1 \leq i \leq n} |S_{i,n}| \leq P n^{-\frac{1}{2+\nu}} \), where \( S_{i,n} := \sum_{i=1}^n (\varepsilon_{(i)} - \sigma_{(i)} \zeta_{(i)}) \).
Then, using summation by parts, \( \sup_{x \in X} | \sum_{i=1}^{n} \mathcal{H}(x, x_i) (\varepsilon_i - \bar{\varepsilon}_i \bar{\sigma}_i) | = \sup_{x \in X} \mathcal{H}(x, x_n) S_{n,n} - \sum_{i=1}^{n-1} S_{i,n} (\mathcal{H}(x, x_{i+1}) - \mathcal{H}(x, x_i)) \leq (\sup_{x \in X} \max_{1 \leq i \leq n} | \mathcal{H}(x, x_i) + \sup_{x \in X} \sum_{i=1}^{n-1} | \mathcal{H}(x, x_{i+1}) - \mathcal{H}(x, x_i) | ) \max_{1 \leq i \leq n} | S_{l,n} | . \]

Since \( \sum_{i=1}^{n-1} | \mathcal{H}(x, x_{i+1}) - \mathcal{H}(x, x_i) | \leq TV_X (\mathcal{H}(x, \cdot)) \), we have \( A(x) = d \mathbb{G}_{n} [ \mathcal{H}(x, x_i) \sigma_i \xi] + o_{P} (r_n^{-1}) \).

When (ii) holds, the proof is the same except that under the stronger moment restriction, \( \max_{1 \leq l \leq n} | S_{l,n} | \leq C \log n \) by [32, Theorem 1].

To prove Theorem 6.1, for each \( j = 0, 1, 2, 3 \), let \( \mathcal{H}(x, u) = \gamma_{q_j}(x) \Pi_j (u) / \sqrt{\Omega_j(x)} \) and observe that \( \sup_{x \in X} \sup_{u \in X} | \mathcal{H}(x, u) | \leq h^{-d/2} \).

By Lemma 8.1, the uniform bound on the total variation of \( \mathcal{H}(x, u) \) can also be verified easily. Alternatively, simply note that \( \| \sum_{i=1}^{n} S_{i,n} (\mathcal{H}(x, x_{i+1}) - \mathcal{H}(x, x_i)) \| \leq \| \gamma_{q_j}(x) \Pi_j (\sqrt{\Omega_j(x)}) \| \| \sum_{i=1}^{n-1} S_{i,n} (\Pi_j (x_{i+1}) - \Pi_j (x_i)) \| \). By Assumption 3 and Lemma 8.1, \( \sup_{x \in X} \sup_{u \in X} | \gamma_{q_j}(x) \Pi_j (u) / \sqrt{\Omega_j(x)} | \leq h^{-d/2} \).

Also, write the \( j \)-th element of \( \Pi_j (\cdot) \) as \( \pi_{j,j}(\cdot) \). Then, \( \max_{1 \leq i \leq K_j} | \sum_{i=1}^{n} (\pi_{j,j}(x_{i+1}) - \pi_{j,j}(x_i)) S_{l,n} | \leq \max_{1 \leq i \leq K_j} | \pi_{j,j}(x_{i+1}) - \pi_{j,j}(x_i) | \max_{1 \leq l \leq n} | S_{l,n} | . \)

By Assumption 3 and 5, \( \max_{1 \leq i \leq K_j} \sum_{i=1}^{n} | \pi_{j,j}(x_{i+1}) - \pi_{j,j}(x_i) | \leq 1 \). Thus, using Lemma 8.2, under the corresponding moment conditions and rate restrictions, there exists an independent standard normal sequence \( \{ \zeta_i : 1 \leq i \leq n \} \) such that \( \mathbb{G}_{n} [\mathcal{H}(x, x_i) \varepsilon_i] = d z_j(x) + o_{P} (r_n^{-1}) \).

Next, note that \( z_j(x) = d x \gamma_{q_j}(x) \Pi_j (\sqrt{\Omega_j(x)}) + \gamma_{q_j}(x) \Pi_j (\sqrt{\Omega_j(x)}) \) where \( \Pi_j (\cdot) \) is a \( K_j \)-dimensional standard normal vector (independent of \( X \)) and \( "= d x " \) denotes that two processes have the same conditional distribution given \( X \). For the second term, by Gaussian Maximal inequality [20, Lemma 13], \( \mathbb{E} \left[ \left\| \hat{\Sigma}_j^{1/2} - \Pi_j (\sqrt{\Omega_j(x)}) \right\|_\infty \right] \leq \sqrt{\log n} \| \hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2} \| . \)

By the same argument given in the proof of Lemma SA-2.1 in §SA-10.1, \( \| \hat{\Sigma}_j - \Sigma_j \| \leq h^{d/2} (\log n / (nh^d))^{1/2} \). Then by [4, Theorem X.1.1], \( \| \hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2} \| \leq h^{d/2} (\log n / (nh^d))^{1/4} \). For \( j = 0, 1, \) a sharper bound is available:

by [4, Theorem X.3.8] and Lemma 8.1, \( \| \hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2} \| \leq \lambda_{\min} (\Sigma_j^{1/2} - \Pi_j (\sqrt{\Omega_j(x)})) \| \| \Pi_j (\sqrt{\Omega_j(x)}) | \| \leq h^{-d/2} \sqrt{\log n} \| \hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2} \| . \)

Thus, it follows that \( \mathbb{E} \left[ \sup_{x \in X} | \gamma_{q_j}(x) \Pi_j (\sqrt{\Omega_j(x)}) | \right] \leq h^{-d/2} \sqrt{\log n} \| \Sigma_j^{1/2} - \Pi_j (\sqrt{\Omega_j(x)}) \| . \)

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 \( \xi_i = \frac{1}{\sqrt{n}} \Pi_j (x_i) \varepsilon_i \). \( \{ \xi_i : 1 \leq i \leq n \} \) is an i.i.d. sequence of \( K_j \)-dimensional random vectors, and \( \sum_{i=1}^{n} \mathbb{E} \left[ \| \xi_i \|_2 \| \xi_i \|_\infty \right] = \mathbb{E} \left[ \| \Pi_j (x_i) \varepsilon_i \|_2 \right] / \sqrt{n} \leq \mathbb{E} \left[ \Pi_j (x_i)^4 \right] / \sqrt{n} \leq \frac{n^{-1/2}}{\sqrt{n}} \) by...
Assumption 3, the moment condition imposed, and Lemma 8.1. On the other hand, let \{g_i : 1 \leq i \leq n\} be a sequence of independent Gaussian vectors with mean zero and variance \(\Sigma_j/n\). By properties of Gaussian random variables and Lemma 8.1, \((\mathbb{E}[\|g_i\|^2])^{1/2} \leq \sqrt{\log n} + \sum_{i=1}^{n} \mathbb{E}[\|g_i\|^4])^{1/2} \leq \text{trace}(\sum_{i=1}^{n} \mathbb{E}[\xi_i \xi_i^\top]) \leq 1\), and thus \(L_n := \sum_{i=1}^{n} \mathbb{E}[\|g_i\|^2]\|g_i\|_{\infty} \leq \sqrt{\log n/n}\). Then, there exists a \(K_j\)-dimensional normal vector \(N_{K_j}\) with variance equal to \(\Sigma_j\) such that for any \(t > 0\), \(\mathbb{P}(\|\sum_{i=1}^{n} \xi_i - N_{K_j}\|_{\infty} > r_n^{-1} 3h^2 t) \leq \min_{\tau \geq 0} (2\mathbb{P}(\|Z\|_{\infty} > \tau) + h^{-3/2} t^{-3/2} L_n \tau^2) \leq t^{-3/2} (\log n)^{3/2}/\sqrt{nh^3d}\) where \(Z\) 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 \(\sup_{x \in X} |\gamma_{q,j}(x)|/\sqrt{\Omega_j(x)}|_{\infty} \leq h^{-d/2}\) again, the result follows.

8.5. **Proof of Theorem 6.3.** For each \(j = 0, 1, 2, 3\),
\[
\hat{Z}_j(x) - Z_j(x) = (\frac{\gamma_{q,j}(x)}{\Omega_j'(x)} \cdot \frac{\gamma_{q,j}(x)}{\Omega_j'(x)}) \Sigma_j^{1/2} N_{K_j} + \frac{\gamma_{q,j}(x)}{\sqrt{\Omega_j'(x)}},
\]
Conditional on the data, each term on the RHS is a mean-zero Gaussian process. The desired results can be obtained by applying the Gaussian Maximal inequality to each term as in Section 8.3 and using Lemma 8.1 and Theorem SA-4.2 in §SA-4.

8.6. **Proof of Theorem 6.4.** In view of Theorem 6.1 and 6.2, there exists a sequence of constants \(\eta_n\) such that \(\eta_n = o(1)\) and \(\mathbb{P}(\sup_{x \in X} |\hat{T}_j(x)| - \sup_{x \in X} |Z_j(x)| > \eta_n/r_n) = o(1)\). Therefore, for any \(u \in \mathbb{R}\), \(\mathbb{P}(\sup_{x \in X} |\hat{T}_j(x)| \leq u) \leq \mathbb{P}(\sup_{x \in X} |\hat{T}_j(x)| - \sup_{x \in X} |Z_j(x)| \leq \eta_n/r_n) + \mathbb{P}(\sup_{x \in X} |\hat{T}_j(x)| - \sup_{x \in X} |Z_j(x)| > \eta_n/r_n) \leq \mathbb{P}(\sup_{x \in X} |Z_j(x)| \leq u + \eta_n/r_n) + o(1) \leq \mathbb{P}(\sup_{x \in X} |Z_j(x)| \leq u) + Cr_n^{-1} \eta_n \mathbb{E}[\sup_{x \in X} |Z_j(x)|] + o(1)\), for some constant \(C > 0\) where the inequality holds by the Anti-Concentration Inequality due to [17]. By Gaussian Maximal inequality, \(\mathbb{E}[\sup_{x \in X} |Z_j(x)|] \leq \sqrt{\log n}\). Since we assume \(r_n = \sqrt{\log n}\), the two terms on the far right of the last inequality is \(o(1)\) and do not depend on \(u\). The reverse of the inequality follows similarly, and thus \(\sup_{u \in \mathbb{R}} \mathbb{P}(\sup_{x \in X} |\hat{T}_j(x)| \leq u) = o(1)\). On the other hand, by Theorem 6.3, \(\hat{Z}_j(\cdot)\) is approximated by the same Gaussian process conditional on the data. Thus, using the same argument given above, the result follows.

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 \( \mathsf{R} \) package \texttt{lspartition} \cite{CattaneoFeng2018}.

**Acknowledgements.** 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 Supplement A for supplementary materials.

**SUPPLEMENTARY MATERIAL**

**Supplement A:** Additional Technical Results, Omitted Proofs, Implementation Details, and Further Simulation Results \( \text{(http://arxiv.org/pdf/1804.04916)} \). 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 (§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.**


Yingjie Feng
Department of Politics
Princeton University
Princeton, NJ 08544
E-mail: yjfeng@umich.edu

Matias D. Cattaneo
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544
E-mail: cattaneo@princeton.edu

Max H. Farrell
Booth School of Business
University of Chicago
Chicago, IL 60637
E-mail: max.farrell@chicagobooth.edu

Yingjie Feng
Department of Politics
Princeton University
Princeton, NJ 08544
E-mail: yjfeng@umich.edu
Table 1  
*Simulation Evidence*  
(a) B-Splines \((m = 2, \tilde{m} = 3, \Delta = \tilde{\Delta}, \text{Evenly Spaced Partition})\)

<table>
<thead>
<tr>
<th>(j)</th>
<th>(\kappa) IMSE</th>
<th>(\hat{\kappa}) ROT</th>
<th>(\hat{\kappa}) DPI</th>
<th>(\kappa)</th>
<th>RMSE</th>
<th>Pointwise</th>
<th>Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.0</td>
<td>4.1</td>
<td>4.7</td>
<td>0.046</td>
<td>0.002</td>
<td>0.008</td>
<td>91.6</td>
</tr>
<tr>
<td>1</td>
<td>3.0</td>
<td>4.1</td>
<td>4.7</td>
<td>0.003</td>
<td>0.008</td>
<td>0.007</td>
<td>94.8</td>
</tr>
<tr>
<td>2</td>
<td>3.0</td>
<td>4.1</td>
<td>4.7</td>
<td>0.016</td>
<td>0.007</td>
<td>0.004</td>
<td>90.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>(j)</th>
<th>(s) IMSE</th>
<th>(\hat{s}) ROT</th>
<th>(\hat{s}) DPI</th>
<th>(s)</th>
<th>RMSE</th>
<th>Pointwise</th>
<th>Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.0</td>
<td>2.8</td>
<td>2.8</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>94.2</td>
</tr>
<tr>
<td>1</td>
<td>3.0</td>
<td>2.8</td>
<td>2.8</td>
<td>0.036</td>
<td>0.036</td>
<td>0.036</td>
<td>93.6</td>
</tr>
<tr>
<td>2</td>
<td>3.0</td>
<td>2.8</td>
<td>2.8</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>94.2</td>
</tr>
</tbody>
</table>

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) \(\kappa\) IMSE and \(s\) IMSE = infeasible IMSE-optimal number of partitions, \(\hat{\kappa}\) ROT and \(\hat{s}\) ROT = feasible rule-of-thumb (ROT) implementation of \(\kappa\) IMSE, \(\hat{\kappa}\) DPI and \(\hat{s}\) DPI = feasible direct plug-in (DP1) implementation of \(\kappa\) IMSE. See §SA-8 and §SA-9 in supplemental appendix for more details.