Large Sample Properties of Partitioning-Based Series Estimators

Matias D. Cattaneo† Max H. Farrell‡ Yingjie Feng§

April 16, 2018

Abstract

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, employing a carefully crafted coupling approach, we develop uniform distributional approximations for \( t \)-statistic processes (indexed over the support of the conditioning variables). These results cover both undersmoothed and bias corrected inference, require seemingly minimal rate restrictions, and achieve fast approximation rates. Using the uniform approximations we construct valid confidence bands which share these same advantages. While our coupling approach exploits specific features of the estimators considered, the underlying ideas may be useful in other contexts more generally. Second, we obtain valid integrated mean square error approximations for partitioning-based estimators and develop feasible tuning parameter selection. We apply our general results to three popular partition-based estimators: splines, wavelets, and piecewise polynomials (generalizing the regressogram). The supplemental appendix includes other general and example-specific technical results that may be of independent interest. A companion \( \text{R} \) package is provided.

Keywords: nonparametric regression, series methods, linear sieve methods, robust bias correction, uniform inference, strong approximation, tuning parameter selection.

---

*We thank Victor Chernozhukov, Denis Chetverikov, Michael Jansson, Xinwei Ma, Whitney Newey, and Andres Santos for useful discussions. Cattaneo gratefully acknowledges financial support from the National Science Foundation (SES 1459931).

†Department of Economics and Department of Statistics, University of Michigan.
‡Booth School of Business, University of Chicago.
§Department of Economics, University of Michigan.
Contents

1 Introduction .......................................................... 1
  1.1 Related Literature ............................................... 5
  1.2 Organization of This Paper ...................................... 6

2 Setup ........................................................................ 6
  2.1 Assumptions ......................................................... 7
  2.2 Characterization and Correction of Bias ......................... 11

3 Uniform Inference ....................................................... 14
  3.1 Strong Approximation .............................................. 15
  3.2 Implementation ..................................................... 19
  3.3 Application: Confidence Bands .................................. 21

4 Integrated Mean Squared Error ........................................ 23

5 Simulations .................................................................. 27

6 Conclusion .................................................................... 28

Appendix A Examples of Popular Partitioning-Based Estimators .... 29
  A.1 B-Splines on Rectangular Partitions ............................. 29
  A.2 Wavelets on Rectangular Partitions .............................. 31
  A.3 Generalized Regressograms on Rectangular Partitions ....... 32
  A.4 Generalized Regressograms on General Partitions ........... 33

Appendix B Further Technical Results for Rectangular Partitions .... 34

Supplemental Appendix ..................................................... 39

SA-7 Setup, Assumptions and Notation .................................. 39
  SA-7.1 Notation .......................................................... 39
  SA-7.2 Setup .............................................................. 40
  SA-7.3 Assumptions ..................................................... 41

SA-8 Technical Lemmas ..................................................... 44

SA-9 Pointwise Estimation and Inference ................................ 47

SA-10 Uniform Estimation and Inference ................................. 49
  SA-10.1 Uniform Convergence ....................................... 49
  SA-10.2 Strong Approximation ....................................... 52
SA-10.3 Implementation ........................................... 55
SA-10.4 Application: Confidence Bands .......................... 58

**SA-11 Discussion and Extensions** ............................... 59
SA-11.1 Connecting Splines and Piecewise Polynomials ......... 59
SA-11.2 Comparison of Bias Correction Approaches ............... 62

**SA-12 Implementation Details** ................................. 65
SA-12.1 Rule-of-Thumb Choice ..................................... 65
SA-12.2 Direct Plug-in Choice ...................................... 66

**SA-13 Simulations** ................................................ 68

**SA-14 Proofs: Supplemental Appendix** ......................... 69
SA-14.1 Proof of Lemma SA-8.1 ..................................... 69
SA-14.2 Proof of Lemma SA-8.2 ..................................... 72
SA-14.3 Proof of Lemma SA-8.3 ..................................... 72
SA-14.4 Proof of Lemma SA-8.4 ..................................... 74
SA-14.5 Proof of Lemma SA-8.5 ..................................... 75
SA-14.6 Proof of Lemma SA-9.1 ..................................... 79
SA-14.7 Proof of Theorem SA-9.1 ................................... 81
SA-14.8 Proof of Lemma SA-10.1 ................................... 81
SA-14.9 Proof of Lemma SA-10.2 ................................... 84
SA-14.10 Proof of Theorem SA-10.1 ................................. 84
SA-14.11 Proof of Theorem SA-10.2 ................................. 85
SA-14.12 Proof of Theorem SA-10.3 ................................. 88
SA-14.13 Proof of Lemma SA-10.3 ................................... 88
SA-14.14 Proof of Lemma SA-10.4 ................................... 89
SA-14.15 Proof of Theorem SA-10.5 ................................. 89
SA-14.16 Proof of Theorem SA-10.4 ................................. 91
SA-14.17 Proof of Theorem SA-10.5 ................................. 91
SA-14.18 Proof of Theorem SA-10.6 ................................. 91
SA-14.19 Proof of Theorem SA-10.7 ................................. 92
SA-14.20 Proof of Theorem SA-10.8 ................................. 93
SA-14.21 Proof of Theorem SA-10.9 ................................. 93
SA-14.22 Proof of Theorem SA-10.10 ............................... 95
SA-14.23 Proof of Theorem SA-10.11 ............................... 96
SA-14.24 Proof of Theorem SA-10.12 ............................... 96
SA-14.25 Proof of Lemma SA-11.1 ................................... 97

**SA-15 Proofs: Main Paper** ....................................... 102
| Section     | Title                        | Page |
|-------------|------------------------------|------|
| SA-15.1     | Proof of Lemma 2.1           | 102  |
| SA-15.2     | Proof of Theorem 3.1         | 102  |
| SA-15.3     | Proof of Theorem 3.2         | 102  |
| SA-15.4     | Proof of Theorem 3.3         | 102  |
| SA-15.5     | Proof of Theorem 3.4         | 102  |
| SA-15.6     | Proof of Theorem 4.1         | 102  |
| SA-15.7     | Proof of Lemma A.1           | 104  |
| SA-15.8     | Proof of Lemma A.2           | 107  |
| SA-15.9     | Proof of Lemma A.3           | 111  |
| SA-15.10    | Proof of Lemma A.4           | 112  |
| SA-15.11    | Proof of Theorem B.1         | 112  |
| SA-15.12    | Proof of Corollary 1         | 117  |
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

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

for a scalar response \( y_i \) and a \( d \)-vector of continuously distributed covariates \( x_i = (x_{1,i}, \ldots, x_{d,i})' \) with compact support \( \mathcal{X} \). The object of interest is the unknown regression function \( \mu(\cdot) \) and its derivatives. In this paper we focus on partitioning-based, or locally-supported, series (linear sieve) least squares regression estimators. These 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 basis function is nonzero on only a small, contiguous set of cells of the partition. This contrasts with, for example, global polynomial approximations. Popular examples of partitioning-based estimators are splines, compact-supported wavelets, and piecewise polynomials (which we call a generalized regressogram). For this class of estimators, we develop novel (pointwise and) uniform estimation and inference results for the the regression function \( \mu(\cdot) \) and its derivatives, with and without bias correction techniques, and integrated mean-square error expansions useful for tuning parameter selection. We also obtain other general and basis-specific technical results.

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 \( \mathcal{X} \)). Formalized below, we restrict \( \delta_l \) to be polyhedral, which allows for tensor products of intervals as well as other popular choices. Based on this partition, the dictionary of \( K \) basis functions, each of order \( m \) (e.g., \( m = 4 \) for cubic splines) is denoted by

\[
x_i \mapsto p(x_i) := p(x_i; \Delta, m) = (p_1(x_i; \Delta, m), \ldots, p_K(x_i; \Delta, m))'.
\]

Popular examples of bases and partitioning schemes are discussed precisely in Appendix A, where new technical results concerning bias approximations are also given. Under the regularity conditions given below, and for a point \( x \in \mathcal{X} \) and a \( q = (q_1, \ldots, q_d)' \in \mathbb{Z}_+^d \), the partial derivative \( \partial^q \mu(x) \) is
estimated by least squares regression
\[ \hat{\partial^q \mu}(x) = \partial^q p(x)' \hat{\beta}, \quad \hat{\beta} \in \text{arg min}_{b \in \mathbb{R}^K} \sum_{i=1}^n (y_i - p(x)' b)^2, \]
where \( \partial^q \mu(x) = \partial^{q_1 + \cdots + q_d} \mu(x)/\partial^{q_1} x_1 \cdots \partial^{q_d} x_d \) (and for boundary points defined from the interior of \( X \) as usual). Using the notation \( \mathbb{E}_n[g(x_i)] = n^{-1} \sum_{i=1}^n g(x_i) \) for a function \( g(\cdot) \), our assumptions will guarantee that the sample matrix \( \mathbb{E}_n[p(x_i)p(x_i)'] \) is nonsingular with probability approaching one in large samples, and thus we will write the estimator as
\[ \hat{\partial^q \mu_0}(x) := \hat{\gamma}_{q,0}(x)' \mathbb{E}_n[\Pi_0(x_i)y_i], \]
where \( \hat{\gamma}_{q,0}(x)' := \partial^q p(x)' \mathbb{E}_n[p(x_i)p(x_i)']^{-1} \) and \( \Pi_0(x_i) := p(x_i) \).

(The subscript "0" will differentiate from the bias-corrected estimators discussed below.) This notation is slightly nonstandard for regression-based estimators, but will be useful in elucidating our inference approach.

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 for this class of nonparametric estimators. Under our assumptions, stated precisely in the next section, \( \bar{\kappa} \to \infty \) 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 \( \bar{\kappa} \to \infty \), \( h^d \) vanishes at the same rate, and with each basis being supported only on a finite number of cells, \( K \) diverges proportionally as well. Depending on the particular basis and partition used, one of the three measures of granularity may be more convenient.

Our first main contribution is a strong approximation for the whole \( t \)-statistic process, indexed by \( x \in X \),
\[ \hat{T}_0(x) = \frac{\hat{\partial^q \mu_0}(x) - \hat{\partial^q \mu}(x)}{\sqrt{\hat{\Omega}_0(x)/n}}, \quad \hat{\Omega}_0(x) = \hat{\gamma}_{q,0}(x)' \hat{\Sigma}_0 \hat{\gamma}_{q,0}(x), \]
where \( \hat{\Omega}_0(x) \) is a consistent estimator for \( n\mathbb{V}[^{\hat{\partial^q \mu}(x)}|x_1, \ldots, x_n] \), in particular with a plug-in estimator \( \hat{\Sigma}_0 \) of the unknown matrix \( \Sigma_0 = \mathbb{E}[\Pi_0(x_i)\Pi_0(x_i)'] \sigma^2(x_i) \). Our distributional approximations rely on a carefully constructed coupling strategy together with sharp bounds for the convergence
of $\Sigma_0^{1/2}$ to $\Sigma_0^{1/2}$.

We obtain this strong approximation to the entire process under seemingly minimal conditions on the tuning parameter $h$ (i.e., on $K$ or $\bar{K}$). Section 3 develops the coupling strategy, which may be of independent interest, in two steps. First, we couple the $t$-statistic process to a process that is Gaussian only conditionally on the covariates $x_1, \ldots, x_n$, but not otherwise. We then construct an unconditionally Gaussian process $\{Z_0(x), x \in \mathcal{X}\}$ that approximates the conditionally Gaussian process. Thus, $\{Z_0(x), x \in \mathcal{X}\}$ can be used to approximate the distribution of the whole $t$-statistic process $\{\hat{T}_0(x), x \in \mathcal{X}\}$. Such a strong approximation can also be used to approximate the sampling distribution of continuous (with respect to the uniform norm) transformations of the $t$-statistic process (e.g., suprema). To illustrate, Section 3.3 constructs valid confidence bands for (derivatives of) the regression function.

Our coupling strategy allows for very weak conditions on the complexity of the partition $\Delta$, and gives fast approximation error rates. For example, if $\varepsilon_i$ possesses bounded fourth (conditional on $x_i$) moments, we require only $(nh^{2d})^{-1} \asymp K^2/n = o(1)$, up to $\log(n)$ terms, while if a subexponential moment bound holds, we can weaken this further to the seemingly minimal restriction that $(nh^d)^{-1} \asymp K/n = o(1)$, up to $\log(n)$ terms. We conjecture that these rate restrictions cannot be relaxed further beyond improvements in the log terms. Our coupling strategy thus delivers weaker restrictions than those known in the literature, in addition to offering an approximation for the entire process (not just a specific functional thereof such as the suprema). For example, [3] recently established an analogous result for the entire process $\{\hat{T}_0(x), x \in \mathcal{X}\}$, and for more general series-based estimators, where $K^5/n = o(1)$, up to $\log(n)$ terms is required regardless of moment assumptions imposed. Closely related are the contributions of [16, 17], where it was shown that if interest is constrained to the supremum of an empirical process, rather than the whole process, then only $K^2/n = o(1)$ (with fourth moment bounded) or $K/n = o(1)$ (with subexponential moment bounded) is required, up to $\log(n)$ terms. In comparison, we show in this paper that for the case of partitioning-based least squares estimators, there is no need to focus on the supremum of $\hat{T}_0(x)$ to obtain such weak rate restrictions and fast approximation error rates. Having such sharp coupling approximation for the whole $t$-statistic process has several theoretical and empirical advantages, including the fact that functionals other than the suprema can be considered directly.

Our second main contribution, in Section 4, is a general integrated mean-square error (IMSE) ex-
pansion for partitioning-based estimators. Although our strong approximation results for \( \{ \hat{T}_0(x), x \in \mathcal{X} \} \) allow for a wide range of partitions and associated tuning parameters \( h \asymp K^{1/d} \), a practical choice is still required. We exploit the structure of these estimators to obtain precise characterizations for the leading bias and variance terms. For simple cases on rectangular partitions, some results exist for splines [1, 65, 66] and generalized regressograms [12]; see Appendix A for more details. In addition to generalizing these results substantially (e.g., allowing for more general support and partitioning schemes), our characterization for compact-supported wavelets appears to be new.

The IMSE-optimal choice of partition granularity obeys \( h_{\text{IMSE}} \asymp n^{-1/(2m+d)} \), which translates to the familiar \( K_{\text{IMSE}} \asymp n^{-d/(2m+d)} \). And while the resulting \( \hat{\gamma}_0(x) \) is optimal from a point estimation perspective, inference using \( h_{\text{IMSE}} \) is invalid. More precisely, the \( t \)-statistic \( \hat{T}_0(x) \) is asymptotically Gaussian with unit variance but with nonzero bias. This leads to the common practice of ad-hoc undersmoothing: conducting inference using a finer partition than the one used to construct the IMSE-optimal point estimator \( \hat{\gamma}_0(x) \).

While undersmoothing is theoretically valid for inference, it is difficult to implement in a principled way. Further, inspired by results proving that undersmoothing is never optimal relative to bias correction for kernel-based nonparametrics [8], we develop three bias-corrected inference procedures using our new bias characterizations for 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. We consider three bias-corrected estimators, denoted by \( \partial q \hat{\gamma}_1(x) \), \( \partial q \hat{\gamma}_2(x) \), and \( \partial q \hat{\gamma}_3(x) \), all of which can be constructed parallelising (1.2):

\[
\partial q \hat{\gamma}_j(x) := \hat{\gamma}_{q,j}(x) \hat{\Sigma}^{-1} \hat{\gamma}_{q,j}(x), \quad j = 1, 2, 3,
\]

(1.3)

where \( \hat{\gamma}_{q,j}(x) \) and \( \hat{\Sigma}^{-1} \hat{\gamma}_{q,j}(x) \) are defined in Section 2.2, and depend on a further basis choice as well as the specific bias-correction strategy. Importantly, we extend the strong approximation results for \( \{ \hat{T}_0(x), x \in \mathcal{X} \} \) to hold for

\[
\hat{T}_j(x) = \frac{\partial q \hat{\gamma}_j(x) - \partial q \mu(x)}{\sqrt{\hat{\Omega}_j(x)/n}}, \quad \hat{\Omega}_j(x) = \hat{\gamma}_{q,j}(x) \hat{\Sigma}_j \hat{\gamma}_{q,j}(x),
\]

where \( \hat{\Sigma}_j \) is given in (3.1) in Section 3. We are able to unify our coupling approach, thereby
delivering uniform inference for these entire processes \((j = 1, 2, 3)\) as well, under similarly weak conditions, which constitutes our third main contribution. Due to the reduction in bias, inference based on \(\hat{T}_j(\cdot), j = 1, 2, 3\), is valid when employing the aforementioned IMSE optimal tuning parameter choice, thus offering a fully data-driven method combining optimal point estimation and valid inference. These bias correction results for partitioning-based estimators, both pointwise and uniformly in \(x\), appear to be new to the literature.

All our methods are available in a general-purpose R package [13], including implementations of the IMSE optimal tuning parameters and other results, which are reported only in the appendix and supplemental appendix to streamline the presentation. Finally, the supplemental appendix includes other general and example-specific technical results that may be of independent interest.

1.1 Related Literature

This paper contributes primarily to two literatures, nonparametric regression and strong approximations. There is a vast literature on nonparametric regression, summarized in many textbook treatments [e.g., 29, 35, 61, 40, 52, 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 [49, 41, 65, 42, 14, 12, 3, 15]. Piecewise polynomial fits on partitions have a long and ongoing tradition in statistics, dating at least to the regressogram of Tukey [59], and since variously named local polynomial regression [55], partitioning estimators/regression [35, 12], and the generalized regressogram herein, and continuing through modern, data-driven partitioning techniques such as regression trees [5, 38], trend filtering [57], and related methods [64]. Partitioning-based methods have also featured as inputs or preprocessing in treatment effects [11, 9], empirical finance [10], and other settings. The bias corrections we develop for series estimation and uniform inference follow recent work in kernel-based nonparametric inference [8, 7]. Our coupling and strong approximation results relate to early work discussed in [28, Chapter 22] and the recent work of [16, 17, 18, 19] and [63], as well as with the results for series estimators in [3]. See also [21] and [22] for recent applications of coupling techniques and related methods. Finally, using our strong approximation and bias-correction results, we develop valid confidence bands for (derivatives of) the regression
function; for recent related work on this topic also see [36] and references therein.

1.2 Organization of This Paper

Section 2 states our assumptions, and in the process precisely defines the partitioning schemes we consider as well as the three bias correction approaches. Section 3 proves strong approximations for all estimators, gives numerical methods for implementation, and applies the results to give confidence bands. Our main IMSE results are presented in Section 4, but Appendix B contains further related results. Section 5 discusses the numerical performance of our methods, and lastly, Section 6 concludes. Popular partitioning-based estimators (splines, wavelets, and the generalized regressionogram) are detailed in Appendix A, wherein we verify that these obey the high-level assumptions in Section 2. Complete theoretical proofs, further technical details, and extensive Monte Carlo studies are collected in the supplemental appendix (SA, hereafter).

2 Setup

We employ the following notation throughout. 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 $\partial^q \mu(x) = \partial^{[q]} \mu(x)/\partial x_1^{q_1} \cdots \partial x_d^{q_d}$. 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 $0 = (0, \ldots, 0)'$ be the length-$d$ zero vector. The tensor product or Kronecker product operator is $\otimes$. The smallest integer greater than or equal to $u$ is $\lfloor u \rfloor$.

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 $\lambda_{\min}(A)$ denotes the minimum eigenvalue of $A$.

We also use standard empirical process notation: $E_n[g(x_i)] = \frac{1}{n} \sum_{i=1}^n g(x_i)$, $E[g(x_i)]$, and $G_n[g(x_i)] = \frac{1}{\sqrt{n}} \sum_{i=1}^n (g(x_i) - E[g(x_i)])$. For sequences of numbers or random variables, we use $a_n \lesssim b_n$ to denote that $\limsup_n |a_n/b_n|$ is finite, $a_n = O_P(b_n)$ to denote $\limsup_{\epsilon \to \infty} \limsup_n \mathbb{P}[|a_n/b_n| \geq \epsilon] = 0$, $a_n = o_P(b_n)$ implies $a_n/b_n \to 0$, and $a_n = o_P(b_n)$ implies that $a_n/b_n \to 0$, where $\to$ denotes convergence in probability. $a_n \asymp b_n$ implies that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two random variables $X$ and $Y$, $X \equiv d Y$ implies that they have the same probability distribution. Finally, we set $\mu(x) := \partial^0 \mu(x)$.
and \( \bar{\mu}_j(x) := \partial^0 \mu_j(x) \) for \( j = 0,1,2,3 \) and collect the covariates as \( X = [x_1, \ldots, x_n]' \).

### 2.1 Assumptions

Our first assumption concerns 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 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 \( S > [q] \), 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 are “locally” supported. These assumptions are general enough to cover many special cases of interest in empirical work, as we illustrate in Appendix A, where we show that these assumptions hold for splines, wavelets, and piecewise polynomials. The first key substantive restriction on the partitioning restricts the shape of the resulting cells, but allows for quite irregular cell shapes.

**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 \( \delta \).

Assumption 2(a) rules out too irregular shapes of partitioning cells and implies that the size of each \( \delta \in \Delta \) can be well characterized by the diameter of \( \delta \). Assumption 2(b) implies 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. It is also clear that under Assumption 2 the volume of each element \( \delta \) is of the same magnitude as \( h^d \). In addition, whenever \( \Delta \) covers only strict subset of \( \mathcal{X} \), all our results hold on that subset.
A special case is a “rectangular” partition. If the support of the regressors is of tensor product form, then each dimension of \( X \) is partitioned marginally into intervals and \( \Delta \) is the tensor product of these intervals. Let \( X_{\ell} = [x_{\ell,0}, x_{\ell}] \) be the support of covariate \( \ell = 1, 2, \ldots, d \) and partition this into \( \kappa_{\ell} \) disjoint subintervals defined by \( \{x_{\ell} = t_{\ell,0} < t_{\ell,1} < \cdots < t_{\ell,\kappa_{\ell} - 1} < t_{\ell,\kappa_{\ell}} = \overline{x}_{\ell}\} \). If this partition of \( X_{\ell} \) is \( \Delta_{\ell} \), then a complete partition of \( X \) can be formed by tensor products of the one-dimensional partitions: \( \Delta = \otimes_{\ell=1}^{d} \Delta_{\ell} \), with \( \kappa = (\kappa_1, \kappa_2, \ldots, \kappa_d)' \) subintervals in each dimension of \( X \) and \( \overline{\kappa} = \kappa_1 \kappa_1 \cdots \kappa_d \). A generic cell of this partition is the rectangle

\[
\delta_{l_1 \ldots l_d} = \{x : t_{l_1, l_1} < x_{l_1} < t_{l_1, l_1+1}, \quad 0 \leq l_1 \leq \kappa_1 - 1 \quad \text{and} \quad 1 \leq l \leq d\}. \tag{2.1}
\]

Given this setup, Assumption 2 is verified by choosing the knot positions/configuration appropriately, often dividing \( X_{\ell} \) uniformly or by empirical quantiles.

Adapting our results to allow for random (data-driven) partitioning, such as marginal empirical quantiles for rectangular partitioning, is beyond the scope of this paper. Some specific results have been obtained in the literature for simple cases, in which data-driven methods yield partitions with shrinking cells (though not necessarily quasi-uniformly, as here), such as \( L_2 \)-consistency of regression trees [5, Ch. 12], \( L_2 \)-consistency uniformly over partitions forming a VC class [50], and univariate IMSE expansion for quantile-based regressograms [9]. Further, general data-dependence can be accommodated if one is willing to split the data, estimate the partition break points in the first sample, and perform inference in the second. In this way, quite general partitions can be used in our results, including data-driven methods such as regression trees and other modern machine learning techniques for which sample splitting is enjoying a renaissance [see e.g. 30, 20, and references therein].

The second assumption on the partitioning-based estimators employs generalized notions of stable local basis [25] and active basis [42]. We say a function \( p(\cdot) \) on \( \mathcal{X} \) is active on \( \delta \in \Delta \) if it is not identically zero on \( \delta \). Concrete examples include \( B \)-splines, compact-supported wavelets, and piecewise polynomials, among other possibilities; see Appendix A for more details.

**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 any \( k = 1, \ldots, K \), both the number of elements involved
in $\mathcal{H}_k$ 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 $\mathbb{E}[p(x_i)p(x_j)']$ 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). Finally, Assumption 3(c) is a standard restriction that 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: $K \asymp \bar{\kappa} \asymp h^{-d}$. We will impose conditions on the growth rate of these tuning parameters so that the $m$-th order least squares partitioning-based series estimator employing a basis of approximation satisfying the above conditions is well-defined in large samples.

To simplify notation, for each $x \in \mathcal{X}$ define $\delta_x$ as the element of $\Delta$ whose closure contains $x$ and $h_x$ for the diameter of this $\delta_x$. 

9
**Assumption 4 (Approximation Error).** For all $\varsigma$ satisfying $[\varsigma] \leq \varsigma$, given in Assumption 3, there exists $s^* \in S_{\Delta,m}$, the linear span of $p(x; \Delta, m)$, and

$$B_{m,\varsigma}(x) = - \sum_{u \in \Lambda_m} \partial^u \mu(x) h^{m-[u]} B_{u,\varsigma}(x)$$

such that

$$\sup_{x \in \mathcal{X}} |\partial^\varsigma \mu(x) - \partial^\varsigma s^*(x) + B_{m,\varsigma}(x)| \lesssim h^{m+e-[\varsigma]}$$

(2.2)

and

$$\sup_{\delta \in \Delta} \sup_{x_1, x_2 \in \text{clo}(\delta)} \frac{|B_{u,\varsigma}(x_1) - B_{u,\varsigma}(x_2)|}{\|x_1 - x_2\|} \lesssim h^{-1}$$

(2.3)

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

Equation (2.2) gives an explicit high-level expression of the leading (smoothing bias) approximation error. This level of precision is needed for our bias correction and IMSE expansion results, where a nonspecific rate assumption such as $\sup_{x \in \mathcal{X}} |\partial^q \mu - \partial^q s^*| \lesssim h^{m-[q]}$ will not suffice. The latter condition is commonly assumed when studying undersmoothing-based inference, and indeed, is implied by our assumptions.

The terms $B_{u,\varsigma}(x)$ in $B_{m,\varsigma}(x)$ are known functions of the point $x$ which depend on the particular partitioning scheme and bases used. The only unknowns in the approximation error $B_{m,\varsigma}$ are the higher-order derivatives of $\mu(\cdot)$. In Appendix A we verify this (and the other assumptions) for several special cases of interest in applications, including splines, wavelets, and generalized regressograms. We also give explicit formulas for the leading error in (2.2) and concrete examples of the set $\Lambda_m$. We assume sufficient smoothness exists to characterize these terms; the alternative case when smoothness constrains inference has also been studied [e.g. 27, 8].

The function $B_{m,\varsigma}$ 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: $B_{m,\varsigma}$ and the accompanying error from the linear projection of $B_{m,0}$ onto $S_{\Delta,m}$. We formalize this result in Lemma 2.1 below. The second bias term is usually ignored in the literature because in several cases the leading approximation error $B_{m,0}$ is approximately orthogonal to $p$ with respect to the Lebesgue mea-
sure. However, for general partitioning-based estimators this orthogonality need not to hold. The following remark formalizes this point, which is discussed in more detail in the upcoming sections.

**Remark 2.1 (Approximate Orthogonality).** Suppose Assumptions 1–4 hold. Then, the basis \( p \) is said to be *approximately orthogonal* to \( B_{m,0} \) if the following condition holds:

\[
\max_{1 \leq k \leq K} \int_{\mathcal{H}_k} p_k(x; \Delta, m) B_{m,0}(x) \, dx = o(h^{m+d}). \tag{2.4}
\]

When (2.4) 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; see Lemma A.1 below for an example), a sufficient condition for (2.4) is simply the orthogonality between \( B_{u,0} \) and \( p \) in \( L_2 \) with respect to the Lebesgue measure, for all \( u \in \Lambda_m \).

As shown in Appendix A, in some simple cases condition (2.4) is automatically satisfied if one constructs the leading error based on a basis representing the orthogonal complement of \( S_{\Delta,m} \), but it is not always easy to do so. For example, condition (2.4) is hard to verify when the partitioning employed is sufficiently uneven, as it is usually the case when employing modern machine learning techniques. All our results hold when this orthogonality fails, and in particular our bias correction approach and IMSE expansion explicitly account for the \( L_2 \) projection of \( B_{m,0} \) onto the approximating space spanned by \( p \), thus allowing for violations of (2.4).

### 2.2 Characterization and Correction of Bias

Under the assumptions above, we can precisely characterize the bias of \( \hat{\partial^q \mu}_0(x) \). This relies, in particular, on Assumption 4, but does not assume that (2.4) holds. Particular cases are discussed in Section 4 and Appendix A. Using this form, we will then describe three bias correction approaches.

**Lemma 2.1 (Conditional Bias).** *Let Assumptions 1–4 hold. If \( \frac{\log n}{nh^d} = o(1) \) and \( h = o(1) \), then*

\[
\mathbb{E}[\hat{\partial^q \mu}_0(x)|X] - \partial^q \mu(x) = \hat{\gamma}_{q,0}(x)'\mathbb{E}_n[\Pi_0(x_i)\mu(x_i)] - \partial^q \mu(x) \tag{2.5}
\]

\[
= \mathcal{B}_{m,q}(x) - \hat{\gamma}_{q,0}(x)'\mathbb{E}_n[\Pi_0(x_i)\mathcal{B}_{m,0}(x_i)] + O_P(h^{m+\rho-|q|}). \tag{2.6}
\]
The proof of this lemma generalizes an idea in [65, Theorem 2.2] to handle partitioning-based series estimators beyond the specific example of $B$-Splines on rectangular partitioning. The first component $\mathcal{B}_{m,q}(x)$ 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 (2.4), when it holds, suffices to guarantee that the projection of leading error is of smaller order (such as for $B$-splines on a rectangular partition). In general the bias will be $O(h^{m-[q]})$ and further, in finite samples both terms may be important even if (2.4) holds.

We will consider three bias correction methods to remove the leading bias terms of Lemma 2.1. 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 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, and most obvious approach, is simply to use the higher order basis in place of the original basis [c.f., 42, Section 5.3]. This is thus named higher-order-basis bias correction and numbered as approach $j = 1$. In complete parallel to (1.2) define

$$\hat{\partial}^q \mu_1(x) := \tilde{\gamma}_{q,1}(x)'E_n[\Pi_1(x_i)y_i],$$

where $\tilde{\gamma}_{q,1}(x)' := \partial^q \tilde{p}(x)'E_n[\tilde{p}(x_i)^2]^{-1}$ and $\Pi_1(x_i) := \tilde{p}(x_i)$.

This approach can be viewed as a bias correction of the original point estimator because, trivially, $\hat{\partial}^q \mu_1(x) = \hat{\partial}^q \mu_0(x) - (\hat{\partial}^q \mu_0(x) - \hat{\partial}^q \mu_1(x))$. Valid inference based on $\hat{\partial}^q \mu_1(x)$ can be viewed as “undersmoothing” applied to the higher-order point estimator, but is distinct from undersmoothing $\hat{\partial}^q \mu_0(x)$ (i.e., using a finer partition $\Delta$ and keeping the order fixed).

Our second approach makes use of the generic expression of the least squares bias in Equation (2.5). The unknown objects in this expression are $\mu$ and $\partial^q \mu$, both of which can be estimated using the higher-order estimator (2.7), with $q = 0$ for the former. By plugging these into Equation (2.5) and subtracting the result from $\hat{\partial}^q \mu_0(x)$ (from (1.2)), we obtain the least-squares bias correction,
numbered as approach 2:

\[
\hat{\varphi} q_{2}\mu_2(x) := \hat{\varphi} q_{0}\mu_0(x) - \left(\gamma q_{0}(x)'E_n[\Pi_0(x_i)\hat{\varphi} q_1(x_i)] - \hat{\varphi} q_{1}\mu_1(x)\right) = \hat{\gamma} q_{2}(x)'E_n[\Pi_2(x_i)\gamma_{i}]
\]

where \( \gamma q_{2}(x)' := \left(\gamma q_{0}(x)', -\gamma q_{0}(x)'E_n[p(x_i)\hat{p}(x_i)]E_n[\gamma_{i}']^{-1} + \gamma q_{1}(x)\right) \) (2.8)

and \( \Pi_2(x_i) := (p(x_i)', \hat{p}(x_i)')' \),

which is exactly of the same form as \( \hat{\varphi} q_{0}\mu_0(x) \) and \( \hat{\varphi} q_{1}\mu_1(x) \) (cf. Equations (1.2) and (2.7)), except for the change in \( \hat{\gamma} q_{j}(x) \) and \( \Pi_j(x_i) \).

Finally, approach number 3 targets the leading terms identified in Equation (2.6). We dub this approach plug-in bias correction, as it specifically estimates the leading bias terms, in fixed-\( n \) form, of \( \hat{\varphi} q_{0}\mu_0(x) \) according to Assumption 4. To be precise, we employ the explicit plug-in bias estimator

\[
\hat{\varphi} m,q(x) = -\sum_{u \in \Lambda_m} (\partial^u \hat{\mu}_1(x)) h_x^{m-|q|} B_{u,q}(x),
\]

with \( |q| < m \) and \( \Lambda_m \) as in Assumption 4, leading to

\[
\hat{\varphi} q_{3}\mu_3(x) := \hat{\varphi} q_{0}\mu_0(x) - \left(\hat{\varphi} m,q(x) - \hat{\gamma} q_{0}(x)'E_n[\Pi_0(x_i)\hat{\varphi} m,0(x_i)]\right) = \hat{\gamma} q_{3}(x)'E_n[\Pi_3(x_i)\gamma_{i}]
\]

where \( \hat{\gamma} q_{3}(x)' := \left(\hat{\gamma} q_{0}(x)', \sum_{u \in \Lambda_m} \left\{\hat{\gamma} u_{i}(x)'h_x^{m-|q|} B_{u,q}(x) - \hat{\gamma} q_{0}(x)'E_n[p(x_i)h_x^{m} B_{u,0}(x_i)\hat{\gamma} u_{i}(x_i)']\right\}\right) \),

and \( \Pi_3(x_i) := (p(x_i)', \hat{p}(x_i)')' \).

(2.9)

Once again, this estimator is of the familiar form \( \hat{\gamma} q_{j}(x)'E_n[\Pi_j(x_i)\gamma_{i}] \) but with different \( \hat{\gamma} q_{j}(x) \) and \( \Pi_j(x_i) \). As in Remark 2.1, when the orthogonality condition (2.4) holds, the second correction term in \( \hat{\varphi} q_{3}\mu_3(x) \) is asymptotically negligible relative to the first. However in finite samples both terms can be important, so we consider the general case.

Our results will require the following conditions on the higher order basis, which essentially codify the idea of bias correction while keeping within partitioning-based series estimators. Namely, we assume the higher-order basis has sufficient approximation power, local support properties with a mesh size proportional to \( h \), and we must handle any covariance between the uncorrected estimator and the estimated leading bias in the distributional approximations (for approaches 2 and 3).
Assumption 5 (Bias Correction). The partition $\tilde{\Delta}$ satisfies Assumption 2, with maximum mesh $\tilde{h}$, and the basis $\tilde{p}(x; \tilde{\Delta}, \tilde{m})$ satisfies Assumptions 3 and 4 with $\zeta = \zeta(\tilde{m}) \geq m$ in place of $\zeta$. Let $\rho := h/\tilde{h}$, which obeys $\rho \to \rho_0 \in (0, \infty)$. In addition, for $j = 3$, either (i) $\hat{p}(x; \tilde{\Delta}, \tilde{m})$ 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 $\hat{p}(x; \tilde{\Delta}, \tilde{m})$ reproduce polynomials of degree $[q]$.

Assumption 5 can be used to guide the selection of the basis for bias estimation. In addition to removing the leading bias, the conditions require 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 required for plug-in bias correction ($j = 3$) due to the more complicated covariance between $\hat{\partial} q \mu_0$ and the estimated leading bias; the condition shown is a mild one under which the asymptotic variance can still be controlled. 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 high-level conditions in Assumption 5 are easy to verify in a case-by-case basis, as shown in Appendix A. The bias correction strategies are compared further in the SA (§SA-5.2).

3 Uniform Inference

We now turn to our strong approximation results. Our goal is to find a valid distributional approximation for the whole process $\{\hat{T}_j(x), x \in \mathcal{X}\}$ for each $j = 0, 1, 2, 3$. Such an approximation is proven using our coupling strategy, which is detailed herein. Following this, we demonstrate several sampling-based feasible implementations of the result. We then close this section with an application of our results to construct asymptotically valid confidence bands for $\partial q \mu(\cdot)$.

The SA (§SA-4.1) shows that under Assumptions 1–4 and if $\mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ for some $\nu > 0$, then the partitioning-based series estimator $\hat{\partial} q \mu_0(\cdot)$ attains the optimal (mean-square and uniform) convergence rate [55], and we give distributional approximations pointwise in $x$. These results were published before for different classes of series estimators under alternative high-level assumptions in [12], [3], and [15]. In this section, they are ingredients in proving uniform inference, and can be useful for tuning parameter choices. Section 4 shows how to choose IMSE-optimal partitions $\Delta$ and mesh sizes $h \asymp n^{-\frac{1}{m+d}}$ (or, equivalently, $K \asymp n^{\frac{d}{m+d}}$) to attain those optimal rates.
3.1 Strong Approximation

The four $t$-statistics discussed previously, and indexed by $j = 0, 1, 2, 3$, are defined as

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

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

recalling the definitions of $\hat{\mu}_j(x)$, $\hat{\gamma}_{q,j}(x)$, and $\Pi_j(x_i)$ in Equations (1.2), (2.7), (2.8), and (2.9). Once the basis functions and partitioning schemes are chosen, the statistic $\hat{T}_j(x)$ is readily implementable. To state the strong approximation results for the processes $\{\hat{T}_j(x) : x \in \mathcal{X}\}$, it will be instructive to go briefly over the several steps involved, so that our coupling strategy can be made clear. The final result is stated afterwards.

The first step is to show that the estimation and sampling uncertainty of $\hat{\gamma}_{q,j}(x)$ and $\hat{\Omega}_j(x)$ are negligible uniformly over $x \in \mathcal{X}$. Let $\gamma_{q,j}(x)$, $j = 0, 1, 2, 3$, be defined as $\hat{\gamma}_{q,j}(x)$ in (1.2), (2.7), (2.8), and (2.9), respectively, but with sample averages and other estimators replaced by their population counterparts, and similarly set $\Omega_j(x) = \gamma_{q,j}(x)\Sigma_j\gamma_{q,j}(x)$ with $\Sigma_j = \mathbb{E}[\Pi_j(x_i)\Pi_j(x_i)'\sigma^2(x_i)]$. We first prove that, for a positive non-vanishing sequence $r_n$, and for each $j = 0, 1, 2, 3$,

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

This step is formalized in Lemmas SA-4.3 and SA-4.4 in the SA.

Next comes our strategy for constructing an unconditionally Gaussian process that approximates the distribution of the whole $t$-statistic processes $\{t_j(x) : x \in \mathcal{X}\}$. Our coupling strategy consists of two parts. We first couple $t_j(x)$ to a process $\{z_j(x) : x \in \mathcal{X}\}$ that is Gaussian only conditionally on $X$ but not unconditionally. Second, we show that there is an unconditionally Gaussian process $\{Z_j(x) : x \in \mathcal{X}\}$ that approximates the distribution of $\{z_j(x) : x \in \mathcal{X}\}$.

To complete the conditional (on $X$) coupling step, we employ a version of the classical KMT inequalities \[44, 45\] that applies to independent but non-identically distributed (i.n.i.d) random variables \[53\]. 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 $X$. This part of our proof is inspired by (but distinct than) the one given in [28, Chapter 22], where a conditional strong
approximation result for smoothing splines is established. Crucially, our approach allows for weak rate restrictions because it effectively corresponds to univariate coupling. Put together, for each \( j = 0, 1, 2, 3 \), we prove that on an appropriately enlarged probability space there exists a copy \( t'_j(x) \) of \( t_j(x) \), and an i.i.d. sequence \( \{ \zeta_i : 1 \leq i \leq n \} \) of standard Normal random variables, such that

\[
\sup_{x \in \mathcal{X}} |t'_j(x) - z_j(x)| = o_P(r_n^{-1}), \quad z_j(x) = \frac{\gamma_{q,j}(x)' \bar{\Sigma}^{1/2}}{\sqrt{\Omega_j(x)}} G_n[\Pi_j(x_j)\sigma(x_j)\zeta]. \tag{3.3}
\]

This first part of our coupling strategy is formalized in Theorems SA-4.4 and SA-4.5 in the SA, where we also present a more general result (Lemma SA-4.5) that may be of independent interest. This intermediate result has the obvious drawback that the coupling process \( \{z_j(x) : x \in \mathcal{X}\} \) is Gaussian only conditionally on \( \mathbf{X} \) but not unconditionally.

The final unconditional coupling step addresses this shortcoming. We use the fact that, for each \( j = 0, 1, 2, 3 \), there exists a standard Normal random vector \( \mathbf{N}_{K_j} \sim \mathcal{N}(0, I_{K_j}) \) with \( K_j = \dim(\Pi_j(x)) \) on a sufficiently rich probability space such that \( z'_j(x) \overset{d}{=} \bar{Z}_j(x) \) conditional on \( \mathbf{X} \), where

\[
\bar{Z}_j(x) = \frac{\gamma_{q,j}(x)' \bar{\Sigma}^{1/2}}{\sqrt{\Omega_j(x)}} \mathbf{N}_{K_j}, \quad \bar{\Sigma}_j := \mathbb{E}_n[\Pi_j(x_j)\Pi_j(x_j)'\sigma^2(x_j)].
\]

Then, we establish

\[
\sup_{x \in \mathcal{X}} |\bar{Z}_j(x) - Z_j(x)| = o_P(r_n^{-1}), \quad Z_j(x) = \frac{\gamma_{q,j}(x)' \Sigma^{1/2}}{\sqrt{\Omega_j(x)}} \mathbf{N}_{K_j}, \tag{3.4}
\]

by showing

\[
\|\bar{\Sigma}_j^{1/2} - \Sigma_j^{1/2}\| = o_P \left( \sup_{x \in \mathcal{X}} \left| \frac{\gamma_{q,j}(x)'}{\sqrt{\Omega_j(x)}} \right| \right)^{-1} \frac{r_n^{-1}}{\sqrt{\log n}},
\]

where \( \sup_{x \in \mathcal{X}} \|\gamma_{q,j}(x)'/\sqrt{\Omega_j(x)}\|_\infty \leq h^{-d/2} \) for the case of partitioning-based estimators (see SA §SA-4.2 for details). It follows that \( \{Z_j(x) : x \in \mathcal{X}\} \) is an unconditional Gaussian process that approximates the distribution of the whole \( t \)-statistic process \( \{T_j(x) : x \in \mathcal{X}\} \). This second and final part of our coupling strategy is formalized in Theorem SA-4.6 in the SA.

The crux in establishing (3.4) is to prove precise (rate) control on \( \|\bar{\Sigma}_j^{1/2} - \Sigma_j^{1/2}\|, j = 0, 1, 2, 3 \). Both \( \bar{\Sigma}_j \) and \( \Sigma_j \) are symmetric and positive semi-definite. Further, for \( j = 0, 1 \), \( \lambda_{\min}(\Sigma_j) \geq h^d \) for
generic partitioning-based estimators under our assumptions (see the SA, §SA-2), 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 immediately obtain unconditional coupling from conditional coupling without additional rate restrictions.

However, for \(j = 2, 3\) this bound cannot be used in general because \(p\) and \(\tilde{p}\) are typically not linearly independent, and hence \(\Sigma_j\) will be singular. To circumvent this technical problem, we employ the following 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 3.1 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 3.4 further below). Furthermore, for certain partitioning-based series estimators it is still possible to use (3.5) even when \(j = 2, 3\) is considered (see Remark 3.1 below).

Putting together the steps above, we obtain a valid distributional approximation for the process \(\{\hat{T}_j(x) : x \in \mathcal{X}\}\) using \(\{Z_j(x) : x \in \mathcal{X}\}\), for \(j = 0, 1, 2, 3\). Schematically, our approach is summarized as follows:
\[ \hat{T}_j(\cdot) \approx p \, t_j(\cdot) \overset{d}{=} t'_j(\cdot) \approx p \, z_j(\cdot) \overset{dX}{=} \tilde{Z}_j(\cdot) \approx p \, Z_j(\cdot), \]
where \(\approx\) refers to uniform convergence in probability of (3.2), (3.3), and (3.4), \(\overset{d}{=}\) refers to equality in distribution in a sufficiently rich probability space, and \(\overset{dX}{=}\) refers to equality in distribution conditional on \(X\) in a sufficiently rich probability space. To save notation, we denote this approximation by \(\hat{T}_j(\cdot) =_d Z_j(\cdot) + o_p(r_n^{-1})\) in \(\mathcal{L}^\infty(\mathcal{X})\), where \(\mathcal{L}^\infty(\mathcal{X})\) refers to the set of all uniformly bounded real functions on \(\mathcal{X}\) equipped with uniform norm. The following theorem gives the precise result.

**Theorem 3.1** (Strong Approximation). Let Assumptions 1–4 hold and let \(r_n > 0\) be a non-
vanishing sequence. Assume one of the following holds:

\[(i) \quad \sup_{x \in X} E[|\varepsilon_i|^3 \exp(|\varepsilon_i|)|x_i = x] < \infty \quad \text{and} \quad \frac{(\log n)^4}{nh^d} = o(r_n^{-2}), \quad \text{or} \]

\[(ii) \quad \sup_{x \in X} E[|\varepsilon_i|^{2+\nu}|x_i = x] < \infty, \quad \text{for a } \nu > 0, \quad \text{and} \quad \frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2+2\nu}{2+\nu}}}{nh^d} = o(r_n^{-2}). \]

Furthermore,

- if \(j = 0\), assume \(nh^{d+2m} = o(r_n^{-2})\);
- if \(j = 1\), assume Assumption 5 holds and \(nh^{d+2m+2\theta} = o(r_n^{-2})\);
- if \(j = 2, 3\), assume Assumption 5 holds, \(nh^{d+2m+2\theta} = o(r_n^{-2})\), and \((\log n)^{3/2}/\sqrt{nh^d} = o(r_n^{-2})\).

Then, for each \(j = 0, 1, 2, 3\), there exists a sequence of \(K_j\)-dimensional multivariate standard Normal vectors \(\{N_{K_j}\}\) in an appropriately enlarged probability space such that \(\hat{T}_j(\cdot) = dZ_j(\cdot) + o_P(r_n^{-1})\) in \(L^\infty(X)\), where \(Z_j(x)\) is given in (3.4).

These strong approximation results for partitioning-based least squares estimation appear to be new in the literature. An alternative unconditional strong approximation for general series estimators is obtained by [3] for the case of undersmoothing inference (\(j = 0\) in Theorem 3.1). Their proof is based on Yurinskii [62]'s coupling inequality that controls the convergence rate of partial sums in terms of Euclidean norm, leading to the rate restriction \(r_n^6/(nh^{5d}) \asymp r_n^6K^5/n \to 0\), up to \(\log n\) terms, which does not depend on \(\nu\) (i.e., the rate restriction is the same regardless of the order of bounded conditional moments of \(\varepsilon_i\)). In contrast, our results employ KMT-type (conditional) coupling together with (3.4) and (3.5), 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 \(r_n^6/(n^{3\nu/(2+\nu)}h^{3d}) \asymp r_n^6K^3/n^{3\nu/(2+\nu)} \to 0\), 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 available in the literature. Under the sub-exponential conditional moment restriction in Theorem 3.1 this can be relaxed all the way to \(K/n \to 0\), up to \(\log n\) terms, which appears to be a minimal condition. Note that this is for the entire \(t\)-statistic process. In addition, Theorem 3.1 gives novel strong approximation results for robust bias-corrected \(t\)-statistic processes.

It is also possible to establish uniform distributional approximations for series/least squares estimators in particular, and series-based and kernel-based statistics more generally, employing the
coupling techniques developed by Massart [46], Koltchinskii [43], and Rio [51], as recently shown by [21] and [22]. This approach is different from ours in that it approximates a sequence of multi-dimensional stochastic processes directly, treats the underlying dimensionality of the problem as $d+1$, and imposes some specific high-level conditions on the underlying probability distribution. As a consequence, this coupling approach requires stronger side rate restrictions and yields slower approximation rates. In contrast, our coupling approach effectively employs a univariate (conditional) coupling strategy, which avoids dimensionality issues and thus achieves shaper results.

**Remark 3.1** (Square-root Convergence and Improved Rates). The additional restriction imposed in Theorem 3.1 for $j = 2, 3$, that $(nh^d)^{-1/2}(\log n)^{3/2} = o(r_n^{-2})$, 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 $\hat{p}$, which obeys Assumption 3, such that $(p(\cdot)',\hat{p}(\cdot)')' = \Upsilon\hat{p}(\cdot)$. In other words, the two bases $p$ and $\hat{p}$ can be expressed in terms of another basis $\hat{p}$ without linear dependence. Then, a positive lower bound holds for $\lambda_{\text{min}}(\Sigma_j), j = 2, 3$, implying that the bound (3.5) can be used instead of (3.6).

For example, for generalized regressograms and $B$-splines with equal knot placements for $p$ and $\hat{p}$, a natural choice of $\hat{p}$ is simply a higher-order polynomial basis on the same partition. Since each function in $p$ and $\hat{p}$ 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. Therefore, in these special but useful cases, Theorem 3.1 holds without the additional rate condition $(nh^d)^{-1/2}(\log n)^{3/2} = o(r_n^{-2})$ for all $j = 0, 1, 2, 3$. See Appendix A and the SA.

### 3.2 Implementation

To conduct inference using the results in Theorem 3.1 it is necessary to approximate the infeasible standardized Gaussian processes $\{Z_j(x) : x \in \mathcal{X}\}$. In this section, we present a plug-in approach and a bootstrap approach to do just that. To be precise, we construct (feasible) Gaussian processes $\{\hat{Z}_j(x) : x \in \mathcal{X}\}$, with distributions known conditional on the data $(y, X)$, such that there exists a copy $\hat{Z}_j(\cdot)$ of $Z_j(\cdot)$ in a sufficiently rich probability space and (i) $\hat{Z}_j(\cdot) =_d Z_j(\cdot)$ conditional on the
data \((y, X)\) and (ii) for some positive non-vanishing sequence \(r_n\) and for all \(\eta > 0\),

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

where \(\mathbb{P}^*[\cdot] = \mathbb{P}[\cdot | y, X]\) denotes the probability operator conditional on the data. When such a feasible process exists, we write \(\hat{Z}_j(\cdot) = d_r Z_j(\cdot) + o_P(r_n^{-1})\) in \(L^\infty(\mathcal{X})\). From a practical perspective, sampling from \(\hat{Z}_j(\cdot)\), conditional on the data, is possible and provides a valid distributional approximation.

The first construction is a direct plug-in approach using the conclusion of Theorem 3.1. All the unknown objects in (3.4) can be consistently estimated using the same estimators already used in the feasible \(t\)-statistics, so we simply show that sampling from

\[
\hat{Z}_j(x) = \frac{\hat{\gamma}_{q,j}(x)' \hat{\Sigma}_j^{1/2} \mathbf{N}_{K_j}}{\sqrt{\hat{\Omega}_j(x)}}, \quad x \in \mathcal{X}, \quad j = 0, 1, 2, 3, \tag{3.7}
\]

conditional on the data, gives a valid distributional approximation. Sampling is from a multivariate standard Gaussian of dimension \(K_j\), not \(n\).

**Theorem 3.2** (Plug-in Approximation). For each \(j = 0, 1, 2, 3\), let the conditions of Theorem 3.1 hold, as appropriate. Furthermore, for \(j = 2, 3\), the rate restrictions in (i) and (ii) are slightly strengthened, respectively, to \((\log n)^{5/2}/\sqrt{nhd} = o(r_n^{-2})\) and \(n^{1/2}(\log n)^{1+\nu}/\sqrt{nhd} = o(r_n^{-2})\). Then, for each \(j = 0, 1, 2, 3\), \(\hat{Z}_j(\cdot) = d_r Z_j(\cdot) + o_P(r_n^{-1})\) in \(L^\infty(\mathcal{X})\), where \(\hat{Z}_j(\cdot)\) is given in (3.7).

The additional rate condition for approaches \(j = 2, 3\) strengthens that of Theorem 3.1 only by logarithmic factors. Moreover, if the structure discussed in Remark 3.1 holds, then this condition can be dropped.

As a second method for feasible inference, we show that the wild (or multiplier) bootstrap can also provide a feasible approximation to \(\{Z_j(x) : x \in \mathcal{X}\}, j = 0, 1, 2, 3\). This is based on resampling the residuals conditionally on the data. For zero-mean unit-variance i.i.d. bounded random variables \(\{\omega_i : 1 \leq i \leq n\}\) independent of the data, we construct bootstrapped \(t\)-statistics:

\[
\hat{z}_j^*(x) = \frac{\hat{\gamma}_{q,j}(x)' \sqrt{n} \left[ \prod_j \omega_i \hat{e}_{i,j} \right]}{\sqrt{\hat{\Omega}_j(x)}}, \quad x \in \mathcal{X}, \quad j = 0, 1, 2, 3, \tag{3.8}
\]
where $\hat{\varepsilon}_{i,j}$ is defined in (3.1), and the bootstrap Studentization $\hat{\Omega}^*_j(x)$ is constructed using $\hat{\Sigma}^*_j = \mathbb{E}_{n}[\Pi_j(x_i)\Pi_j(x_i)'/\omega_i^2\hat{\varepsilon}^2_{i,j}]$.

**Theorem 3.3** (Bootstrap Approximation). For each $j = 0, 1, 2, 3$, let the conditions of Theorem 3.1 hold, as appropriate. Furthermore,

- if $j = 0, 1$, the rate restriction in (ii) is slightly strengthened to $n^{2/(5 + \nu)}(\log n)^2/(nh^d) = o(r_n^{-2})$;
- if $j = 2, 3$, the rate restrictions in (i) and (ii) are slightly strengthened, respectively, to $(\log n)^{5/2}/\sqrt{nh^d} = o(r_n^{-2})$ and $n^{1/(4 + 2\nu)}(\log n)^{4 + 3\nu}/\sqrt{nh^d} = o(r_n^{-2})$.

Then, for each $j = 0, 1, 2, 3$, $\hat{\gamma}^*_j(\cdot) = d^* Z_j(\cdot) + o_P(r_n^{-1})$ in $L^\infty(\mathcal{X})$, where $\hat{\gamma}^*_j(\cdot)$ is given in (3.8).

The additional rate restriction imposed in this theorem for $j = 0, 1$ strengthens those already imposed in Theorem 3.1 only by logarithmic factors. This condition is needed for technical reasons only (i.e., to ensure that certain bootstrap-based matrices converge in probability; see the SA).

**Remark 3.2** (Alternative Plug-In Approximation). In the SA (§SA-4.3), we also prove validity of another plug-in approach, one that is based on approximating the intermediate (conditionally on $X$ Gaussian) process $\{z_j(x) : x \in \mathcal{X}\}$. Specifically, we show that if $\hat{\sigma}^2(x)$ is a uniformly consistent estimator of $\sigma^2(x)$ at the appropriate rate, then sampling conditional on the data from

$$\hat{\gamma}_j(x) = \frac{\hat{\gamma}^{a,j}(x)'}{\sqrt{\hat{\Omega}_j(x)}} \mathbb{E}_{n}[\Pi_j(x_i)\hat{\sigma}(x_i)\zeta_i]$$

gives a valid distributional approximation, where sampling is based on $n$ standard Gaussian variables.

**3.3 Application: Confidence Bands**

A natural application of Theorems 3.1, 3.2, and 3.3 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

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

where $\hat{T}_j(x)$ is constructed using $\hat{\Sigma}^*_j = \mathbb{E}_{n}[\Pi_j(x_i)\Pi_j(x_i)'/\omega_i^2\hat{\varepsilon}^2_{i,j}]$. 


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

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

The following theorem 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 3.4** (Confidence Bands). Let the conditions of Theorem 3.1 hold with $r_n = \sqrt{\log n}$. If the corresponding conditions of Theorem 3.2 for each $j = 0, 1, 2, 3$ hold, then

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

If the corresponding conditions of Theorem 3.3 for each $j = 0, 1, 2, 3$ hold, then

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

[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. This argument 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., $j = 0$ in the Theorem above). 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 3.4 under the side rate condition $K/n^{1-2/(2+\nu)} \asymp n^{-1+2/(2+\nu)} h^{-d} = o(1)$, up to log $n$ terms (with $q = 2 + \nu$ in their notation). In comparison, our result for $j = 0$ in Theorem 3.4, under the same moment conditions, requires exactly the same side condition, up to log $n$ terms. However, comparing Theorems 3.1 and 3.4 shows 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. The same result holds for sub-exponential moments, where the rate condition becomes minimal: $K/n \asymp \frac{1}{nh^d} = o(1)$, up to log $n$ factors. We are able to achieve such sharp rate restrictions and approximation rates by exploiting the specific features of the estimator together
with the new coupling approach discussed above, and with the help of the key anti-concentration idea introduced by [17]. In addition, Theorem 3.4 gives new inference results for bias-corrected estimators \((j = 1, 2, 3)\).

Finally, the strong approximation result for the entire t-statistic processes given in Theorem 3.1, and related technical results given in the SA, can also be used to construct other types of confidence bands for the regression function and its derivatives; e.g., [33, 32]. We do not elaborate further on this to conserve space.

4 Integrated Mean Squared Error

From a practical point of view, the methods above give a complete implementation of partitioning-based nonparametric inference for a user-chosen partition. To fill the final gap in implementation, we present IMSE expansions for partitioning-based least squares estimators. We first give a very general result, which then we specialized to a more detailed result for the special case of a rectangular partition (a tensor product of marginally-formed intervals).

A chief advantage of robust bias corrected inference is that (I)MSE-optimal tuning parameters (and related choices such as those obtained from cross-validation) are valid for inference, which is not the case for the standard approach unless ad-hoc undersmoothing is used. This allows researchers to combine an optimal estimate of the function, \(\widehat{\partial^q \mu_0}(\cdot)\) based on the IMSE-optimal \(h_{\text{IMSE}} \asymp n^{-1/(2m+d)}\), with inference based on the same tuning parameter choices (and hence employing the same partitioning scheme). For this reason, we present only the IMSE results for the standard estimator \(\widehat{\partial^q \mu_0}(\cdot)\); similar results can be shown for the bias-corrected estimators.

Our general result, which holds for any partition \(\Delta\) satisfying Assumption 2, is the following.

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

\[
\int_X \mathbb{E}[(\partial^q_{\Delta} \mu_0(x) - \partial^q \mu(x))^2 | X] w(x) \, dx = \frac{1}{n} \left( \gamma_{\Delta, q} + o_P(h^{-d-2[q]}) \right) + \left( \beta_{\Delta, q} + o_P(h^{2m-2[q]}) \right)
\]
where

\[
\begin{align*}
\mathcal{V}_{\Delta, \mathbf{q}} &= \text{trace} \left( \Sigma_0 \int_{\mathcal{X}} \gamma_{\mathbf{q},0}(\mathbf{x}) \gamma_{\mathbf{q},0}(\mathbf{x})' w(\mathbf{x}) d\mathbf{x} \right) \asymp h^{-d-2[q]}, \\
\mathcal{B}_{\Delta, \mathbf{q}} &= \int_{\mathcal{X}} \left( \mathcal{B}_{m, \mathbf{q}}(\mathbf{x}) - \gamma_{\mathbf{q},0}(\mathbf{x})' \mathbb{E}[\mathbf{p}(\mathbf{x}_i) \mathcal{B}_{m,0}(\mathbf{x}_i)] \right)^2 w(\mathbf{x}) d\mathbf{x} \lesssim h^{2m-2[q]}.
\end{align*}
\]

This theorem shows that the leading term in the integrated (and pointwise) variance of \( \hat{\partial \mathbf{q} \mu_0}(\mathbf{x}) \) 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. It is easy to see that this rate bound is sharp in general.

The quantities \( \mathcal{V}_{\Delta, \mathbf{q}} \) and \( \mathcal{B}_{\Delta, \mathbf{q}} \) are two sequences of non-random constants 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 (\( \mathcal{B}_{\Delta, \mathbf{q}} \neq 0 \)), Theorem 4.1 implies that the IMSE-optimal mesh size \( h_{\text{IMSE}} \) is proportional to \( n^{-1} h^{-d-2[q]} \), or equivalently, the IMSE-optimal number of series terms \( K_{\text{IMSE}} \propto n^{\frac{d}{2m+d}} \). 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 a IMSE sense. It is possible to consider the optimization problem

\[
\min_{\Delta \in \mathcal{D}} \left\{ \frac{1}{n} \mathcal{V}_{\Delta, \mathbf{q}} + \mathcal{B}_{\Delta, \mathbf{q}} \right\}
\]

as a way of selecting an “optimal” partitioning scheme among the class of partitioning schemes \( \mathcal{D} \).

Theorem 4.1 generalizes prior work substantially. Existing results cover only special cases, such as generalized regressograms [12] or splines [1, 65, 66] on rectangular partitions only, and often restricting to \( d = 1 \) or \([q] = 0\). To the best of our knowledge, covering nonrectangular partitions and other series functions such as wavelets is new to the literature. Furthermore, uniform inference results on series methods (as cited above) often contain no practical guidance.

To illustrate the usefulness of this result in applications, we will consider the special case of a rectangular partitioning where the “tuning parameter” \( \Delta \) reduces to the vector of partitioning knots \( \mathbf{\kappa} = (\kappa_1, \ldots, \kappa_d)' \), where \( \kappa_\ell \) is the number of subintervals used for the \( \ell \)-th covariate. We further assume that \( \Delta \) and \( \mathbf{p}(\cdot) \) obey the following regularity conditions, so that the limiting constants in the IMSE approximation can be characterized. Appendix A verifies this condition for splines,
wavelets, and generalized regressograms. Appendix B gives more details and further results for rectangular partitions.

**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 rectangular partition. For $x \in [0,1]^d$, let $\delta_x \in \Delta$ be the cell containing $x$, denoted $\delta_x = \{ t_{\ell,x} \leq x_\ell \leq t_{\ell,x+1}, 1 \leq \ell \leq d \}$ with $l_x < \kappa_\ell$ (c.f. Equation (2.1) and surrounding discussion). Let $b_x = (b_{x,1}, \ldots, b_{x,d})$ collect the interval lengths $b_{x,\ell} = |t_{\ell,x+1} - t_{\ell,x}|$. 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

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

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 $J + \bar{J} < \infty$ groups such that $\{\delta_{\tau_{s,k}}\}_{k=1}^{K_s}$, $s = 1, \ldots, J + \bar{J}$, $\sum_{s=1}^{J+\bar{J}} K_s = K$, and the following conditions hold: (i) For all $1 \leq s \leq J$, $\{\delta_{\tau_{s,k}}\}_{k=1}^{K_s}$ are pairwise disjoint and $\text{vol} \left([0,1]^d \setminus \bigcup_{k=1}^{K_s} \delta_{\tau_{s,k}}\right) = o(1)$; and (ii) for all $J + 1 \leq s \leq J + \bar{J}$, $\text{vol} \left(\bigcup_{k=1}^{K_s} \delta_{\tau_{s,k}}\right) = o(1)$.

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 rectangular partitions, instead of the more general diameter used in Section 2. Appendix A gives examples. 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 $J + \bar{J}$ groups. The supports of functions in each of the first $J$ groups constitute “almost” complete covers of $X$. 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 generalized regressograms 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 $K \approx h^{-d} \to \infty$.

Under this assumption, we have the following result for $\hat{\mu}_0(x)$.

**Corollary 1 (Asymptotic IMSE).** Suppose that the conditions in Theorem 4.1 and Assumption 6 hold. Then, for $[q] = 0$,

$$
\gamma_{\kappa,0} = \left( \prod_{\ell=1}^{d} \kappa_\ell \right) \gamma_0 + o(h^{-d}), \quad \gamma_0 = J \int_{[0,1]^d} \frac{\sigma^2(x)}{f(x)} \left( \prod_{\ell=1}^{d} g_\ell(x) \right) w(x) \, dx,
$$

and, provided that (2.4) holds,

$$
\mathcal{B}_{\kappa,0} = \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1+u_2)} \mathcal{B}_{u_1, u_2, 0} + o(h^{2m}), \quad \mathcal{B}_{u_1, u_2, 0} = \eta_{u_1, u_2, 0} \int_{[0,1]^d} \frac{\partial^{u_1} \mu(x) \partial^{u_2} \mu(x)}{g(x)^{u_1+u_2}} w(x) \, dx.
$$

The bias approximation requires the approximate orthogonality condition (2.4) which is satisfied by $B$-splines, wavelets, and piecewise polynomials (see Appendix A). It appears to be an open question whether $\gamma_{\kappa,q}$ and $\mathcal{B}_{\kappa,q}$ converge to a well-defined limit when general basis functions are considered. [12] showed convergence to a well-defined limits for the special case of (what we here call) generalized regressograms, but their result is not easy to extend to cover other bases functions without imposing $q = 0$ and the approximate orthogonality condition (2.4). See Appendix B for more details and discussion. This is the reason why Corollary 1 only considers $q = 0$ (i.e., the IMSE of $\hat{\mu}_0(x)$) and imposes condition (2.4).

Corollary 1 immediately justifies the IMSE-optimal choice of number of knots:

$$
\kappa_{\text{IMSE},0} = \arg \min_{\kappa \in \mathbb{Z}^d_+} \left\{ \frac{1}{n} \left( \prod_{\ell=1}^{d} \kappa_\ell \right) \gamma_0 + \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1+u_2)} \mathcal{B}_{u_1, u_2, 0} \right\}.
$$
and, in particular, when the same number of knots is used in all margins,

\[
\kappa_{\text{IMSE},0} = \left[ \frac{2m \sum_{u_1,u_2 \in \Lambda_m} R_{u_1,u_2,0}}{dV_0} \right]^{\frac{1}{2m+d}} \frac{1}{n^{\frac{1}{2m+d}}}
\]

Data-driven versions of this IMSE-optimal choice, and extensions to derivative estimation, are discussed in the SA (§SA-6) and fully implemented in our companion general-purpose R package lspartition [13].

5 Simulations

This section reports a summary of the results from a simulation study investigating the finite sample performance of our methods. The full set of results is available in the SA (§SA-7). All numerical results were obtained using our companion R package lspartition [13].

For simplicity and comparability, we considered the three univariate \((d = 1)\) data generating processes recently used in [36], and also a bivariate \((d = 2)\) model used in [12]. Presently we report the results from only one, deferring the rest to the SA. 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 \(x_i \sim \text{U}[0,1]\) and \(\varepsilon_i \sim \text{N}(0,1)\), independent of each other. We consider 5,000 simulated datasets with \(n = 1,000\) each time. We use linear splines \((m = 2)\) to form the point estimator \(\hat{\mu}_0(x)\), and quadratic splines \((\tilde{m} = 3)\) for bias correction. We investigate both evenly- and quantile-spaced knot placements, and for simplicity point estimators and bias correction employ the same partitioning scheme \((\Delta = \tilde{\Delta})\).

The results are presented in Tables 1–3. Tables 1 and 2 present pointwise estimation and inference results for three fixed evaluation points. These tables include (simulated) root mean squared error (RMSE) for point estimators, and coverage rate (CR) and average interval length (IL) for 95% nominal confidence intervals. Table 3 presents the main uniform inference results. This table includes the three measures previously used by [36]: proportion of values covered with probability at least 95\% (CP), average coverage errors (ACE), and average width of confidence band (AW). The more stringent criterion of uniform coverage rate (UCR) is also presented. In all cases, the tables report results for both evenly-spaced and quantile-spaced partitions, employing either the
infeasible IMSE-optimal size choice ($\kappa_{\text{IMSE}}$), a rule-of-thumb estimate ($\hat{\kappa}_{\text{ROT}}$), or a direct plug-in estimate ($\hat{\kappa}_{\text{DPI}}$). See §SA-6 in SA for more details. Finally, each table reports all four (estimation and) inference methods discussed in this paper indexed by $j = 1, 2, 3, 4$.

All the numerical findings are consistent with our theoretical results. To conserve space, we only summarize the overarching findings: robust bias-correction seems to perform quite well, always delivering close-to-correct coverage, both pointwise (Tables 1–2) and uniformly (Table 3). The numerical performance of our rule-of-thumb (ROT) and direct plug-in (DPI) knot selection procedures for rectangular partitions also worked well in our simulation study. More details and additional results are reported in §SA-7 in the SA.

6 Conclusion

We presented new asymptotic results for partitioning-based least squares regression estimators. The first main contribution gave a strong approximations to both standard and bias-corrected $t$-statistic processes indexed by $x \in \mathcal{X}$, with seemingly minimal rate restrictions. The second main contribution gave a general IMSE expansion for partitioning-based estimators, which generalized previous special cases available in the literature as well as gave new results previously unavailable. These main results were illustrated using three popular special cases: $B$-splines, compact-supported wavelets, and the generalized regressogram (piecewise polynomial regression). All methods discussed herein are available in the general purpose R package `lspartition` [13].

In addition, the SA gives several additional technical results that may be of independent interest, including a general (conditional) coupling result (Lemma SA-4.5), pointwise and uniform stochastic linearization useful in semiparametric settings (§SA-3 and §SA-4.1; see, in particular, Remark SA-3.1), other technical lemmas (§SA-2), theoretical comparisons between bias-correction approaches (§SA-5.2), and a discussion of the relationship between $B$-Splines and polynomials (§SA-5.1). Finally, in research underway, we are applying our proposed coupling strategy to more general ultra-high-dimensional linear least squares estimation and related settings, where $d \to \infty$ (with possibly $d \gg n$).
Appendix A  Examples of Popular Partitioning-Based Estimators

We now illustrate how Assumptions 3, 4, and 5 are verified for popular basis choices, B-splines, wavelets, and piecewise polynomials (i.e., generalized regressogram), and show when (2.4) holds. Note well that even in the cases when (2.4) holds, we will not make use of this property neither herein nor in our software implementation [13], since both bias terms of Lemma 2.1 may be important in finite samples.

The first three subsections illustrate each of these assuming rectangular support, \( \mathcal{X} = \otimes_{\ell=1}^{d} \mathcal{X}_\ell \subset \mathbb{R}^d \), and correspondingly use rectangular partitions. We note that these three in fact overlap in the special cases of \( m = 1 \) on a rectangular partition, leading to so-called zero-degree splines, Haar wavelets, and regressograms (piecewise constant). The final subsection considers a general partition. However, recall that whenever \( \Delta \) covers only strict subset of \( \mathcal{X} \), all our results hold on that subset; the first three subsections may also be interpreted in this light.

Some additional notation is required. For a multi-index let \( q! = q_1! \cdots q_d! \). Denote the \( d \)-vector of ones by \( \mathbf{1} = (1, \ldots, 1)' \). When \( \Delta \) is a rectangular partition, \( \mathbf{b}_\mathcal{X} \) is the vector collecting the interval lengths of \( \delta_\mathcal{X} \) as defined in Assumption 6, and \( t^{L}_\mathcal{X} \) is used to denote the start point of \( \delta_\mathcal{X} \). For \( \ell = 1, \ldots, d \), for a generic cell \( \delta_{\ell} \) in a rectangular partition (see Equation (2.1)), we write \( b_{\ell} = t_{\ell,t} - t_{\ell,t-1} \), \( b_{\ell} = \max_{0 \leq t \leq \kappa_{\ell}} b_{\ell} \), \( b_{\ell} = \max_{1 \leq \ell \leq d} b_{\ell} \) (\( h \) by Assumption 2). Moreover, \( \|g\|_{L^\infty(\mathcal{X})} = \text{ess sup}_{x \in \mathcal{X}} |g(x)| \) for a real-valued function \( g(x) \), and \( \mathcal{C}^s(\mathcal{X}) \) denotes the space of \( s \)-times continuously differentiable functions on \( \mathcal{X} \). Finally, \( \otimes \) denotes the entrywise division operator (Hadamard division) for matrices.

A.1  B-Splines on Rectangular Partitions

A univariate spline is a piecewise polynomial satisfying certain smoothness constraints. For some integer \( m_\ell \geq 2 \), let \( S_{\Delta_\ell,m_\ell} \) be the set of splines of order \( m_\ell \) with univariate partition \( \Delta_\ell \). Then

\[
S_{\Delta_\ell,m_\ell} = \left\{ s \in C^{m_\ell-2}(\mathcal{X}_\ell) : s(x) \text{ is a polynomial of degree } (m_\ell-1) \text{ on each subinterval } [t_{\ell,t}, t_{\ell,t+1}] \right\},
\]

and hence \( S_{\Delta_\ell,m_\ell} \) is a vector space and can be spanned by many equivalent representing bases. B-splines as a local basis are well studied in literature and enjoy many nice properties [48, 54].

Define an extended knot sequence \( \Delta_{\ell,e} \) such that

\[
t_{\ell,-m_\ell+1} = t_{\ell,-m_\ell+2} = \cdots = t_{\ell,0} = t_{\ell,1} = \cdots = t_{\ell,\kappa_\ell-1} = \cdots = t_{\ell,\kappa_\ell+m_\ell-1},
\]

Then, the \( m_\ell \)-th order B-spline with knot sequence \( \Delta_{\ell,e} \) is

\[
p_{\ell,m_\ell}(x_\ell) = (t_{\ell,t} - t_{\ell,t-m_\ell})(t_{\ell,t-m_\ell} - \cdots - t_{\ell,0})^{m_\ell-1}, \quad l_\ell = 1, \ldots, K_\ell = \kappa_\ell + m_\ell - 1,
\]

where \( (a)_{+} = \max\{a,0\} \) and \( [t_1, t_2, \ldots, t_s] \) \( \mu(t, x) \) denotes the divided difference of \( \mu(t, x) \) with respect to \( t \), given a sequence of knots \( t_1 \leq t_2 \leq \cdots \leq t_s \).

When there is no chance of confusion, we shall write \( p_{\ell}(x_\ell) \) instead of \( p_{\ell,m_\ell}(x_\ell) \). Accordingly, the space of tensor product polynomial splines of order \( m = (m_1, \ldots, m_d) \) with partition \( \Delta \) is spanned by the tensor products of univariate B-splines

\[
S_{\Delta,m} = \otimes_{\ell=1}^{d} S_{\Delta_\ell,m_\ell} = \text{span}\{p_{l_1}(x_1)p_{l_2}(x_2) \cdots p_{l_d}(x_d) : l_1 = 1, \ldots, K_1 \cdots K_d = 1\}.
\]

We have a total of \( K = \prod_{\ell=1}^{d} K_\ell \) basis functions. The order of univariate basis could vary across dimensions, but for simplicity we assume that \( m_1 = \cdots = m_d = m \) and hence we write \( S_{\Delta,m} = \text{span}\{p_{1}(x_1)p_{1}(x_2) \cdots p_{1}(x_d) : l_1 = 1, \cdots, K_1 \cdots K_d = 1\} \).
\( S_{\Delta,m} \). Also, let \( p_{l_1\ldots l_d}(x) = p_{l_1}(x_1)\cdots p_{l_d}(x_d) \) to further simplify notation.

Arrange this set of basis functions by first increasing \( l_d \) from 1 to \( K_d \) with other \( l'_\ell \)'s fixed at 1 and then increasing \( l_\ell \) sequentially. As a result, we construct a one-to-one correspondence \( \varphi \) mapping from \( \{(l_1,\ldots,l_d) : 1 \leq l_\ell \leq K_\ell, \ell = 1,\ldots,d\} \) to \( \{1,\ldots,K\} \). Then, we can write \( p_k(x) = p_{\varphi^{-1}(k)}(x) \), \( k = 1,\ldots,K \), which is consistent with our notation in Section 1.

The following lemma shows that Assumptions 3 and 4 hold for B-splines.

**Lemma A.1 (B-Splines Estimators).** Let \( p(x) \) be a tensor-product B-Spline basis of order \( m \), and suppose Assumptions 1 and 2 hold with \( m \leq S \). Then:

(a) \( p(x) \) satisfies Assumption 3.

(b) If, in addition,
\[
\max_{0 \leq l \leq \kappa \ell - 2} |b_{l,l+1} - b_{l,l}| = O(b^{1+\nu}), \quad \ell = 1,\ldots,d, \tag{A.1}
\]
then Assumption 4 holds with \( \varsigma = m - 1 \) and
\[
\mathcal{B}_{m,\varsigma}(x) = - \sum_{u \in \Lambda_m} \frac{\partial^{u} \mu(x) h_{x}^{m-|s|} b_{x}^{u-\varsigma}}{(u-\varsigma)!} b_{x}^{m-|s|} \mathcal{B}_{u,\varsigma}^{S}((x-t^{L}) \otimes b_{x})
\]
where \( \Lambda_m = \{ u \in \mathbb{Z}^d_+ : |u| = m, \text{ and } u_{\ell} = m \text{ for some } 1 \leq \ell \leq d \} \) and \( B_{u}^{S}(x) \) is the product of univariate Bernoulli polynomials; that is, \( B_{u}^{S}(x) := \prod_{\ell=1}^{d} B_{u_{\ell}}(x_{\ell}) \) with \( B_{u_{\ell}}(\cdot) \) being the \( u_{\ell} \)-th Bernoulli polynomial and \( B_{u}^{S} = 0 \) if \( u \) contains negative elements. Furthermore, Equation (2.4) holds.

(c) Let \( \bar{p}(x) \) be a tensor-product B-Spline basis of order \( \bar{m} > m \) on the same partition \( \Delta \), and assume \( \bar{m} \leq S \). Then Assumption 5 is satisfied.

Equation (A.1) gives a precise definition of the strong quasi-uniform condition on the partition scheme alluded to in Remark 2.1 that is required for B-splines. Assumption 2 requires that the volume of all cells vanish at the same rate but allows for any constant proportionality between neighboring cells. Presently, cells are further restricted to asymptotically be the same volume, and further, a specific rate is required that is related to the smoothness of \( \mu(\cdot) \). Note that, for example, equally spaced knots satisfy this conditional trivially. For other schemes, additional work may be needed. Under (A.1), [2] obtained an expression for the leading asymptotic error of univariate splines, which was later used by [65] and [66], among others. Lemma A.1 extends previous results to the multi-dimensional case, in addition to showing that the high-level conditions in Assumption 3 and 4 hold for B-Splines.

When (A.1) fails, it may be possible to obtain results, but additional cumbersome notation would be needed and the results may be less useful. However, it is straightforward to verify that without (A.1), there will still exist \( s^{*} \in S_{\Delta,m} \) such that for all \( [s] \leq \varsigma \),
\[
\sup_{x \in \mathcal{X}} |\partial^{s} \mu - \partial^{s} s^{*}| \lesssim h^{m-|s|}, \tag{A.2}
\]
where recall that \( S_{\Delta,m} \) denotes the linear span of \( p(x;\Delta,m) \). However, this result does not allow for bias correction or IMSE expansion.

Finally, \( \Lambda_m \) contains only the multi-indices corresponding to \( m \)-th order partial derivatives of \( \mu(x) \). This is due to the fact that, as a variant of polynomial approximation, the total order of tensor product B-splines is not fixed at \( m \), i.e. some higher-order components are retained that
approximate terms with $m$-th order cross partial derivatives. This feature distinguishes tensor product splines from the multivariate splines which do control the total order of approximating basis.

### A.2 Wavelets on Rectangular Partitions

Our results apply to compactly supported wavelets, such as Cohen-Daubechies-Vial wavelets [24]. For more background details see [47, 37, 23], and references therein.

To describe these estimators, we first introduce the general definition of wavelets and then show that a large class of orthogonal wavelets satisfy our high-level assumptions. For some integer $m \geq 1$, we call $\phi$ a (univariate) scaling function or “father wavelet” of degree $m - 1$ if (i) $\int_{\mathbb{R}} \phi(x) \, dx = 1$, (ii) $\phi$ and all its derivatives up to order $m - 1$ decrease rapidly as $|x| \to \infty$, and (iii) $\{\phi(x_l - l) : l \in \mathbb{Z}\}$ forms a Riesz basis for a closed subspace of $L_2(\mathbb{R})$. A real-valued function $\psi$ is called a (univariate) “mother wavelet” of degree $m - 1$ if (i) $\int_{\mathbb{R}} x^v \psi(x) \, dx = 0$ for $0 \leq v \leq m - 1$, (ii) $\psi$ and all its derivatives decrease rapidly as $|x| \to \infty$, and (iii) $\{2^{s/2}(2^s x - l) : s, l \in \mathbb{Z}\}$ forms a Riesz basis of $L_2(\mathbb{R})$.

We restrict our attention to orthogonal wavelets with compact support. The father wavelet $\phi$ and the mother wavelet $\psi$ are both compactly supported, and for any integer $s_0 \geq 0$, any function in $L_2(\mathbb{R})$ admits the following $(m - 1)$-regular wavelet multiresolutional expansion:

$$
\mu(x_\ell) = \sum_{l=-\infty}^{\infty} a_{s_0 l} \phi_{s_0 l}(x_\ell) + \sum_{s=s_0}^{\infty} \sum_{l=-\infty}^{\infty} b_{s l} \psi_{s l}(x_\ell), \quad x_\ell \in \mathbb{R},
$$

where

$$
a_{s l} = \int_{\mathbb{R}} \mu(x) \phi_{s l}(x) \, dx, \quad \phi_{s l}(x) = 2^{s/2} \phi(2^s x - l),
$$

$$
b_{s l} = \int_{\mathbb{R}} \mu(x) \psi_{s l}(x) \, dx, \quad \psi_{s l}(x) = 2^{s/2} \psi(2^s x - l),
$$

and $\{\phi_{s_0 l}, l \in \mathbb{Z}; \psi_{s l}, s \geq s_0, l \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$. Accordingly, to construct an orthonormal basis of $L_2([0, 1])$, one can pick those basis functions supported in the interior of $[0, 1]$ and add some boundary correction functions. For details of construction of such a basis see, for example, [24]. With a slight abuse of notation, in what follows we write $\{\phi_{s_0 l}, l \in \mathcal{L}_{s_0}; \psi_{s l}, s \geq s_0, l \in \mathcal{L}_s\}$ for an orthonormal basis of $L_2([0, 1])$ rather than $L_2(\mathbb{R})$ and $\mathcal{L}_{s_0}$ and $\mathcal{L}_s$ are some proper index sets depending on $s_0$ and $s$ respectively. Again, we use tensor product to form a multidimensional wavelet basis, and then it fits into the context of our analysis.

The compactness of the father wavelet is needed to verify Assumption 3. To see this, consider a standard multiresolutional analysis setting. A large space, say $L_2([0, 1])$, is decomposed into a sequence of nested subspaces $\{0\} \subset \cdots \subset V_{-1} \subset V_0 \subset V_1 \cdots \subset L_2([0, 1])$. Generally, $\{\phi_{s l}, l \in \mathcal{L}_s\}$ constitutes a basis for $V_s$, and $\{\psi_{s l}, l \in \mathcal{L}_s\}$ forms a basis for the orthogonal complement $W_s$ of $V_s$. Since the support of $\phi$ is compact and $\{\phi_{s l}\}$ is generated simply by dilation and translation of $\phi$, $[0, 1]$ can be viewed as implicitly partitioned. Specifically, by increasing the resolution level from $s$ to $s + 1$, the length of the support of $\phi_{s l}$ is halved. Hence, it is equivalent to placing additional partitioning knots at the midpoint of each subinterval. In addition, as the sample size $n$ grows, we allow the resolution level $s$ to increase (thus written as $s_0$ when needed), but each basis function $\phi_{s_0 l}$ is supported by only a finite number of subintervals since the generating scaling function is compactly supported.
The above discussion connects the generating process of wavelets at different levels with a partitioning scheme. It is easy to see that Assumption 2 is automatically satisfied in this case, and given a resolution level \( s_n \), the mesh width \( b = 2^{-s_n} \). On the other hand, given a starting level \( s_0 \), \( \{ \phi_{s_l}, l \in \mathcal{L}_{s_l}; \psi_{s_l}, s_0 \leq s \leq s_n, l \in \mathcal{L}_s \} \) does not satisfy Assumption 3, since some functions in such a basis have increasing supports as the resolution level \( s_n \to \infty \). Nevertheless, they span the same space as \( \{ \phi_{s_n,l}, l \in \mathcal{L}_{s_n} \} \) and hence there exists a linear transformation which connects the two equivalent bases, and least squares estimators are invariant to such a transformation. Therefore, we only need to work with a tensor-product (father) wavelet basis

\[
p(x) := \otimes_{\ell=1}^{d} 2^{-s_n/2} \phi_{s_n}(x_{\ell})
\]

where \( \phi_{s_n} \) is a vector containing all functions in \( \{ \phi_{s_n,l}, l \in \mathcal{L}_{s_n} \} \). By multiplying the basis by \( 2^{-s_n/2} \), we drop the normalizing constants that originally appear in the construction of orthonormal basis. As the next lemma shows, this large class of orthogonal wavelet bases satisfy our assumptions.

**Lemma A.2 (Wavelets Estimators).** Let \( \phi \) and \( \psi \) be a scaling function and a wavelet function of degree \( m - 1 \) with \( q + 1 \) continuous derivatives, \( p(x) \) be the tensor product orthogonal (father) wavelet basis of degree \( m - 1 \) generated by \( \phi \), and suppose Assumption 1 holds with \( m \leq S \).

(a) \( p(x) \) satisfies Assumption 3.

(b) Assumption 4 holds with \( \varsigma = q \) and

\[
\rho_{m, \varsigma}(x) = - \sum_{\mathbf{u} \in \Lambda_m} \frac{\partial^{\mathbf{u}} \mu(x) \mathcal{P}^{m-|\mathbf{u}|} \mathcal{P}^{m-|\mathbf{u}|}}{\mathcal{P}^{m-|\mathbf{u}|} \mathcal{P}^{m-|\mathbf{u}|}} B_{\mathbf{u}, \varsigma}(x/b)
\]

where \( \Lambda_m = \{ \mathbf{u} \in \mathbb{Z}_+^d : [\mathbf{u}] = m \), and \( u_\ell = m \) for some \( 1 \leq \ell \leq d \} \) and \( B_{\mathbf{u}, \varsigma}(x) = \sum_{s \geq 0} \partial^{s} \xi_{\mathbf{u}, s}(x) \) converges uniformly, with \( \xi_{\mathbf{u}, s}(\cdot) \) being a linear combination of products of univariate father wavelet \( \phi \) and the mother wavelet \( \psi \); its exact form is notationally cumbersome and is given in the SA (Equation SA-9.4). Furthermore, Equation (2.4) holds.

(c) Let \( \tilde{\phi} \) be a scaling function of degree \( \tilde{m} - 1 \) with \( m + 1 \) continuous derivatives for some \( \tilde{m} > m \), \( \tilde{p}(x) \) be the tensor-product orthogonal wavelet basis generated by \( \tilde{\phi} \) having the same resolution level as \( p(x) \), and assume \( \tilde{m} \leq S \). Then Assumption 5 is satisfied.

In addition to verifying that our high-level assumptions hold for wavelets, this result gives a novel asymptotic error expansion for multidimensional compact-supported wavelets. Our derivation employs ideas in [56] and exploits the tensor product structure of the wavelet basis. The end result is similar to tensor product splines, in the sense that the total order of the approximating basis is not fixed at \( m \) and thus \( \Lambda_m \) is the same as that for \( B \)-splines.

### A.3 Generalized Regressograms on Rectangular Partitions

To construct the generalized regressogram, we first define the piecewise polynomials. What will distinguish these from splines is that (i) each polynomial is supported on exactly one cell, and relatedly (ii) no continuity is assumed between cells. First, for some fixed integer \( m \in \mathbb{Z}_+ \), let \( r(x_\ell) = (1, x_\ell, \ldots, x_\ell^{m-1})' \) denote a vector of powers up to degree \( m - 1 \). To extend it to a multidimensional basis, we take the tensor product of \( r(x_\ell) \), denoted by a column vector \( R(x) \). The total order of such a basis is not fixed, and its behavior is more similar to tensor-product \( B \)-splines. Following [12, and references therein], we exclude all terms with degree greater than \( m - 1 \) in
\( \mathbf{R}(\mathbf{x}) \). Hence the remaining elements in \( \mathbf{R}(\mathbf{x}) \) are given by \( \mathbf{x}^{\alpha} = x_1^{\alpha_1} \cdots x_d^{\alpha_d} \) for a unique \( d \)-tuple \( \alpha \) such that \( |\alpha| \leq m - 1 \). Then we "localize" this basis by restricting it to a particular subrectangle \( \delta_{l_1} \cdots l_d \). Specifically, we write \( p_{l_1} \cdots l_d(\mathbf{x}) = \mathbf{1}_{\delta_{l_1} \cdots l_d}(\mathbf{x}) \mathbf{R}(\mathbf{x}) \), where \( \mathbf{1}_{\delta_{l_1} \cdots l_d}(\mathbf{x}) \) is equal to 1 if \( \mathbf{x} \in \delta_{l_1} \cdots l_d \) and 0 otherwise. Finally, we rotate the basis by centering each basis function at the start point of the corresponding cell and scale it by interval lengths. We can arrange the basis functions according to a particular ordering \( \varphi \) first order the subrectangles the same way as that for \( B \)-splines, and then within each subrectangle \( \delta_{l_1} \cdots l_d \) the basis functions in \( \mathbf{R}(\mathbf{x}) \) are ordered ascendingly in \( \alpha \) and \( \ell = 1, \ldots, d \). Using the same notation as in Section 1, we will write the basis as \( \mathbf{p}(\mathbf{x}) = (p_{1}(\mathbf{x}), \cdots, p_{K}(\mathbf{x}))' \).

The following lemma shows that Assumptions 3 and 4 hold for generalized regressograms.

**Lemma A.3 (Generalized Regressograms).** Let \( \mathbf{p}(\mathbf{x}) \) be the rotated piecewise polynomial basis of degree \( m - 1 \) based on Legendre polynomials, and suppose that Assumptions 1 and 2 hold with \( m \leq S \). Then,

(a) \( \mathbf{p}(\mathbf{x}) \) satisfies Assumption 3.

(b) Assumption 4 holds with \( \varsigma = m - 1 \) and

\[
\mathcal{B}_{m,\varsigma}(\mathbf{x}) = - \sum_{\mathbf{u} \in \Lambda_m} \frac{\partial^n \mu(\mathbf{x}) h_{\mathbf{x}}^{m-|\mathbf{u}|}}{(u-\varsigma)!} \frac{\mathbf{b}_{\mathbf{x}}^{u-\varsigma}}{h_{\mathbf{x}}^{m-|\mathbf{u}|}} B_{\mathbf{u}}^p ((\mathbf{x} - t_{\mathbf{x}}^\alpha) \odot \mathbf{b}_{\mathbf{x}}),
\]

where \( \Lambda_m = \{ \mathbf{u} : |\mathbf{u}| = m \} \) and

\[
B_{\mathbf{u}}^p(\mathbf{x}) := \prod_{\ell=1}^d \left( \frac{2u_\ell}{u_\ell} \right)_{-1}^{-1} P_{u_\ell}(x_\ell),
\]

with \( P_{u_\ell}(\cdot) \) being the \( u_\ell \)-th shifted Legendre polynomial orthogonal on \([0,1] \), and \( B_{\mathbf{u}}^p = 0 \) if \( \mathbf{u} \) contains negative elements. Furthermore, Equation (2.4) holds.

(c) Let \( \mathbf{p}(\mathbf{x}) \) be a piecewise polynomial basis of degree \( \tilde{m} - 1 \) on the same partition \( \Delta \) for some \( \tilde{m} > m \), and assume \( \tilde{m} \leq S \). Then Assumption 5 is satisfied.

The leading asymptotic error obtained in Lemma A.3 differs from the one in [12] because it is expressed in terms of orthogonal polynomials. Here we employ Legendre polynomials \( \tilde{P}_m(x) \) orthogonal on \([-1,1]\) with respect to Lebesgue measure, and then shift them to \( P_m(x) = \tilde{P}_m(2x - 1) \), thus giving the shifted Legendre polynomials orthogonality on \([0,1]\). Further, we allow for more general partitioning schemes.

**A.4 Generalized Regressograms on General Partitions**

Moving away from rectangular partitions will impede verification of Assumption 4 in general. A more typical result, such as Equation (A.2) can be more easily proven in generality. In practice, however, many data-driven partitioning selection procedures, such as those induced by regression trees, do not necessarily lead to a rectangular partition. There is also a large literature in approximation theory discussing bases constructed on triangulations or other general partition schemes [e.g., 48]. To demonstrate the power of our theory, we show presently that for the generalized regressogram we can make obtain concrete results on non-rectangular partitions.

The basis is as described above, but now allowing for non-rectangular cells. Let \( t_\delta \) be an arbitrary point in each \( \delta \in \Delta \). Hence, \( t_{\delta_x} \) is simply a point in the cell containing \( x \). Construct the rotated
piecewise polynomial basis as in Section A.3, but centering the basis at \( t_\delta \) and rescaling it by the diameter of \( \delta \).

**Lemma A.4.** Let \( p(x) \) be the rotated piecewise polynomial basis of degree \( m-1 \) on a general partition \( \Delta \), and suppose that Assumptions 1 and 2 hold with \( m \leq S \). Then,

(a) \( p(x) \) satisfies Assumption 3.

(b) Assumption 4 holds with \( \varsigma = m - 1 \) and

\[
\mathcal{B}_{m, \varsigma}(x) = - \sum_{u \in \Lambda_m} \frac{\partial^\varsigma \mu(x) h_{x}^{m-\varsigma}}{(\mathbf{u} - \varsigma)!} \left( \frac{x - t_\delta}{h_x} \right)^{\varsigma} \mathbf{1}(\mathbf{u} \geq \varsigma),
\]

where \( \Lambda_m = \{ \mathbf{u} : [\mathbf{u}] = m \} \) and \( \mathbf{u} \geq \varsigma \iff u_1 \geq \varsigma_1, \ldots, u_d \geq \varsigma_d \).

(c) Let \( \tilde{p}(x) \) be a piecewise polynomial of degree \( \tilde{m} - 1 \) on the same partition \( \Delta \) for some \( \tilde{m} > m \), and assume \( \tilde{m} \leq S \). Then Assumption 5 is satisfied.

It is a challenging task to construct orthogonal polynomial basis on non-rectangular domains, which makes Equation (2.4) hard to satisfy when employing partitioning estimators on non-rectangular partitions. Thus, our more general characterization (and correction) of the bias is quite useful in this case.

**Appendix B  Further Technical Results for Rectangular Partitions**

We now state an IMSE result that specializes Theorem 4.1 to the case of rectangular partitions but is less restricted than Corollary 1. In particular, we allow for any \( q \) and do not rely on Assumption 6(c). This result gives the key ingredient to characterize the limit \( \mathcal{V}_{\Delta, q} \) and \( \mathcal{B}_{\Delta, q} \) in some generality.

**Theorem B.1** (IMSE for Rectangular Partitions). Suppose that the conditions in Theorem 4.1 and Assumption 6(a) and 6(b) hold. Then, for any arbitrary sequence of points \( \{ \tau_k \}_{k=1}^K \) such that \( \tau_k \in \text{supp}(p_k(\cdot)) \) for each \( k = 1, \ldots, K \),

\[
\mathcal{V}_{\Delta, q} = \mathcal{V}_{\kappa, q} + o(h^{-d-2[q]}), \quad \text{and} \quad \mathcal{B}_{\Delta, q} = \mathcal{B}_{\kappa, q} + o(h^{2m-2[q]}),
\]

with

\[
\mathcal{V}_{\kappa, q} = \kappa^{1+2q} \sum_{k=1}^K \left[ \frac{\sigma^2(\tau_k) w(\tau_k)}{f(\tau_k)} \prod_{l=1}^d g_e(\tau_k) \right] e_k^T H_1^{-1} H_q e_k \text{vol}(\delta_{\tau_k}) \asymp h^{-d-2[q]},
\]

\[
\mathcal{B}_{\kappa, q} = \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1+u_2-2q)} \left\{ \mathcal{B}_{u_1, u_2, q} + v'_{u_1, 0} H_0^{-1} v_{u_2, 0} - 2v'_{u_1, q} H_0^{-1} v_{u_2, 0} \right\} \lesssim h^{2m-2[q]},
\]

where the \( K \)-dimensional vector \( e_k \) is the \( k \)-th unit vector (i.e., \( e_k \) has a 1 in the \( k \)-th position and 0 elsewhere), \( q(x) = (g_1(x), \ldots, g_d(x))' \),

\[
H_q = \kappa^{-2q} \int_{[0,1]^d} \partial^q p(x) \partial^q p(x)' dx, \quad \mathcal{B}_{u_1, u_2, q} = \eta_{u_1, u_2, q} \int_{[0,1]^d} \frac{\partial^{u_1} \mu(x) \partial^{u_2} \mu(x)}{\mathbf{g}(x)^{u_1+u_2-2q}} w(x) dx,
\]

and the \( K \)-dimensional vector \( v_{u_1, q} \) has \( k \)-th typical element

\[
\sqrt{w(\tau_k)} \frac{\partial^u \mu(\tau_k)}{\kappa^{q} \mathbf{g}(\tau_k)^{u-q}} \int_{\mathcal{X}} \frac{h_x^{m-[q]}}{b_x^{u-q}} \partial^q p_k(x) B_{u, q}(x) dx.
\]
This theorem takes advantage of the assumed rectangular structure of the partition $\Delta$ to express the leading bias and variance as proportional to the number of subintervals used for each regressor. The accompanying constants are expressed as sums over the local contributions of the basis function $p(\cdot)$ used, and are easy to be shown bounded in general (they may still not converge to any well-defined limit at this level of generality).

Assumption 6(a) used in Theorem B.1 (and in Corollary 1) slightly strengthens the quasi-uniform condition imposed in Assumption 2. It can be verified by choosing knot positions/configuration appropriately. Specifically, let $\Delta$ be a rectangular partition (see (2.1)) with (marginal) knots chosen as

$$t_{\ell,l} = G_{\ell}^{-1}\left(\frac{l}{\kappa_{\ell}}\right), \quad l = 0, 1, \ldots, \kappa_{\ell}, \quad \ell = 1, 2, \ldots, d,$$

where $G_{\ell}(\cdot)$ is a univariate continuously differentiable distribution function and $G_{\ell}^{-1}(v) = \inf\{x \in \mathbb{R} : G_{\ell}(x) \geq v\}$. In this case the function $g_{\ell}(\cdot)$ in Assumption 6(a) is simply the density of $G_{\ell}(\cdot)$.

Two examples commonly used in practice are: (i) evenly-spaced partitions, denoted by $\Delta_{ES}$, where $G_{\ell}(x) = x$ and $g_{\ell}(x) = 1$, and (ii) quantile-spaced partitions, denoted by $\Delta_{QS}$, where $G_{\ell}(x) = \hat{F}_{\ell}(x)$ with $\hat{F}_{\ell}(x)$ the empirical distribution function for the $\ell$-th covariate, $\ell = 1, 2, \ldots, d$. For the case of quantile-spaced partitioning, if $\hat{F}_{\ell}(x)$ convergences to $F_{\ell}(x) = \mathbb{P}[x_{\ell} \leq x]$ in a suitable sense, $g_{\ell}(x) = dF_{\ell}(x)/dx$, i.e., the marginal density of $x_{\ell}$. See [1] and [2] for a slightly more high-level condition in the context of univariate $B$-splines.

Assumption 6(b) in Theorem B.1 (and in Corollary 1) involves a scaling factor $h_{x_{\ell}^{-2q}}/b_{x_{\ell}^{-2q}}$. In Section 2 we do not add specific restrictions on the shape of cells in $\Delta$ and thus the diameter of a cell is used to conveniently express the order of approximation error (denoted by $h$), but for rectangular partitions the approximation error is usually characterized by the lengths of intervals on each axis (denoted by $b_x$), as illustrated in Appendix A. Therefore, the scaling factor makes the above theorem immediately apply to the three leading examples in Appendix A based on rectangular partitioning schemes (splines, wavelets, piecewise polynomials).

Using Theorem B.1, the IMSE-optimal tuning parameter selector for partitioning-based series estimators becomes

$$\kappa_{\text{IMSE}, q} = \arg \min_{\kappa \in \mathbb{Z}_+^d} \left\{\frac{1}{n} \mathcal{V}_{\kappa, q} + \mathcal{B}_{\kappa, q}\right\},$$

which is still given in implicit form because of the leading constants are not known to converge at this level of generality. Nevertheless, it follows that $\kappa_{\text{IMSE}, q} = (\kappa_{\text{IMSE}, q, 1}, \ldots, \kappa_{\text{IMSE}, q, d})'$ with $\kappa_{\text{IMSE}, q, \ell} \propto n^{\frac{1}{2m+d} - 1}$, $\ell = 1, 2, \ldots, d$. 

35
Table 1: Pointwise Estimation and Inference, B-Splines, Evenly-Spaced Partition.

| j  | \(\kappa\) | \(x = 0.2\) | \(x = 0.5\) | \(x = 0.8\) |
|----|-------------|-------------|-------------|-------------|
|    | RMSE | CR | IL | RMSE | CR | IL | RMSE | CR | IL |
| 0  | \(\kappa\text{IMSE}\) | 3.0 | 0.007 | 94.8 | 0.252 | 0.046 | 91.5 | 0.328 | 0.004 | 94.7 | 0.221 | 0.002 | 94.4 | 0.400 | 0.009 | 94.9 | 0.199 |
|   | \(\hat{\kappa}\text{ROT}\) | 4.9 | 0.008 | 95.1 | 0.328 | 0.009 | 94.6 | 0.317 | 0.009 | 94.2 | 0.328 |
|   | \(\hat{\kappa}\text{DPI}\) | 5.1 | 0.006 | 95.2 | 0.323 | 0.007 | 94.4 | 0.318 | 0.008 | 94.5 | 0.323 |
| 1  | \(\kappa\text{IMSE}\) | 3.0 | 0.002 | 95.3 | 0.304 | 0.003 | 94.8 | 0.226 | 0.001 | 94.8 | 0.304 |
|   | \(\hat{\kappa}\text{ROT}\) | 4.9 | 0.001 | 95.3 | 0.304 | 0.006 | 95.0 | 0.298 | 0.008 | 95.0 | 0.304 |
|   | \(\hat{\kappa}\text{DPI}\) | 5.1 | 0.001 | 95.0 | 0.315 | 0.006 | 95.1 | 0.306 | 0.007 | 94.7 | 0.315 |
| 2  | \(\kappa\text{IMSE}\) | 4.0 | 0.004 | 94.5 | 0.314 | 0.009 | 94.8 | 0.311 | 0.001 | 94.9 | 0.314 |
|   | \(\hat{\kappa}\text{ROT}\) | 4.9 | 0.001 | 94.7 | 0.323 | 0.003 | 95.0 | 0.336 | 0.002 | 95.2 | 0.323 |
|   | \(\hat{\kappa}\text{DPI}\) | 5.1 | 0.001 | 94.5 | 0.332 | 0.003 | 94.9 | 0.342 | 0.001 | 95.1 | 0.332 |
| 3  | \(\kappa\text{IMSE}\) | 4.0 | 0.001 | 94.6 | 0.403 | 0.008 | 94.5 | 0.250 | 0.034 | 92.9 | 0.341 |
|   | \(\hat{\kappa}\text{ROT}\) | 4.9 | 0.002 | 95.0 | 0.318 | 0.021 | 93.8 | 0.391 | 0.003 | 94.8 | 0.318 |
|   | \(\hat{\kappa}\text{DPI}\) | 5.1 | 0.000 | 95.3 | 0.339 | 0.005 | 94.3 | 0.331 | 0.005 | 94.8 | 0.321 |

Notes: (i) \(n = 1,000;\) 5,000 replications; \(m = 2,\) \(\hat{m} = 3,\) \(\Delta = \hat{\Delta};\) (iii) RMSE = root mean squared error of point estimator, CR = coverage rate of 95% nominal confidence intervals, IL = average interval length of 95% nominal confidence intervals; (iv) \(\kappa_{\text{IMSE}}\) = infeasible IMSE-optimal number of partitions, \(\hat{\kappa}_{\text{ROT}}\) = feasible rule-of-thumb (ROT) implementation of \(\kappa_{\text{IMSE}},\) \(\hat{\kappa}_{\text{DPI}}\) = feasible direct plug-in (DPI) implementation of \(\kappa_{\text{IMSE}}.\) See §8A-6 and §8A-7 in supplemental appendix for more details.
|       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|
|       |       |       |       |       |       |       |
|       | $x = 0.2$ |       | $x = 0.5$ |       | $x = 0.8$ |       |
| $\kappa$ | RMSE | CR | IL | RMSE | CR | IL | RMSE | CR | IL |
| $j = 0$ |       |       |       |       |       |       |       |       |       |
| 1.0   | 0.084 | 56.5 | 0.185 | 0.046 | 87.6 | 0.240 | 0.068 | 69.6 | 0.185 |
| 2.0   | 0.054 | 79.0 | 0.190 | 0.045 | 81.9 | 0.176 | 0.003 | 94.2 | 0.190 |
| $\kappa_{\text{IMSE}}$ | 3.0 | 0.006 | 94.8 | 0.252 | 0.036 | 92.1 | 0.308 | 0.036 | 90.8 | 0.253 |
|       | 4.0   | 0.021 | 94.3 | 0.352 | 0.003 | 94.5 | 0.225 | 0.018 | 94.0 | 0.352 |
|       | 5.0   | 0.003 | 94.6 | 0.322 | 0.013 | 94.5 | 0.364 | 0.001 | 94.9 | 0.321 |
| $\kappa_{\text{ROT}}$ | 4.9 | 0.006 | 94.8 | 0.323 | 0.006 | 94.7 | 0.302 | 0.007 | 94.3 | 0.322 |
| $\kappa_{\text{DPI}}$ | 5.1 | 0.005 | 95.0 | 0.318 | 0.005 | 94.1 | 0.305 | 0.006 | 94.6 | 0.317 |
| $j = 1$ |       |       |       |       |       |       |       |       |       |
| 1.0   | 0.032 | 90.7 | 0.205 | 0.030 | 89.5 | 0.187 | 0.056 | 80.7 | 0.205 |
| 2.0   | 0.025 | 93.3 | 0.269 | 0.012 | 94.2 | 0.251 | 0.033 | 92.3 | 0.269 |
| $\kappa_{\text{IMSE}}$ | 3.0 | 0.002 | 95.4 | 0.303 | 0.004 | 94.6 | 0.228 | 0.001 | 94.8 | 0.303 |
|       | 4.0   | 0.003 | 95.1 | 0.281 | 0.008 | 94.6 | 0.307 | 0.010 | 94.4 | 0.280 |
|       | 5.0   | 0.001 | 94.7 | 0.298 | 0.008 | 94.7 | 0.274 | 0.008 | 94.7 | 0.298 |
| $\kappa_{\text{ROT}}$ | 4.9 | 0.002 | 95.5 | 0.305 | 0.006 | 94.9 | 0.300 | 0.008 | 95.0 | 0.306 |
| $\kappa_{\text{DPI}}$ | 5.1 | 0.001 | 94.9 | 0.316 | 0.006 | 95.0 | 0.307 | 0.006 | 94.7 | 0.316 |
| $j = 2$ |       |       |       |       |       |       |       |       |       |
| 1.0   | 0.029 | 91.2 | 0.205 | 0.019 | 93.1 | 0.194 | 0.054 | 82.0 | 0.206 |
| 2.0   | 0.031 | 92.4 | 0.270 | 0.013 | 94.2 | 0.251 | 0.038 | 91.2 | 0.270 |
| $\kappa_{\text{IMSE}}$ | 3.0 | 0.002 | 95.0 | 0.314 | 0.004 | 94.5 | 0.264 | 0.001 | 94.8 | 0.313 |
|       | 4.0   | 0.003 | 94.6 | 0.310 | 0.009 | 94.6 | 0.310 | 0.000 | 95.0 | 0.310 |
|       | 5.0   | 0.000 | 94.3 | 0.314 | 0.002 | 94.5 | 0.329 | 0.005 | 94.8 | 0.314 |
| $\kappa_{\text{ROT}}$ | 4.9 | 0.001 | 94.7 | 0.323 | 0.004 | 94.9 | 0.330 | 0.002 | 95.2 | 0.323 |
| $\kappa_{\text{DPI}}$ | 5.1 | 0.001 | 94.5 | 0.332 | 0.004 | 94.8 | 0.337 | 0.001 | 94.9 | 0.332 |
| $j = 3$ |       |       |       |       |       |       |       |       |       |
| 1.0   | 0.046 | 85.0 | 0.198 | 0.022 | 35.3 | 0.254 | 0.038 | 88.3 | 0.198 |
| 2.0   | 0.005 | 95.0 | 0.223 | 0.012 | 93.9 | 0.198 | 0.043 | 87.6 | 0.223 |
| $\kappa_{\text{IMSE}}$ | 3.0 | 0.005 | 94.9 | 0.249 | 0.012 | 90.0 | 0.299 | 0.037 | 90.7 | 0.249 |
|       | 4.0   | 0.000 | 94.8 | 0.341 | 0.008 | 94.7 | 0.251 | 0.005 | 93.6 | 0.341 |
|       | 5.0   | 0.002 | 94.9 | 0.325 | 0.003 | 93.8 | 0.355 | 0.004 | 94.8 | 0.324 |
| $\kappa_{\text{ROT}}$ | 4.9 | 0.002 | 94.9 | 0.326 | 0.006 | 94.8 | 0.311 | 0.001 | 94.7 | 0.324 |
| $\kappa_{\text{DPI}}$ | 5.1 | 0.001 | 94.7 | 0.324 | 0.005 | 94.4 | 0.316 | 0.000 | 94.8 | 0.324 |

Notes: (i) $n = 1,000$; 5,000 replications; $m = 2$, $\tilde{m} = 3$, $\Delta = \tilde{\Delta}$; (ii) RMSE = root mean squared error of point estimator; CR = coverage rate of 95% nominal confidence intervals; IL = average interval length of 95% nominal confidence intervals; (iii) $\kappa_{\text{IMSE}}$ = infeasible IMSE-optimal number of partitions, $\kappa_{\text{ROT}}$ = feasible rule-of-thumb (ROT) implementation of $\kappa_{\text{IMSE}}$, $\kappa_{\text{DPI}}$ = feasible direct plug-in (DPI) implementation of $\kappa_{\text{IMSE}}$. See §SA-6 and §SA-7 in supplemental appendix for more details.
Table 3: Uniform Inference: B-Splines

|                  | Evenly-spaced |                     | Quantile-spaced |                     |
|------------------|---------------|---------------------|-----------------|---------------------|
|                  | CP  | ACE  | AW   | UCR  | CP  | ACE  | AW   | UCR  |
| Plug-in, $j = 0$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 0.92 | 0.017 | 0.384 | 79.68 | 0.92 | 0.017 | 0.383 | 79.96 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.005 | 0.469 | 92.22 | 1.00 | 0.005 | 0.468 | 92.22 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.006 | 0.478 | 91.40 | 1.00 | 0.006 | 0.478 | 91.56 |
| Plug-in, $j = 1$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00 | 0.005 | 0.426 | 93.86 | 1.00 | 0.005 | 0.426 | 93.52 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.506 | 93.72 | 1.00 | 0.004 | 0.505 | 93.44 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.003 | 0.514 | 93.44 | 1.00 | 0.003 | 0.515 | 93.44 |
| Plug-in, $j = 2$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00 | 0.005 | 0.443 | 94.06 | 1.00 | 0.005 | 0.443 | 94.02 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.003 | 0.536 | 93.78 | 1.00 | 0.003 | 0.536 | 93.72 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.003 | 0.546 | 93.34 | 1.00 | 0.003 | 0.547 | 93.50 |
| Plug-in, $j = 3$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00 | 0.008 | 0.413 | 88.98 | 1.00 | 0.008 | 0.413 | 88.76 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.499 | 93.56 | 1.00 | 0.004 | 0.498 | 93.34 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.509 | 93.04 | 1.00 | 0.004 | 0.509 | 93.08 |
| Bootstrap, $j = 0$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 0.92 | 0.017 | 0.382 | 79.50 | 0.92 | 0.017 | 0.382 | 79.62 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.006 | 0.466 | 91.82 | 1.00 | 0.006 | 0.464 | 91.74 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.006 | 0.475 | 91.16 | 1.00 | 0.006 | 0.474 | 90.92 |
| Bootstrap, $j = 1$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00 | 0.005 | 0.424 | 93.64 | 1.00 | 0.005 | 0.423 | 93.54 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.501 | 93.06 | 1.00 | 0.004 | 0.501 | 92.76 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.510 | 92.76 | 1.00 | 0.004 | 0.510 | 92.60 |
| Bootstrap, $j = 2$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00 | 0.005 | 0.440 | 94.04 | 1.00 | 0.005 | 0.440 | 93.82 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.531 | 93.20 | 1.00 | 0.004 | 0.531 | 93.12 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.541 | 92.74 | 1.00 | 0.003 | 0.542 | 93.04 |
| Bootstrap, $j = 3$ |     |      |      |      |     |      |      |      |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00 | 0.008 | 0.411 | 88.36 | 1.00 | 0.008 | 0.411 | 88.48 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.495 | 93.22 | 1.00 | 0.004 | 0.494 | 92.76 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.504 | 92.58 | 1.00 | 0.004 | 0.504 | 92.56 |

Notes: (i) $n = 1,000$; 5,000 replications; $m = 2$, $\tilde{m} = 3$, $\Delta = \tilde{\Delta}$; (iii) CP = proportion of values covered with probability at least 95%, ACE = average coverage errors of 95% nominal confidence intervals, AW = average width of 95% nominal confidence band, UCR = uniform coverage rate of 95% nominal confidence band; (iv) $\hat{\kappa}_{\text{IMSE}}$ = infeasible IMSE-optimal number of partitions, $\hat{\kappa}_{\text{ROT}}$ = feasible rule-of-thumb (ROT) implementation of $\kappa_{\text{IMSE}}$, $\hat{\kappa}_{\text{DPI}}$ = feasible direct plug-in (DPI) implementation of $\kappa_{\text{IMSE}}$. See §SA-6 and §SA-7 in supplemental appendix for more details.
Supplemental Appendix (For Online Publication Only)

This supplement gives all theoretical proofs of the results discussed in the main paper, and additional simulation evidence. Furthermore, other technical results and discussion not in the main text are presented, which may be of independent interest. In particular, Section SA-9 studies pointwise estimation and inference, which are directly applicable to nonparametric inference. Section SA-10.1 proves uniform convergence rates and uniform stochastic linearizations, which are useful in semiparametric settings. The contributions of these two sections are not detailed in the main text. A general version of our coupling approach (used to establish Equation (3.3) and Theorem 3.1 in the main paper) is stated in Lemma SA-10.5 (page 53).

Beyond this, Section SA-11 gives additional technical discussions of the spline and piecewise polynomial bases, and of the bias correction strategies discussed in the main paper. Implementation and other numerical issues are discussed in Section SA-12, and results from a simulation study are presented in Section SA-13; see also the companion R package lpartition detailed in [13] and available at 

http://sites.google.com/site/nppackages/lspartition.

Finally, Section SA-14 and SA-15 contain all proofs. This supplemental appendix is largely self-contained. A detailed table of contents appears below.

SA-7 Setup, Assumptions and Notation

SA-7.1 Notation

For a $d$-tuple $q = (q_1, \ldots, q_d) \in \mathbb{Z}_+^d$, define $[q] = \sum_{\ell=1}^d q_\ell$, $x^q = x_1^{q_1} x_2^{q_2} \cdots x_d^{q_d}$ 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 $0 = (0, \ldots, 0)'$ be the length-$d$ zero vector. The tensor product or Kronecker product operator is $\otimes$, and the entrywise division operator (Hadamard division) is $\oslash$. The smallest integer greater than or equal to $u$ is $\lceil u \rceil$. 

39
We use several norms. For a column vector \( \mathbf{v} = (v_1, \ldots, v_M)' \in \mathbb{R}^M \), we write \( \|\mathbf{v}\| = (\sum_{i=1}^M v_i^2)^{1/2} \), \( \|\mathbf{v}\|_\infty = \max_{1 \leq i \leq M} |v_i| \) and \( \dim(\mathbf{v}) = M \). For a matrix \( \mathbf{A} \in \mathbb{R}^{M \times N} \), \( \|\mathbf{A}\|_1 = \max_{1 \leq j \leq N} \sum_{i=1}^M |a_{ij}| \), \( \|\mathbf{A}\| = \max_i \sigma_i(\mathbf{A}) \) and \( \|\mathbf{A}\|_\infty = \max_{1 \leq i \leq M} \sum_{j=1}^N |a_{ij}| \) for operator norms induced by \( L_1 \), \( L_2 \) and \( L_\infty \) norms respectively, where \( \sigma_i(\mathbf{A}) \) is the \( i \)-th singular value of \( \mathbf{A} \). \( \lambda_{\max}(\mathbf{A}) \) and \( \lambda_{\min}(\mathbf{A}) \) denote the maximum and minimum eigenvalues of \( \mathbf{A} \). For a real-valued function \( g(x) \), \( \|g\|_{L_p(\mathcal{X})} = (\int_\mathcal{X} |g(x)|^p \, dx)^{1/p} \), and \( \|g\|_{L_\infty(\mathcal{X})} = \text{ess sup}_{x \in \mathcal{X}} |g(x)| \).

We also use standard empirical process notation: \( \mathbb{E}_n[g] = \mathbb{E}_n[g(x_i)] = \frac{1}{n} \sum_{i=1}^n g(x_i) \), \( \mathbb{E}[g] = \mathbb{E}[g(x_i)] \), and \( \mathbb{G}_n[g] = \mathbb{G}_n[g(x_i)] = \frac{1}{\sqrt{n}} \sum_{i=1}^n (g(x_i) - \mathbb{E}[g(x_i)]) \). For sequences of numbers or random variables, we use \( a_n \lesssim b_n \) to denote \( \limsup_n [a_n/b_n] \) is finite, and \( a_n \lesssim_p b_n \) or \( a_n = O_p(b_n) \) to denote \( \limsup_{\epsilon \to 0} \limsup_n [P[a_n/b_n] \geq \epsilon] = 0 \). \( a_n = o(b_n) \) implies \( a_n/b_n \to 0 \), and \( a_n = o_P(b_n) \) implies that \( a_n/b_n \to_p 0 \) where \( \to_p \) denotes convergence in probability. \( a_n \asymp b_n \) implies that \( a_n \lesssim b_n \) and \( b_n \lesssim a_n \). \( \sim \) denotes convergence in distribution, and for two random variables \( X \) and \( Y \), \( X \sim_d Y \) implies that they have the same probability distribution.

We set \( \mu(x) := \partial^0 \mu(x) \) and likewise \( \mu_{ij}(x) := \partial^0 \mu_{ij}(x) \) for \( j = 0, 1, 2, 3 \). Let \( \mathbf{X} = [x_1, \ldots, x_n]' \).

Finally, let \( C, C_1, C_2, \ldots \) denote universal constants which may be difference in different uses.

**SA-7.2 Setup**

As a complete reference, this section repeats the setup, assumptions and notation used in the main paper.

We study the standard nonparametric regression setup, where \( \{(y_i, x_i'), i = 1, \ldots, n\} \) is a random sample from the model

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

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. Let \( \Delta = \{\delta_l : 1 \leq l \leq \tilde{k}\} \) be a collection of open and disjoint subsets of \( \mathcal{X} \) such that the closure of their union is \( \mathcal{X} \). Every \( \delta_l \in \Delta \) is restricted to be polyhedral. An important special
case is a “rectangular” partition, in which case the support of the regressors is of tensor product form and each dimension of $\mathcal{X}$ is partitioned marginally into intervals.

Based on a particular partition, the dictionary of $K$ basis functions, each of order $m$ (e.g., $m = 4$ for cubic splines) is denoted by

$$x_i \mapsto p(x_i) := p(x_i; \Delta, m) = (p_1(x_i; \Delta, m), \ldots, p_K(x_i; \Delta, m))'. $$

For a point $x \in \mathcal{X}$ and $q = (q_1, \ldots, q_d)' \in \mathbb{Z}_+^d$, the partial derivative $\partial^q \mu(x)$ is estimated by least squares regression

$$\hat{\partial^q \mu}(x) = \partial^q p(x)' \hat{\beta}, \quad \hat{\beta} \in \arg \min_{b \in \mathbb{R}^K} \sum_{i=1}^n (y_i - p(x_i)' b)^2,$$

Our assumptions will guarantee that the sample matrix $\mathbb{E}_n[p(x_i)' p(x_i)']$ is nonsingular with probability approaching one in large samples, and thus we write the estimator as

$$\hat{\partial^q \mu_0}(x) := \tilde{\gamma}_{q,0}(x)' \mathbb{E}_n [\Pi_0(x_i)y_i],$$

where $\tilde{\gamma}_{q,0}(x)' := \partial^q p(x)' \mathbb{E}_n[p(x_i)' p(x_i)']^{-1}$ and $\Pi_0(x_i) := p(x_i)$.

The order $m$ of the basis is usually fixed in practice, and thus the tuning parameter for this class of nonparametric estimators is $\Delta$. As $n \to \infty$, $\kappa \to \infty$, and the volume of each $\delta_l$ shrinks proportionally to $h^d$, where $h = \max\{\text{diam}(\delta) : \delta \in \Delta\}$. Under our assumptions below, $K$ diverges proportionally as well.

**SA-7.3 Assumptions**

We list our main assumptions in what follows. The first assumption concerns the data generating process.

**Assumption SA-1** (Data Generating Process).

(a) $\{(y_i, x'_i) : 1 \leq i \leq n\}$ are i.i.d. satisfying (SA-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 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 $S > [q]$, and all $\partial^S \mu(\cdot)$, $[s] = S$, are Hölder continuous with exponent $\varrho > 0$.

The next assumption, often referred to as a “quasi-uniformity” condition in literature, restricts the shape of the resulting cells.

**Assumption SA-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 $\delta$.

Given such a partition, the next assumption requires that the basis is “locally” supported. We employ the notion of active basis: a function $p(\cdot)$ on $X$ is active on $\delta \in \Delta$ if it is not identically zero on $\delta$.

**Assumption SA-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 $H_k$. For any $k = 1, \ldots, K$, both the number of elements involved in $H_k$ and the number of basis functions which are active on $H_k$ are bounded by a constant.

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

$$
a' \int_{H_k} p(x; \Delta, m)p(x; \Delta, m)' dx a \gtrsim a^2 h^d, \quad k = 1, \ldots, K.
$$

(c) For an integer $\varsigma \in \mathbb{Z}$, 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 $[\varsigma] = \varsigma + 1$,

$$
\sup_{\delta \in \Delta} \sup_{x \in \text{clo}(\delta)} \|\partial^\varsigma p(x; \Delta, m)\| \lesssim h^{-\varsigma - 1}.
$$
We remind readers that Assumption SA-2 and SA-3 implicitly relate the number of approximating series terms, the number of cells in \( \Delta \) and the maximum mesh size: \( K \propto \kappa \propto h^{-d} \).

The next assumption gives an explicit high-level expression of the leading approximation error which is needed for bias correction and integrated mean squared error (IMSE) expansion. To simplify notation, for each \( x \in \mathcal{X} \) define \( \delta_x \) as the element of \( \Delta \) whose closure contains \( x \) and \( h_x \) for the diameter of this \( \delta_x \).

**Assumption SA-4 (Approximation Error).** For all \( \varsigma \) satisfying \( [\varsigma] \leq \varsigma \), given in Assumption SA-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

\[
\sup_{x \in \mathcal{X}} |\partial^\varsigma \mu(x) - \partial^\varsigma s^*(x) + \mathcal{B}_{m,\varsigma}(x)| \lesssim h^{m+\varsigma}[\varsigma] \tag{SA-3}
\]

and

\[
\sup_{\delta \in \Delta} \sup_{x_1, x_2 \in \text{clo}(\delta)} \frac{|B_{u,\varsigma}(x_1) - B_{u,\varsigma}(x_2)|}{||x_1 - x_2||} \lesssim h^{-1} \tag{SA-4}
\]

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

Our last assumption concerns the basis used for bias correction. Specifically, for some \( \bar{m} > m \), let \( \bar{p}(x) := \bar{p}(x; \bar{\Delta}, \bar{m}) \) be a basis of order \( \bar{m} \) defined on partition \( \bar{\Delta} \) which has maximum mesh \( \bar{h} \). Objects accented with a tilde always pertain to this secondary basis and partition for bias correction.

**Assumption SA-5 (Bias Correction).** The partition \( \bar{\Delta} \) satisfies Assumption SA-2, with maximum mesh \( \bar{h} \), and the basis \( \bar{p}(x; \bar{\Delta}, \bar{m}) \) satisfies Assumptions SA-3 and SA-4 with \( \varsigma = \varsigma(\bar{m}) \geq m \) in place of \( \varsigma \). Let \( \rho := \bar{h}/\bar{h} \), which obeys \( \rho \to \rho_0 \in (0, \infty) \). In addition, for \( j = 3 \), either (i) \( \bar{p}(x; \bar{\Delta}, \bar{m}) \) 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 \( \bar{p}(x; \bar{\Delta}, \bar{m}) \) reproduce polynomials of degree \([q]\).
SA-8 Technical Lemmas

In this section we present a series of technical lemmas that will be used for pointwise and uniform analysis. To begin with, we introduce additional notation which greatly simplify our expressions.

The following are some important outer-product matrices:

\[
\begin{align*}
Q_m &= E[p(x_i)p(x_i)'], \\
\hat{Q}_m &= E_n[p(x_i)p(x_i)'], \\
\tilde{Q}_m &= E[p(\tilde{x}_i)\tilde{p}(\tilde{x}_i)'], \\
\hat{\tilde{Q}}_m &= E_n[\tilde{p}(x_i)\tilde{p}(x_i)'], \\
Q_{m,\tilde{m}} &= E[p(x_i)\tilde{p}(x_i)'], \\
\hat{Q}_{m,\tilde{m}} &= E_n[p(x_i)\tilde{p}(x_i)'].
\end{align*}
\]

In the main paper, we define four nonparametric estimators based on the partitioning method, numbered as \(j = 0, 1, 2, 3\), which can be written in exactly the same form:

\[
\frac{\partial q}{\partial \mu_j}(x) := \frac{\partial q}{\partial \gamma_{q,j}}(x)\hat{Q}_m^{-1} \quad \text{and} \quad \Pi_j(x_i) := p(x_i) \quad (\text{SA-1})
\]

For \(j = 1\) (high-order-basis bias correction),

\[
\frac{\partial q}{\partial \gamma_{q,1}}(x) := \frac{\partial q}{\partial \gamma_{q,0}}(x)\hat{Q}_{\tilde{m}}^{-1}, \quad \text{and} \quad \Pi_1(x_i) := \tilde{p}(x_i) \quad (\text{SA-2})
\]

For \(j = 2\) (least squares bias correction),

\[
\frac{\partial q}{\partial \gamma_{q,2}}(x) := \left(\frac{\partial q}{\partial \gamma_{q,0}}(x)' - \frac{\partial q}{\partial \gamma_{q,0}}(x)\hat{Q}_{m,\tilde{m}}\hat{Q}_{\tilde{m}}^{-1} + \frac{\partial q}{\partial \gamma_{q,1}}(x)\right)' \quad \text{and} \quad \Pi_2(x_i) := (p(x_i)', \tilde{p}(x_i))' \quad (\text{SA-3})
\]
For $j = 3$ (plug-in bias correction),

$$\hat{\gamma}_{q,3}(x)' = \left(\hat{\gamma}_{q,0}(x)', \sum_{u \in \Lambda_m} \left\{ \hat{\gamma}_{u,1}(x)'h_x^{m-[q]}B_{u,q}(x) - \hat{\gamma}_{q,0}(x)'E_n[p(x_i)h_x^mB_{u,0}(x_i)\hat{\gamma}_{u,1}(x_i)'] \right\} \right),$$

and

$$\Pi_3(x_i) := (p(x_i)', \tilde{p}(x_i)')'.$$

(SA-4)

For $j = 0, 1, 2, 3$, $\gamma_{q,j}(x)$ is defined as $\hat{\gamma}_{q,j}(x)$ in (SA-1), (SA-2), (SA-3) and (SA-4) but with sample averages replaced by their population counterparts. Finally, define

$$\Sigma_j = E[\Pi_j(x_i)\Pi_j(x_i)'\sigma^2(x_i)], \quad \tilde{\Sigma}_j = E_n[\Pi_j(x_i)\Pi_j(x_i)'\sigma^2(x_i)]$$

(SA-5)

and

$$\Omega_j(x) = \gamma_{q,j}(x)'\Sigma_j\gamma_{q,j}(x), \quad \tilde{\Omega}_j(x) = \hat{\gamma}_{q,j}(x)'\tilde{\Sigma}_j\hat{\gamma}_{q,j}(x).$$

The next two lemmas establish the convergence rate and boundedness for $\hat{Q}_m$ and other relevant matrices which will be crucial for both pointwise and uniform analysis. Since the orders of bases used to construct bias corrected estimators are fixed and mesh size ratio $\rho \to \rho_0 \in (0, \infty)$ by Assumption SA-5, the same conclusions apply to $\hat{Q}_{\tilde{m}}$ and $Q_{\tilde{m}}$, though we do not make it explicit in the statement.

**Lemma SA-8.1.** Under Assumptions SA-1–SA-3,

$$h^d \lesssim \lambda_{\min}(Q_m) \leq \lambda_{\max}(Q_m) \lesssim h^d, \quad \|Q_m^{-1}\|_{\infty} \lesssim h^{-d}.$$

Moreover, when $\frac{\log n}{nh^d} = o(1)$, we have

$$\|\hat{Q}_m - Q_m\|_{\infty} \lesssim_P h^d \sqrt{\log n/(nh^d)} = o_P(h^d),$$

$$\|\tilde{Q}_m - Q_m\| \lesssim_P h^d \sqrt{\log n/(nh^d)} = o_P(h^d),$$

$$\|Q_m\| \lesssim_P h^d, \quad \|\hat{Q}_m^{-1}\| \lesssim_P h^{-d} \quad \text{and} \quad \|\tilde{Q}_m^{-1}\| \lesssim_P h^{-d}. $$
Lemma SA-8.2. Under Assumptions SA-1–SA-3, if \( \frac{\log n}{nh^d} = o(1) \), then

\[
\| \hat{Q}_m^{-1} - Q_m^{-1} \|_{\infty} \lesssim_P h^{-d} \sqrt{\log n/(nh^d)} = o_P(h^{-d})
\]
\[
\| \hat{Q}_m^{-1} - Q_m^{-1} \| \lesssim_P h^{-d} \sqrt{\log n/(nh^d)} = o_P(h^{-d}), \text{ and}
\]
\[
\| \hat{\Sigma}_0 - \Sigma_0 \| \lesssim_P h^d \sqrt{\log n/(nh^d)} = o_P(h^d).
\]

The next lemma shows that the asymptotic error expansion specified in Assumption SA-4 translates into the bias expansion for the classical point estimator \( \hat{\partial}q\mu_0(\cdot) \), which is also presented in Lemma 2.1 of the main paper. Importantly, it should be noted that the following lemma also constructs a uniform bound on the conditional bias which is crucial for least squares bias correction \((j = 2)\) and uniform inference. Let \( X = [x_1, \ldots, x_n]' \).

Lemma SA-8.3 (Conditional Bias). Let Assumptions SA-1–SA-4 hold. If \( \frac{\log n}{nh^d} = o(1) \), then,

\[
E[\hat{\partial}q\mu_0(x)|X] - \partial q\mu(x) = \mathcal{B}_{m,q}(x) - \hat{\gamma}_{q,0}(x)' E_n[\Pi_0(x_i)\mathcal{B}_{m,0}(x_i)] + O_P\left(h^{m+\varepsilon-[q]}\right)
\]

and

\[
\sup_{x \in \mathcal{X}} \left| E[\hat{\partial}q\mu_0(x)|X] - \partial q\mu(x) \right| \lesssim_P h^{m-[q]}.
\]

Moreover, if the following condition is also satisfied,

\[
\max_{1 \leq k \leq K} \int_{H_k} p_k(x; \Delta, m) \mathcal{B}_{m,0}(x) \, dx = o(h^{m+d}), \quad (SA-6)
\]

then \( \| \hat{\gamma}_{q,0}(x)' E_n[\Pi_0(x_i)\mathcal{B}_{m,0}(x_i)] \|_{L_{\infty}(\mathcal{X})} = o_P(h^{m-[q]}) \).

Equation (SA-6) implies that the leading approximation error is approximately orthogonal to \( p(\cdot) \), in which case the \( L_{\infty} \) approximation error coincides with the leading smoothing bias of \( \hat{\partial}q\mu_0(\cdot) \) (see Remark 1 in the main paper).

The next lemma shows that \( \hat{\gamma}_{q,j}(x) \) in (SA-1), (SA-2), (SA-3) and (SA-4) converge to their population counterpart \( \gamma_{q,j}(x) \) in a proper sense, and \( \gamma_{q,j}(x) \)'s are well bounded in terms of both \( L_{\infty} \) - and \( L_2 \)-operator norms.
**Lemma SA-8.4.** Let Assumptions SA-1–SA-3 and SA-5 hold. If $\frac{\log n}{nh^d} = o(1)$, then,

$$
\sup_{x \in \mathcal{X}} \|\gamma_{q,j}(x)'\|_\infty \lesssim h^{-d-[q]}, \quad \sup_{x \in \mathcal{X}} \|\hat{\gamma}_{q,j}(x)' - \gamma_{q,j}(x)'\|_\infty \lesssim \frac{\sqrt{n \log n}}{h h^d},
$$

$$
\sup_{x \in \mathcal{X}} \|\gamma_{q,j}(x)\| \lesssim h^{-d-[q]}, \quad \inf_{x \in \mathcal{X}} \|\gamma_{q,j}(x)\| \gtrsim h^{-d-[q]},
$$

$$
\sup_{x \in \mathcal{X}} \|\hat{\gamma}_{q,j}(x)' - \gamma_{q,j}(x)'\| \lesssim \frac{\sqrt{n \log n}}{h h^d}.
$$

The last lemma in this section proves that the asymptotic variance of classical and bias-corrected estimators is properly bounded. Importantly, we need not only an upper bound, but also a lower bound since the variance term appears in the denominator of $t$-statistics.

**Lemma SA-8.5.** Let Assumptions SA-1–SA-5 hold. If $\frac{\log n}{nh^d} = o(1)$,

$$
\sup_{x \in \mathcal{X}} \Omega_j \lesssim h^{-d-2[q]} \quad \text{and} \quad \inf_{x \in \mathcal{X}} \Omega_j(x) \gtrsim h^{-d-2[q]}.
$$

This lemma also implies that bias correction does not change the order of asymptotic variance.

**SA-9 Pointwise Estimation and Inference**

This section reports the pointwise results for classical and bias-corrected estimators. The first lemma establishes pointwise linearization.

**Lemma SA-9.1 (Pointwise Linearization).** Let Assumptions SA-1–SA-5 hold. If $\frac{\log n}{nh^d} = o(1)$, then

(a) \textit{(Classical Estimator) for each} $x \in \mathcal{X}$,

$$
\partial^q \mu_0(x) - \partial^q \mu(x) = \gamma_{q,0}(x)' E_n[\Pi_0(x_i)\varepsilon_i] + R_{1n,q}(x) + R_{2n,q}(x)
$$

where

$$
R_{1n,q}(x) := (\hat{\gamma}_{q,0}(x)' - \gamma_{q,0}(x)') E_n[\Pi_0(x_i)\varepsilon_i] \lesssim \frac{\sqrt{n \log n}}{n h^d + [q]},
$$

$$
R_{2n,q}(x) := E[\partial^q \mu_0(x)|\mathcal{X}] - \partial^q \mu(x) \lesssim n^{n-[q]}.
$$
(b) (Bias Correction) for each \( x \in \mathcal{X} \) and \( j = 1, 2, 3 \),

\[
\hat{\partial q}_j \mu_j(x) - \partial^q \mu(x) = \gamma_{q,j}(x)' \mathbb{E}_n[\Pi_j(x_i)\varepsilon_i] + R_{1n,q}(x) + R_{2n,q}(x)
\]

where

\[
R_{1n,q}(x) := (\gamma_{q,j}(x) - \gamma_{q,j}(x)') \mathbb{E}_n[\Pi_j(x_i)\varepsilon_i] \leq_p \frac{\sqrt{\log n}}{nh^d + |q|},
\]

\[
R_{2n,q}(x) := \mathbb{E}[\hat{\partial q}_j \mu_j(x)|\mathcal{X}] - \partial^q \mu(x) \lesssim_p h^{m+q-[q]}.
\]

**Remark SA-9.1.** The terms denoted by \( R_{2n,q}(x) \) in parts (a) and (b) of the pointwise linearization Lemma SA-9.1 capture the conditional bias of the point estimator. We establish a sharp bound on them by applying Bernstein’s maximal inequality to control the largest element of \( \mathbb{E}_n[\Pi_0(x_i)\mathcal{B}_{m,0}(x_i)] \) and employing the bound on the uniform norm of \( \hat{Q}^{-1}_m \) derived in Lemma SA-8.1. This result improves on previous results in the literature and, in particular, confirms a conjecture posed by [3, Comment 4.2(ii), p. 352], for the case of partitioning-based series estimation. To be precise, in their setup, \( \ell_{kc} \) can be understood as an uniform bound on the \( L_2 \) approximation error (orthogonal to the approximating basis). Thus, using our results, we obtain

\[
p(x)'(\hat{Q}^{-1} - Q^{-1}) G_n[p(x_i)\ell_{kc}] \lesssim_p \sqrt{\frac{\log n}{nh^d}} \cdot \sqrt{\frac{\log n}{h^d}} \ell_{kc},
\]

which coincides with the Equation (4.15) of [3], up to a normalization, thereby improving on the approximation established in their Equation (4.12). See our proof of Lemmas SA-8.3 and SA-9.1 for more details.

Given the above pointwise linearization, under proper rate restrictions the impacts of unknown design and conditional bias are negligible, and we can restrict our attention to the first term in the linearization. The next theorem constructs the desired asymptotic normality.

**Theorem SA-9.1 (Asymptotic Normality).** Suppose that Assumptions SA-1–SA-5 hold. In addition, \( \sup_{x \in \mathcal{X}} \mathbb{E}[\varepsilon_i^2 I \{|\varepsilon_i| > M\}|x_i = x] \to 0 \) as \( M \to \infty \), and \( \frac{\log n}{nh^d} = o(1) \).
(a) (Undersmoothing) If \( nh^{2m+d} = o(1) \), then

\[
\frac{\hat{\partial}^q \mu_0(x) - \partial^q \mu(x)}{\sqrt{\Omega_0(x)/n}} \rightsquigarrow N(0, 1); \quad (SA-1)
\]

(b) (Bias Correction) If \( nh^{2m+d} \leq 1 \), then for each \( j = 1, 2, 3 \),

\[
\frac{\hat{\partial}^q \mu_j(x) - \partial^q \mu(x)}{\sqrt{\Omega_j(x)/n}} \rightsquigarrow N(0, 1). \quad (SA-2)
\]

SA-10 Uniform Estimation and Inference

SA-10.1 Uniform Convergence

In this section, uniform convergence of the classical and bias-corrected estimators are obtained. In what follows we will derive all desired results under two versions of moment conditions: one assumes that the \((2 + \nu)\)th moment of errors is bounded for some \( \nu > 0 \); the other assumes an exponential moment is bounded. To begin with, we upgrade the pointwise linearization in Lemma SA-9.1 to a uniform version.

Lemma SA-10.1 (Uniform Linearization). Let Assumptions SA-1–SA-5 hold, and \( E[|\epsilon_i|^{2+\nu}] < \infty \) for some \( \nu > 0 \). If \( \frac{n^{\frac{\nu}{2+\nu}} (\log n)^{\frac{\nu}{2+\nu}}}{nh^d} \lesssim 1 \), then

(a) for the classical estimator,

\[
\frac{\hat{\partial}^q \mu_0(x) - \partial^q \mu(x)}{\sqrt{\Omega_0(x)/n}} = \gamma_q \mu_0(x) \quad E_n[\Pi_0(x_i)\epsilon_i] + R_{1n,q}(x) + R_{2n,q}(x)
\]

where \( R_{1n,q}(x) \) and \( R_{2n,q}(x) \) are defined in part (a) of Lemma SA-9.1 and

\[
\sup_{x \in \mathcal{X}} |R_{1n,q}(x)| \lesssim_{\mathbb{P}} \frac{\log n}{nh^{d+|q|}} =: \tilde{R}_{1n,q},
\]

\[
\sup_{x \in \mathcal{X}} |R_{2n,q}(x)| \lesssim_{\mathbb{P}} h^{m-|q|} =: \tilde{R}_{2n,q};
\]

(b) for bias correction approach \( j = 1, 2, 3 \),

\[
\frac{\hat{\partial}^q \mu_j(x) - \partial^q \mu(x)}{\sqrt{\Omega_j(x)/n}} = \gamma_q \mu_j(x) \quad E_n[\Pi_j(x_i)\epsilon_i] + R_{1n,q}(x) + R_{2n,q}(x)
\]
where $R_{1n,q}(x)$ and $R_{2n,q}(x)$ are defined in part (b) of Lemma SA-9.1 and

$$\sup_{x \in \mathcal{X}} |R_{1n,q}(x)| \lesssim \mathbb{P} \frac{\log n}{nh^{d+\lfloor q \rfloor}} =: \bar{R}_{1n,q},$$

$$\sup_{x \in \mathcal{X}} |R_{2n,q}(x)| \lesssim \mathbb{P} h^{m-\lfloor q \rfloor} =: \bar{R}_{2n,q}.$$

**Lemma SA-10.2.** Let Assumptions SA-1–SA-5 hold, and $\mathbb{E}[|\epsilon_i|^3 \exp(|\epsilon_i|)] < \infty$. If $\frac{(\log n)^3}{nh^2} \lesssim 1$, then the same results in Lemma SA-10.1 still hold.

Given the uniform linearization, the impacts of unknown designs and conditional bias are well bounded uniformly, and now we are ready to establish the uniform convergence of classical and bias-corrected estimators.

**Theorem SA-10.1 (Uniform Convergence).** Under the conditions of either Lemma SA-10.1 or Lemma SA-10.2,

(a) for the classical estimator,

$$\sup_{x \in \mathcal{X}} |\tilde{\partial^q \mu_0}(x) - \partial^q \mu(x)| \lesssim \mathbb{P} h^{-\lfloor (d/2+\lfloor q \rfloor)} \sqrt{\log n \left( \frac{\log n}{n} + \bar{R}_{1n,q} + \bar{R}_{2n,q} \right)}$$

$$\lesssim \mathbb{P} h^{-\lfloor (d/2+\lfloor q \rfloor)} \sqrt{\log n \left( \frac{\log n}{n} + h^{m-\lfloor q \rfloor} \right)} =: R_{uc_{q,0}}$$

where $\bar{R}_{1n,q}$ and $\bar{R}_{2n,q}$ are defined in Lemma SA-10.1(a);

(b) for bias-corrected estimators $j = 1, 2, 3$,

$$\sup_{x \in \mathcal{X}} |\tilde{\partial^q \mu_j}(x) - \partial^q \mu(x)| \lesssim \mathbb{P} h^{-\lfloor (d/2+\lfloor q \rfloor)} \sqrt{\log n \left( \frac{\log n}{n} + \bar{R}_{1n,q} + \bar{R}_{2n,q} \right)}$$

$$\lesssim \mathbb{P} h^{-\lfloor (d/2+\lfloor q \rfloor)} \sqrt{\log n \left( \frac{\log n}{n} + h^{m+\lfloor q \rfloor} \right)} =: R_{uc_{q,j}}$$

where $\bar{R}_{1n,q}$ and $\bar{R}_{2n,q}$ are defined in Lemma SA-10.1(b).

For inference purpose, one may interested in the following $t$-statistic

$$\hat{T}_j(x) = \frac{\tilde{\partial^q \mu_j}(x) - \mu(x)}{\sqrt{\hat{\Omega}_j(x)/n}}.$$
The denominator is an estimate of conditional variance of $\hat{q}_{\mu_j}(x)$ defined in Section SA-8. The next theorem proves the uniform consistency of these estimates.

**Theorem SA-10.2** (Variance Estimate). Let Assumptions SA-1–SA-5 hold. If $\mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ for some $\nu > 0$, and $\frac{n^{\frac{1}{2\nu}}(\log n)^{\nu}}{nh^d} = o(1)$, then

(a) for the classical estimator,

$$
|\hat{\Sigma}_0 - \Sigma_0| \lesssim_P h^d \left( R^\text{uc}_{0,0} + \frac{n^{\frac{1}{2\nu}}(\log n)^{\nu}}{\sqrt{nh^d}} \right) = o_P(h^d), \text{ and }
$$

$$
\sup_{x \in \mathcal{X}} |\hat{\Omega}_0(x) - \Omega_0(x)| \lesssim_P h^{-d-2\lfloor q \rfloor} \left( R^\text{uc}_{0,0} + \frac{n^{\frac{1}{2\nu}}(\log n)^{\nu}}{\sqrt{nh^d}} \right) = o_P(h^{-d-2\lfloor q \rfloor})
$$

where $R^\text{uc}_{0,0}$ is the uniform convergence rate given in Theorem SA-10.1(a) with $q = 0$;

(b) for bias-corrected estimators $j = 1, 2, 3$,

$$
|\hat{\Sigma}_j - \Sigma_j| \lesssim_P h^d \left( R^\text{uc}_{0,j} + \frac{n^{\frac{1}{2\nu}}(\log n)^{\nu}}{\sqrt{nh^d}} \right) = o_P(h^d), \text{ and }
$$

$$
\sup_{x \in \mathcal{X}} |\hat{\Omega}_j(x) - \Omega_j(x)| \lesssim_P h^{-d-2\lfloor q \rfloor} \left( R^\text{uc}_{0,j} + \frac{n^{\frac{1}{2\nu}}(\log n)^{\nu}}{\sqrt{nh^d}} \right) = o_P(h^{-d-2\lfloor q \rfloor})
$$

where $R^\text{uc}_{0,j}$ is the uniform convergence rate given in Theorem SA-10.1(b) with $q = 0$.

**Theorem SA-10.3.** Let Assumptions SA-1–SA-5 hold. If $\mathbb{E}[|\varepsilon_i|\exp(|\varepsilon_i|)] < \infty$ and $\frac{(\log n)^3}{nh^d} = o(1)$, then

(a) for the classical estimator,

$$
|\hat{\Sigma}_0 - \Sigma_0| \lesssim_P h^d \left( R^\text{uc}_{0,0} + \frac{(\log n)^{3/2}}{\sqrt{nh^d}} \right) = o_P(h^d), \text{ and }
$$

$$
\sup_{x \in \mathcal{X}} |\hat{\Omega}_0(x) - \Omega_0(x)| \lesssim_P h^{-d-2\lfloor q \rfloor} \left( R^\text{uc}_{0,0} + \frac{(\log n)^{3/2}}{\sqrt{nh^d}} \right) = o_P(h^{-d-2\lfloor q \rfloor}).
$$

(b) for bias-corrected estimators $j = 1, 2, 3$,

$$
|\hat{\Sigma}_j - \Sigma_j| \lesssim_P h^d \left( R^\text{uc}_{0,j} + \frac{(\log n)^{3/2}}{\sqrt{nh^d}} \right) = o_P(h^d), \text{ and }
$$

51
\[ \sup_{x \in \mathcal{X}} \left| \hat{\Omega}_j(x) - \Omega_j(x) \right| \lesssim \mathbb{P} \left( R_{0,j}^{\text{uc}} + \frac{(\log n)^{3/2}}{\sqrt{nh^d}} \right) = o_P(h^{-d-2[q]}) . \]

### SA-10.2 Strong Approximation

We move on to uniform inference. Let \( r_n \) be a positive non-vanishing sequence, which will be used to denote the approximation error rate in the following analysis. As the first step, we employ Lemma SA-8.4 and Theorem SA-10.2 (or SA-10.3) to show that the sampling and estimation uncertainty of \( \hat{\gamma}_q(x) \) and \( \hat{\Omega}_j(x) \) are negligible uniformly over \( x \in \mathcal{X} \). Specifically, define

\[ t_j(x) = \frac{\gamma_{q,j}(x)}{\sqrt{\Omega_j(x)}} \mathbb{P}_{\mathbb{N}}[ \Pi_j(x_i) \varepsilon_i ] , \quad x \in \mathcal{X}, \quad j = 0, 1, 2, 3. \]

The next lemma shows that \( \hat{T}_j(\cdot) \) can be approximated by \( t_j(\cdot) \) uniformly.

**Lemma SA-10.3.** Let Assumptions SA-1-SA-5 hold. In addition, assume \( \mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty \) for some \( \nu > 0 \) and

\[ \frac{n^{\frac{2}{2+\nu}} (\log n)^{\frac{2+2\nu}{2+\nu}}}{nh^d} = o(r_n^{-2}) . \]

(a) For the classical estimator, if \( nh^{d+2m} = o(r_n^{-2}) \), then

\[ \sup_{x \in \mathcal{X}} | \hat{T}_0(x) - t_0(x) | = o_P(r_n^{-1}) \]

(b) For bias-corrected estimators, if \( nh^{d+2m+2\delta} = o(r_n^{-2}) \), then

\[ \sup_{x \in \mathcal{X}} | \hat{T}_j(x) - t_j(x) | = o_P(r_n^{-1}) , \quad j = 1, 2, 3. \]

**Lemma SA-10.4.** Let Assumptions SA-1-SA-5 hold. In addition, assume \( \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] < \infty \) and \( \frac{(\log n)^6}{nh^d} = o(r_n^{-2}) \). Then the results for \( j = 0 \) and \( j = 1, 2, 3 \) in Lemma SA-10.3 still hold under respective conditions.

Now our task reduces to construct valid distributional approximation to \( t_j(\cdot) \) for each \( j = 0, 1, 2, 3 \) in a proper sense. Specifically, we want to show that on a sufficiently rich enough probability space

52
there exists a copy $t'_j(\cdot)$ of $t_j(\cdot)$ and a Gaussian process $Z_j(\cdot)$ such that

$$\sup_{x \in \mathcal{X}} |t'_j(x) - Z_j(x)| = o_P(r_n^{-1}).$$

When such a construction is possible, we can use $Z_j(\cdot)$ to approximate the distribution of $t_j(\cdot)$, as well as $\hat{T}_j(\cdot)$ in view of Lemma SA-10.3 and SA-10.4. To save our notation, we denote this strong approximation by $\hat{T}_j(\cdot) = d Z_j(\cdot) + o_P(r_n^{-1})$ in $L^\infty(\mathcal{X})$ where $L^\infty(\mathcal{X})$ refers to the set of all uniformly bounded real functions on $\mathcal{X}$ equipped with uniform norm.

As an intermediate step towards the final Gaussian approximation to $\hat{T}_j(\cdot)$, our next theorem will construct approximating processes which are Gaussian conditional on the covariates. Our proof strategy relies on KMT coupling techniques that apply to the partial sum of independent but not necessarily identically distributed random variables. Specifically, we will employ a slightly modified version to account for the lack of i.i.d. in the data, conditional on $\mathcal{X}$, which is a direct application of [53]. We first give a more general result about conditional coupling, stated in the following lemma, and then apply it to the partitioning-based estimators considered in our paper.

**Lemma SA-10.5.** Suppose that Assumption SA-1(a) holds. Let $\{G_n[\mathcal{K}(x, x_i)\varepsilon_i] : x \in \mathcal{X}\}$ be a stochastic process of interest with $\mathcal{K}(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ an $n$-varying function, which could also be a function of $X$ but this extra generality is not needed in this paper.

**(a)** Assume that $\sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^2 + \nu |x_i = x] < \infty$ for some $\nu > 0$. If

$$\sup_{x \in \mathcal{X}} \max_{1 \leq i \leq n} |\mathcal{K}(x, x_i)| = o_P(r_n^{-1} n^{-\frac{1}{2+\nu} + \frac{3}{2}}),$$

then there exists a sequence of i.i.d standard normal random variables $\zeta_i$, $i = 1, 2, \ldots, n$ such that $G_n[\mathcal{K}(x, x_i)\varepsilon_i] \to_d G_n[\mathcal{K}(x, x_i)\sigma(x_i)\zeta_i] + o_P(r_n^{-1})$ in $L^\infty(\mathcal{X})$.

**(b)** Assume that $\sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)|x_i = x] < \infty$. If

$$\sup_{x \in \mathcal{X}} \max_{1 \leq i \leq n} |\mathcal{K}(x, x_i)| = o_P(r_n^{-1} (\log n)^{-1} \sqrt{n}),$$

then there exists a sequence of i.i.d standard normal random variables $\zeta_i$, $i = 1, 2, \ldots, n$ such that $G_n[\mathcal{K}(x, x_i)\varepsilon_i] \to_d G_n[\mathcal{K}(x, x_i)\sigma(x_i)\zeta_i] + o_P(r_n^{-1})$ in $L^\infty(\mathcal{X})$. 53
For the local basis considered in this paper, it is clear that for each \( j = 0, 1, 2, 3, \)

\[
\sup_{x \in \mathcal{X}} \sup_{u \in \mathcal{U}} |K(x, u)| \lesssim h^{-d/2}, \quad \text{with} \quad K(x, u) = \frac{\gamma_{q,j}(x)\Pi_j(u)}{\sqrt{\Omega_j(x)}},
\]

by which we can directly verify the conditions in Lemma SA-10.5 and obtain the following theorems.

**Theorem SA-10.4 (Conditional Coupling).** Let Assumptions SA-1–SA-5 hold. Assume \( \mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty, \sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}] |x_i = x| < \infty \) for some \( \nu, \bar{\nu} > 0 \). In addition,

\[
\frac{n^2 (\log n)^{2+2\nu}}{nh^d} = o(r_n^{-2}) \quad \text{and} \quad \frac{n^2 \nu}{nh^d} = o(r_n^{-2}).
\]

(a) (Undersmoothing) If \( nh^{d+2m} = o(r_n^{-2}) \), then there exists a sequence of i.i.d standard normal random variables \( \zeta_i, i = 1, 2, \ldots, n \) such that

\[
\hat{T}_0(\cdot) = d z_0(\cdot) + o_P(r_n^{-1}) \quad \text{in} \quad \mathcal{L}^\infty(\mathcal{X}), \quad z_0(\cdot) = \frac{\gamma_{q,0}(\cdot)}{\sqrt{\Omega_0(\cdot)}} G_n[\Pi_0(x_i)\zeta_i\sigma(x_i)]. \quad \text{(SA-1)}
\]

(b) (Robust Bias Correction) If \( nh^{d+2m+2\nu} = o(r_n^{-2}) \), then for \( j = 1, 2, 3, \) there exists a sequence of i.i.d standard normal random variables \( \zeta_i, i = 1, 2, \ldots, n \) such that

\[
\hat{T}_j(\cdot) = d z_j(\cdot) + o_P(r_n^{-1}) \quad \text{in} \quad \mathcal{L}^\infty(\mathcal{X}), \quad z_j(\cdot) = \frac{\gamma_{q,j}(\cdot)}{\sqrt{\Omega_j(\cdot)}} G_n[\Pi_j(x_i)\zeta_i\sigma(x_i)]. \quad \text{(SA-2)}
\]

**Theorem SA-10.5.** Let Assumptions SA-1–SA-5 hold. Assume \( \sup_{x \in \mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] |x_i = x| < \infty \). In addition, \( \frac{(\log n)^4}{nh^d} = o(r_n^{-2}) \).

(a) (Undersmoothing) If \( nh^{d+2m} = o(r_n^{-2}) \), then the result in \( \text{(SA-1)} \) still holds.

(b) (Robust Bias Correction) If \( nh^{d+2m+2\nu} = o(r_n^{-2}) \), then the result in \( \text{(SA-2)} \) still holds.

It can be seen from the proof of Theorem SA-10.4 or SA-10.5 that conditional on all covariates, our task reduces to constructing Gaussian approximation for the partial sum of errors, which is a one-dimensional strong approximation problem. This important feature enables us to avoid the more stringent conditions needed for multivariate coupling.

The next theorem completes the construction of the (unconditional) Gaussian approximation, which is stated as Theorem 3.1 in the main paper. Specifically, the approximating processes are
given by
\[
Z_j(x) = \frac{\gamma_{q,j}(x)\Sigma_{1/2}^{1/2}}{\sqrt{\Omega_j(x)}} N_{K_j}, \quad x \in \mathcal{X}, \quad \text{for} \quad j = 0, 1, 2, 3.
\]

where \(N_{K_j}\) is \(K_j\)-dimensional standard normal vector with \(K_j = \dim(\Pi_j(\cdot))\).

**Theorem SA-10.6** (Unconditional Coupling). Suppose the conditions of Theorem SA-10.4 or Theorem SA-10.5 hold.

(a) (Undersmoothing) If \(nh^{d+2m} = o(r_n^{-2})\), then there exists a sequence of \(K_0\)-dimensional standard normal vectors \(\{N_{K_0}\}\) in a properly enlarged probability space such that
\[
\hat{T}_0(\cdot) = Z_0(\cdot) + o_P(r_n^{-1}) \quad \text{in} \quad \mathcal{L}^\infty(\mathcal{X}).
\]

(b) (Robust Bias Correction) Assume that \(nh^{d+2m+2\eta} = o(r_n^{-2})\) for \(j = 1, 2, 3\). In addition, for \(j = 2, 3\), further assume \(\frac{(\log n)^{3/2}}{\sqrt{nh^d}} = o(r_n^{-2})\). Then there exists a sequence of \(K_j\)-dimensional standard normal vectors \(\{N_{K_j}\}\) in a properly enlarged probability space such that
\[
\hat{T}_j(\cdot) = Z_j(\cdot) + o_P(r_n^{-1}) \quad \text{in} \quad \mathcal{L}^\infty(\mathcal{X}), \quad j = 1, 2, 3.
\]

**SA-10.3 Implementation**

\(Z_j(\cdot)\), as an approximating process for \(\hat{T}_j(\cdot)\), is still infeasible. Our next objective is to construct practicably feasible distributional approximation for \(\hat{T}_j(\cdot)\). To be precise, we construct Gaussian processes \(\{\hat{Z}_j(x) : x \in \mathcal{X}\}\), with distributions known conditional on the data \((y, X)\), such that there exists a copy \(\hat{Z}_j(\cdot)\) of \(\hat{Z}_j(\cdot)\) in a sufficiently rich probability space and (i) \(\hat{Z}_j(\cdot) =_d Z_j(\cdot)\) conditional on the data \((y, X)\) and (ii) for some positive non-vanishing sequence \(r_n\) and for all \(\eta > 0\),
\[
P^*[\sup_{x \in \mathcal{X}} |\hat{Z}_j(x) - Z_j(x)| \geq \eta r_n^{-1}] = o_P(1),
\]

where \(P^*[\cdot] = P[\cdot|y, X]\) denotes the probability operator conditional on the data. When such a feasible process exists, we write \(\hat{Z}_j(\cdot) =_d Z_j(\cdot) + o_P(r_n^{-1})\) in \(\mathcal{L}^\infty(\mathcal{X})\). From a practical perspective, sampling from \(\hat{Z}_j(\cdot)\), conditional on the data, is possible and provides a valid distributional approximation.
Our first construction is a direct plug-in approach using the conclusion of Theorem SA-10.6. All unknown objects are replaced by consistent estimators already used in the feasible $t$-statistics:

$$
\hat{Z}_j(x) = \frac{\hat{\gamma}_{q,j}(x)'\hat{\Sigma}_j^{1/2}N_{K_j}, \quad x \in \mathcal{X}, \quad j = 0, 1, 2, 3}
$$

**Theorem SA-10.7** (Plug-in Approximation). Let the conditions in Theorem SA-10.4 hold.

(a) **(Undersmoothing)** Assume that $nh^{d+2m} = o(r_n^{-2})$. Then $\hat{Z}_0(\cdot) = d^* \hat{Z}_0(\cdot) + o_p(r_n^{-1})$ in $L^\infty(\mathcal{X})$.

(b) **(Robust Bias Correction)** Assume that $nh^{d+2m+2\nu} = o(r_n^{-2})$ for $j = 1, 2, 3$. In addition, for $j = 2, 3$, further assume

$$
\frac{n^{\frac{3}{4+2\nu}}(\log n)^{\frac{5}{2}}}{\sqrt{nh^d}} = o(r_n^{-2}).
$$

Then $\hat{Z}_j(\cdot) = d^* Z_j(\cdot) + o_p(r_n^{-1})$ in $L^\infty(\mathcal{X})$, $j = 1, 2, 3$.

**Theorem SA-10.8.** Let the conditions of Theorem SA-10.5 hold.

(a) **(Undersmoothing)** Assume that $nh^{d+2m} = o(r_n^{-2})$. Then $\hat{Z}_0(\cdot) = d^* \hat{Z}_0(\cdot) + o_p(r_n^{-1})$ in $L^\infty(\mathcal{X})$.

(b) **(Robust Bias Correction)** Assume that $nh^{d+2m+2\nu} = o(r_n^{-2})$ for $j = 1, 2, 3$. In addition, for $j = 2, 3$, further assume

$$
\frac{(\log n)^{\frac{5}{2}}}{\sqrt{nh^d}} = o(r_n^{-2}).
$$

Then $\hat{Z}_j(\cdot) = d^* Z_j(\cdot) + o_p(r_n^{-1})$ in $L^\infty(\mathcal{X})$ for $j = 1, 2, 3$.

Our second construction, which is not reported in the main paper, is based on the intermediate approximating processes $z_j(\cdot)$ in (SA-1) and (SA-2). Again, $\gamma_{q,j}(x)$ and $\Omega_j(x)$ can be simply replaced by their sample analogues, and $\zeta_i$’s can be generated by sampling from an $n$-dimensional standard Gaussian vector independent of the data. The key unknown quantity is the conditional variance function $\sigma^2(x) = E[\epsilon_i^2 | x_i = x]$. If the errors are homoskedastic, one can simply let $\sigma^2(\cdot) = 1$. If not, we need an estimator of the conditional variance satisfying a mild condition on uniform convergence, and then we can construct

$$
\tilde{z}_j(x) = \frac{\tilde{\gamma}_{q,j}(x)'\tilde{\Omega}_j^{1/2}\mathbb{G}_n[\Pi_j(x_i)\zeta_i\tilde{\sigma}(x_i)]}{\sqrt{\tilde{\Omega}_j(x)}}, \quad x \in \mathcal{X}, \quad j = 0, 1, 2, 3.
$$

In practice one may simply construct a nonparametric estimator of $\sigma^2(\cdot)$ by using, for example,
regression splines or smoothing splines. See [28, Chapter 22.4] for more details. We do not elaborate on this issue, but the next theorem gives the key ingredient of this approach.

**Theorem SA-10.9** (Plug-in Approximation: Conditional Coupling). Let the conditions in either Theorem SA-10.4 or SA-10.5 hold. In addition, \( \hat{\sigma}^2(\cdot) \) satisfies \( \max_{1 \leq i \leq n} |\hat{\sigma}^2(x_i) - \sigma^2(x_i)| = o_P(r_n^{-1}(\log n)^{-1/2}) \).

(a) **(Undersmoothing)** Assume that \( nh^{d+2m} = o(r_n^{-2}) \). Then \( \hat{z}_0(\cdot) = d^* Z_0(\cdot) + o_{P^*}(r_n^{-1}) \) in \( L^\infty(\mathcal{X}) \).

(b) **(Robust Bias Correction)** Assume that \( nh^{d+2m+2\theta} = o(r_n^{-2}) \). In addition, for \( j = 2, 3 \), further assume \( \frac{(\log n)^{3/2}}{\sqrt{nh^d}} = o(r_n^{-2}) \). Then \( \hat{z}_j(\cdot) = d^* Z_j(\cdot) + o_{P^*}(r_n^{-1}) \) in \( L^\infty(\mathcal{X}) \), \( j = 1, 2, 3 \).

Our third approach to approximating the infeasible \( Z_j(\cdot) \) employs an easy-to-implement wild bootstrap procedure. Specifically, for zero-mean unit-variance i.i.d. bounded random variables \( \{\omega_i : 1 \leq i \leq n\} \) independent of the data, we construct bootstrapped \( t \)-statistics:

\[
\hat{z}_j^*(x) = \frac{\hat{\gamma}_{\alpha,j}(x)}{\sqrt{\hat{\Omega}_j^*(x)}} G_n[\Pi_j(x_i) \omega_i \hat{\varepsilon}_{i,j}], \quad x \in \mathcal{X}, \quad j = 0, 1, 2, 3,
\]

where \( \hat{\varepsilon}_{i,j} \) is defined in (SA-5), and the bootstrap Studentization \( \hat{\Omega}_j^*(x) \) is constructed using \( \hat{\Sigma}_j^* = \mathbb{E}_n[\Pi_j(x_i) \Pi_j(x_i)' \omega_i^2 \hat{\varepsilon}_{i,j}^2] \).

**Theorem SA-10.10** (Wild Bootstrap). Suppose that the conditions of Theorem SA-10.4 hold. Let \( \{\omega_i : 1 \leq i \leq n\} \) be an i.i.d sequence of bounded random variables with \( \mathbb{E}[\omega_i] = 0 \) and \( \mathbb{E}[\omega_i^2] = 1 \) independent of the data.

(a) **(Undersmoothing)** Assume that \( nh^{d+2m} = o(r_n^{-2}) \) and \( n^{2/\nu}(\log n)^2/(nh^d) = o(r_n^{-2}) \). Then \( \hat{z}_0^*(\cdot) = d^* Z_0(\cdot) + o_{P^*}(r_n^{-1}) \) in \( L^\infty(\mathcal{X}) \).

(b) **(Robust Bias Correction)** Assume that \( nh^{d+2m+2\theta} = o(r_n^{-2}) \) for \( j = 1, 2, 3 \). In addition,

- for \( j = 1 \), assume \( n^{2/\nu}(\log n)^2/(nh^d) = o(r_n^{-2}) \);
- for \( j = 2, 3 \), assume \( n^{1/\nu}(\log n)^{4+2\theta}/\sqrt{nh^d} = o(r_n^{-2}) \).

Then \( \hat{z}_j^*(\cdot) = d^* Z_j(\cdot) + o_{P^*}(r_n^{-1}) \) in \( L^\infty(\mathcal{X}) \) for \( j = 1, 2, 3 \).
Theorem SA-10.11. Suppose that the conditions of Theorem SA-10.5 hold. Let \( \{\omega_i : 1 \leq i \leq n\} \) be an i.i.d sequence of bounded random variables with \( \mathbb{E}[\omega_i] = 0 \) and \( \mathbb{E}[\omega_i^2] = 1 \) independent of the data.

(a) (Undersmoothing) Assume that \( nh^{d+2m} = o(r_n^{-2}) \). Then \( \hat{Z}_0^*(\cdot) = d_+ Z_0^*(\cdot) + o_P(r_n^{-1}) \).

(b) (Robust Bias Correction) Assume that \( nh^{d+2m+2m} = o(r_n^{-2}) \). In addition, for \( j = 2, 3 \), further assume \( \frac{(\log n)^{5/2}}{\sqrt{nh^2}} = o(r_n^{-2}) \). Then \( \hat{Z}_j^*(\cdot) = d_+ Z_j^*(\cdot) + o_P(r_n^{-1}) \) for \( j = 1, 2, 3 \).

SA-10.4 Application: Confidence Bands

A natural application of Theorems SA-10.6, SA-10.7, SA-10.9, and SA-10.10 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

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

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

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

The following theorem establishes a valid distributional approximation for the suprema of the \( t \)-statistic processes \( \{\hat{T}_j(x) : x \in \mathcal{X}\} \) using a [17, Lemma 2.4] to convert our strong approximation results into convergence of distribution functions in terms of Kolmogorov distance.

Theorem SA-10.12 (Confidence Band). Let the conditions of Theorem SA-10.6 hold with \( r_n = \sqrt{\log n} \). If the corresponding conditions of Theorem SA-10.7 or SA-10.8 for each \( j = 0, 1, 2, 3 \) hold, then

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

If the corresponding conditions in Theorem SA-10.9 hold for each \( j = 0, 1, 2, 3 \), then

\[
\sup_{u \in \mathbb{R}} \left| P \left[ \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| \leq u \right] - P^* \left[ \sup_{x \in \mathcal{X}} |\hat{Z}_j(x)| \leq u \right] \right| = o_P(1).
\]
If the corresponding conditions in Theorem SA-10.10 or SA-10.11 hold for each $j = 0, 1, 2, 3$, then
\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left[ \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| \leq u \right] - \mathbb{P}^* \left[ \sup_{x \in \mathcal{X}} |z_j^*(x)| \leq u \right] \right| = o_P(1).
\]

### SA-11 Discussion and Extensions

#### SA-11.1 Connecting Splines and Piecewise Polynomials

There is an important connection between splines and piecewise polynomial bases. Essentially, the former can be viewed as a piecewise polynomial basis with certain continuity restrictions. Therefore, the estimators based on the two bases are linked by utilizing the well-known results about regression with linear constraints. See [6] for an illustration of cubic splines as a special case of restricted least squares.

Formally, let us start with the (rotated) piecewise polynomials $p$ discussed in Appendix A. For expositional simplicity, we only discuss rectangular partition here. Clearly, $p$ spans a vector space $\mathcal{P}$ containing all “piecewise” polynomials with degree no greater than $m - 1$ on $\mathcal{X}$:
\[
\mathcal{P} := \left\{ s(\cdot) : s(\cdot) = \sum_{k=1}^{K} a_k p_k(\cdot), a_k \in \mathbb{R} \right\}.
\]

Functions in this space are continuous within each subrectangle, but might have “jumps” along boundaries of cells. Then by imposing certain continuity restrictions on functions in $\mathcal{P}$, we can construct a subspace $\mathcal{S} \subset \mathcal{P}$
\[
\mathcal{S} := \left\{ s(\cdot) : s(\cdot) = \sum_{k=1}^{K} a_k p_k(\cdot), a_k \in \mathbb{R}, \text{ and } \partial^\varsigma s(\cdot) \in C(\mathcal{X}), \forall [\varsigma] \leq \iota \right\}.
\]

where $\iota \leq m - 2$ is a positive integer that controls the smoothness of the basis (and thus the smoothness of the estimated function). Since the derivatives of polynomials are linear in coefficients, the continuity constraints are linear.

Now consider the following restricted least squares:
\[
\hat{\beta}_r = \arg \min_{b \in \mathbb{R}^N} \| y - Pb \|^2 \quad \text{s.t.} \quad Rb = 0
\]
where \( P = (p(x_1), \ldots, p(x_n))' \) and \( R \) is a \( \vartheta \times K \) restriction matrix. \( \vartheta \) denotes the number of restrictions depending on the required smoothness \( \iota \). If there are no redundant constraints, \( R \) has full row rank. It is well known that given the restriction matrix \( R \), the least squares estimator can be written as

\[
\hat{\beta}_r = [I - (P'P)^{-1}R'(R(P'P)^{-1}R')^{-1}R](P'P)^{-1}P'y.
\]

Since the unrestricted least squares estimator \( \hat{\beta}_{ur} = (P'P)^{-1}P'y \), the above equation establishes the relation between restricted and unrestricted estimators:

\[
\hat{\beta}_r = [I - (P'P)^{-1}R'(R(P'P)^{-1}R')^{-1}R]\hat{\beta}_{ur} =: (I - U)\hat{\beta}_{ur}.
\]

Therefore,

\[
\hat{\mu}_r(x) = p(x)'(I - U)\hat{\beta}_{ur} = \hat{\mu}_{ur}(x) - p(x)'U\hat{\beta}_{ur}
\]

where \( \hat{\mu}_{ur}(x) := p(x)'\hat{\beta}_{ur} \) is the unrestricted estimator. With this relation we can derive an expression of bias and variance for the restricted estimator. Clearly, the conditional variance

\[
\mathbb{V}[\hat{\mu}_r(x)|X] = p(x)'(I - U)\mathbb{V}[\hat{\mu}_{ur}(x)|X](I - U)p(x).
\]

On the other hand, as shown in Lemma A.3 of Appendix A, there exists \( s^*(x) := p(x)'\beta^* \) such that \( \|\mu - s^* + \mathcal{R}_{m,0}\|_{L_{\infty}(X)} = o(h^m) \). Hence

\[
\mathbb{E}[\hat{\mu}_r(x)|X] - \mu(x) = p(x)'(I - U)(P'P)^{-1}P'[P\beta^* - \mathcal{R}_{m,0} + o(h^m)] - \mu(x)
\]

\[
= \mathcal{R}_{m,0}(x) - p(x)'(I - U)(P'P)^{-1}P'[\mathcal{R}_{m,0} + o(h^m)] - p(x)'U\beta^* + o(h^m)
\]

where \( \mathcal{R}_{m,0} = (\mathcal{R}_{m,0}(x_1), \ldots, \mathcal{R}_{m,0}(x_n))' \). \( p(x)'U\beta^* \) can be viewed as a measure of to what extent the continuity constraints are satisfied. When \( s^*(x) \) is continuous up to order \( \iota \), \( R\beta^* \) is exactly 0 and thus this term vanishes.

To repeat the previous analysis for such a restricted estimator, the main challenge is to analyze the asymptotic properties of the outer product of the restriction matrix. Once we know its limiting eigenvalue distributions (e.g., bounds on extreme eigenvalues), \( U \) can be properly bounded, and then the conclusions in previous sections may be established with similar proofs. Unfortunately, for general multidimensional cases, it is difficult and tedious to specify a non-redundant set of continuity
constraints and analyze the eigenvalue distributions of $RR'$, thus damping the usefulness of such a method, whereas when $d = 1$, the restriction matrix is quite straightforward and well bounded.

Formally, let $R$ be a restriction matrix corresponding to $\bar{\iota} = \iota + 1$ continuity constraints at each partitioning knot, implying that $s \in S$ has $\iota$ continuous derivatives. We write the $\ell$th restriction at the $k$th knot as $r_{k\ell}$, corresponding to the requirement that the $(\ell - 1)$th derivatives of functions in $S$ are continuous. Explicitly, the entire restriction matrix admits the following structure:

$$
R = \begin{bmatrix}
    r'_{11} \\
    r'_{12} \\
    \vdots \\
    r'_{k\ell} \\
    \vdots \\
    r'_{\kappa1} \\
    \vdots \\
    r'_{\kappa\ell}
\end{bmatrix}
$$

Restrictions at the 1st knot

Restrictions at the $\kappa$th knot

(SA-2)

As $\kappa$ increases, the dimension of $R$ also grows. Moreover, $\iota$ cannot exceed $m - 2$ since when $m$ continuity constraints are imposed, $S$ degenerates to a space of global polynomials of degree no greater than $m - 1$.

The asymptotic behavior of the restricted estimator is closely related to the bounds on extreme eigenvalues of $RR'$. For a general restriction matrix $R$ with fixed dimensions, if it only contains non-redundant constraints, the minimum eigenvalue of $RR'$ is nonzero. When the number of constraints $\vartheta \iota \to \infty$, however, the limit of the minimum eigenvalue does not have to be nonzero, and its limiting behavior depends on the specific structure of constraints. The next lemma shows that for the particular restricted estimators considered here, the eigenvalues of $RR'$ is indeed bounded and bounded away from zero uniformly over the number of knots for $\iota \leq m - 2$.

**Lemma SA-11.1** (Restriction Matrix). *Let $R$ be the restriction matrix described as (SA-2) with $\iota \leq m - 2$. Then*

$$
1 \lesssim \lambda_{\min}(RR') \leq \lambda_{\max}(RR') \lesssim 1.
$$

The proof of this lemma employs the specific structure of the restriction matrix. Generally, the
outer product of $\mathbf{R}$ takes the following form:

$$
\mathbf{R} \mathbf{R}' = \begin{bmatrix}
\mathbf{A} & \mathbf{B} & 0 & \cdots & 0 \\
\mathbf{B}' & \mathbf{A} & \mathbf{B} & 0 & \cdots \\
& \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \mathbf{B}' & \mathbf{A} & \mathbf{B} \\
0 & \cdots & \mathbf{B}' & \mathbf{A}
\end{bmatrix}
$$

(SA-3)

where

$$
\mathbf{A} = \begin{bmatrix}
r'_{11r_{11}} & r'_{11r_{12}} & \cdots & r'_{11r_{1k}} \\
r'_{12r_{11}} & r'_{12r_{12}} & \cdots & r'_{12r_{1k}} \\
& \ddots & \ddots & \vdots \\
r'_{1kr_{11}} & r'_{1kr_{12}} & \cdots & r'_{1kr_{1k}}
\end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix}
r'_{11r_{21}} & r'_{11r_{22}} & \cdots & r'_{11r_{2k}} \\
r'_{12r_{21}} & r'_{12r_{22}} & \cdots & r'_{12r_{2k}} \\
& \ddots & \ddots & \vdots \\
r'_{kr_{21}} & r'_{kr_{22}} & \cdots & r'_{kr_{2k}}
\end{bmatrix}
$$

Importantly, the form described in (SA-3) is usually referred to as a (tridiagonal) block Toeplitz matrix, meaning that it is a tridiagonal block matrix containing blocks repeated down the diagonals. It is well known that its asymptotic eigenvalue distribution is characterized by the Fourier transform

$$
\mathcal{F}_\ell(\omega) = \mathbf{A} + (\mathbf{B} + \mathbf{B}') \cos \omega, \quad \omega \in [0, 2\pi].
$$

As $\kappa \to \infty$, $\lambda_{\min}(\mathbf{R} \mathbf{R}')$ converges to the minimum attained by the minimum eigenvalue of $\mathcal{F}_\ell(\omega)$ as a function of $\omega$ on $[0, 2\pi]$. Similarly, the limit of $\lambda_{\max}(\mathbf{R} \mathbf{R}')$ is the maximum attained by the maximum eigenvalue of $\mathcal{F}_\ell(\omega)$.

### SA-11.2 Comparison of Bias Correction Approaches

We make some comparison of the three bias correction approaches considered in this paper.

First, the higher-order correction ($j = 1$) and least-squares correction ($j = 2$) are closely related. Let $\mathbf{P} = (\mathbf{p}(x_1), \ldots, \mathbf{p}(x_n))'$ and $\tilde{\mathbf{P}} = (\tilde{\mathbf{p}}(x_1), \ldots, \tilde{\mathbf{p}}(x_n))'$. Clearly,

$$
\hat{\mu}_2(x) = \hat{\mu}_1(x) + \mathbf{p}(x)'(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'(\mathbf{I} - \tilde{\mathbf{P}}(\tilde{\mathbf{P}}'\tilde{\mathbf{P}})^{-1}\tilde{\mathbf{P}})\mathbf{y}
= \hat{\mu}_1(x) + \mathbf{p}(x)'(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{M}\tilde{\mathbf{p}}\mathbf{y}
$$
where $M_{\tilde{p}} := I - \tilde{P}(\tilde{P}'\tilde{P})^{-1}\tilde{P}'$. Importantly, when $p$ and $\tilde{p}$ generate nested models, i.e., there exists a transformation matrix $\Upsilon$ such that $p(\cdot) = \Upsilon \tilde{p}(\cdot)$, it is easy to see that $M_{\tilde{p}} P = 0$. Thus higher-order and least-squares bias correction approaches are equivalent. When $\tilde{p}$ and $p$ are not nested bases, the two methods will typically differ in variance and bias.

To compare their variance, we generally have

$$V[\hat{\mu}_2(x)|X] = V[\hat{\mu}_1(x)|X] + V[p(x)'(P'P)^{-1}P'M_{\tilde{p}}y|X] + 2Cov[\tilde{p}(x)'(\tilde{P}'\tilde{P})^{-1}\tilde{P}y, \ p(x)'(P'P)^{-1}P'M_{\tilde{p}}y|X].$$

When $\sigma^2(x) = \sigma^2$, the covariance term is 0, and thus $\hat{\mu}_2$ has variance no less than that of $\hat{\mu}_1$. The same conclusion is true for asymptotic variance as shown in the proof of Lemma SA-8.5.

Regarding their bias, let $\B_{\tilde{m},0}(x) = E[\hat{\mu}_1(x)|X] - \mu(x)$ denote the conditional bias of $\hat{\mu}_1(x)$. Then

$$E[\hat{\mu}_2(x)|X] - \mu(x) = E[\hat{\mu}_1(x)|X] - \mu(x) - p(x)'(P'P)^{-1}P'\B_{\tilde{m},0}$$

$$= \B_{\tilde{m},0}(x) - p(x)'(P'P)^{-1}P'\B_{\tilde{m},0}$$

where $\B_{\tilde{m},0} := (\B_{\tilde{m},0}(x_1), \ldots, \B_{\tilde{m},0}(x_n))'$. Clearly, the second term will asymptotically get close the projection of $\B_{\tilde{m},0}(x)$ onto the space spanned by $p$:

$$\mathcal{L}_p[\B_{\tilde{m},0}](x) := p(x)'(E[p(x)p(x)'])^{-1}E[p(x)\B_{\tilde{m},0}(x)]$$

where $\mathcal{L}_p[\cdot]$ denotes the projection operator. Therefore, when $p$ and $\tilde{p}$ are not nested bases, we typically have the bias of $\hat{\mu}_2$ no greater than that of $\hat{\mu}_1$ in terms of $\|\cdot\|_{F,L_2(\mathcal{X})}$ where for a real-valued function $g(\cdot)$ on $\mathcal{X}$, $\|g\|_{F,L_2(\mathcal{X})} = (\int_{\mathcal{X}} |g(x)|^2 dF(x))^{1/2}$.

According to the discussion above, higher-order and least-squares bias correction approaches do not dominate each other in general, and whether one is preferred to the other depends on the data generating process and the relation between two approximating spaces (or more precisely, the approximating power of $\tilde{p}$ to functions in the linear span of $p$). Then a natural question follows: is there an optimal weighting scheme when $p$ and $\tilde{p}$ are not nested? Again, assume $\varepsilon_i$'s
are homoskedastic for simplicity, and we take a weighted average of \( \hat{\mu}_1 \) and \( \hat{\mu}_2 \):

\[
\hat{\mu}_{w,\text{bc}} := w\hat{\mu}_2 + (1-w)\hat{\mu}_1
\]

where \( w \in [0,1] \). Then using a conclusion in the proof of Lemma SA-8.5, the change in the integrated asymptotic variance (weighted by the design density \( f(x) \)) is

\[
w^2\sigma^2 \int_X p(x)'Q_m^{-1}(Q_m - Q_{m,\tilde{m}}Q_m^{-1}Q_{m,\tilde{m}})Q_m^{-1}p(x)f(x)dx =: w^2 \tilde{\nu}.
\]

On the other hand, by the property of projection operator, the change in the integrated squared bias (weighted by the design density \( f(x) \)) is

\[
(w^2 - 2w) \int_X (\mathcal{L}_p[B_{m,0}](x))^2 f(x)dx =: (w^2 - 2w)\tilde{\beta}.
\]

It is easy to see the optimal weight is \( w^* = \tilde{\beta}/(\tilde{\beta} + \tilde{\nu}) \). Clearly, when variance is less important (e.g., \( \sigma^2 \) is small), \( w^* \) is close to 1 and \( \hat{\mu}_2 \) is preferred, whereas when bias is small, \( w^* \) is close to 0 and one may want to use \( \hat{\mu}_1 \).

Next, the comparison of plug-in bias correction with the other two is more complicated since \( \hat{\mu}_3 \) generally cannot be viewed as regression with additional covariates and the covariance between \( \hat{\mu}_0 \) and the estimated bias does not vanish. For piecewise polynomials, however, all three bias correction approaches are simply equivalent under certain conditions.

To see this, suppose \( p \) and \( \hat{p} \) are constructed on the same partitioning scheme \( \Delta \), but the order of basis increases from \( m \) to \( m + 1 \). \( \hat{\mu}_0 \) and \( \hat{\mu}_1 \) are linked by (SA-1) since \( \hat{\mu}_0 \) can be viewed as a restricted regression estimate compared with \( \hat{\mu}_1 \). Specifically, one can construct a polynomial series of order \( m + 1 \) (with degree no greater than \( m \)) within each cell \( \delta \in \Delta \), and then implement a local regression restricting the coefficients of the polynomial terms of degree \( m \) to be 0. Then the restriction matrix in this case takes the following form:

\[
R = [0 \ I_\vartheta]
\]

where \( \vartheta \) denotes the number of polynomial terms of degree \( m \) and thus \( R \) is a \( \vartheta \times \tilde{K} \) matrix. Plug
it in (SA-1), and then use the formula for matrix inverse in block form to obtain \((\tilde{P}'\tilde{P})^{-1}\). It is easy to see that the second term on the RHS of (SA-1) \((p(x)'\hat{M}_j\hat{\beta}_{ur})\) is exactly the same as the leading bias derived in [12] (see their proof of Theorem 3) with the \(m\)th derivative estimated by piecewise polynomials of order \(m + 1\). As explained in the proof of Lemma A.3, the leading approximation error can be alternatively expressed in terms of Legendre polynomials. Asymptotically, the two expressions are equivalent since the “locally” orthogonalized polynomials of degree \(m\) will converge to the \(m\)th Legendre polynomials given by Lemma A.3.

For splines or wavelets, we do not generally have the above equivalence since bases of different orders do not generate nested spaces, and the relative performance of the three approaches depends on the relation between these approximating spaces.

**SA-12 Implementation Details**

In this section, we briefly discuss implementation details about choosing the IMSE-optimal tuning parameters. We restrict our attention to rectangular partitions with the same number of knots used in every dimension. Thus the tuning parameter reduces to a scalar \(\kappa\) which denotes the number of subintervals used in every dimension. We offer two approaches: rule-of-thumb (ROT) and direct plug-in (DPI).

**SA-12.1 Rule-of-Thumb Choice**

The rule-of-thumb choice is based on the special case considered in Corollary 1 in the main paper. Specifically, we assume \(q = 0\) and knots are evenly spaced. The implementation steps are summarized as follows.

- **Preliminary regression.** Implement a preliminary regression using a global polynomial of degree \((m + 4)\), and denote this estimate of \(\mu(\cdot)\) by \(\hat{\mu}_{pre}(\cdot)\).
- **Bias constant:** Let the weighting function \(w(x)\) be the density function of \(x_i\). Use the preliminary regression coefficients to obtain an estimate of the \(m\)th derivatives of \(\mu(\cdot)\), i.e.,
\( \partial^u \mu(\cdot) = \partial^u \mu_{\text{pre}}(\cdot) \), for each \( u \in \Lambda_m \). Then an estimate of the bias constant is

\[
\hat{B}_{u_1, u_2, 0} = \eta_{u_1, u_2, 0} \times \frac{1}{n} \sum_{i=1}^{n} \hat{\partial}^u \mu(x_i) \hat{\partial}^u \mu(x_i).
\]

**Variance constant.** Implement another regression of \( y_i^2 \) on \( x_i \), using global polynomials of degree \( (m + 4) \), leading to an estimate \( \hat{\mathbb{E}}[y_i^2|x_i = x] \). Combining it with \( \hat{\mu}_{\text{pre}}(\cdot) \), we obtain an estimate of the conditional variance function, denoted by \( \hat{\sigma}^2(\cdot) \), since

\[
\sigma^2(x) = \mathbb{E}[y_i^2|x_i = x] - (\mathbb{E}[y_i|x_i = x])^2.
\]

Then an estimate of the variance constant is

\[
\hat{V}_0 = \begin{cases} 
\frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^2(x_i) & \text{for splines and wavelets,} \\
\left( \frac{d+m-1}{m-1} \right) \times \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^2(x_i) & \text{for piecewise polynomial.}
\end{cases}
\]

**Rule-of-thumb \( \hat{\kappa}_{\text{ROT}} \).** Using the above results, a simple rule-of-thumb choice of \( \kappa \) is

\[
\hat{\kappa}_{\text{ROT}} = \left[ \left( \frac{2(m - \lfloor q \rfloor)}{d + 2\lfloor q \rfloor} \right) \sum_{u_1, u_2 \in \Lambda_m} \hat{\mathbb{R}}_{u_1, u_2, 0} \right]^{\frac{1}{d+m+\beta}}.
\]

Clearly, this choice of \( \kappa \) is derived based on many strong assumptions, but it still has the correct rate \(( \approx n^{\frac{1}{2m+\beta}} \) in other cases.

**SA-12.2 Direct Plug-in Choice**

Assume that the weighting function \( w(x) \) is equal to the density function of \( x_i \). We propose a direct-plug-in (DPI) procedure summarized in the following steps.

- **Preliminary choice of \( \kappa \):** Implement the rule-of-thumb procedure to obtain \( \hat{\kappa}_{\text{ROT}} \).

- **Preliminary regression.** Given the user-specified basis (splines, wavelets, or piecewise polynomials), knot placement scheme (“uniform” or “quantile”) and rule-of-thumb choice \( \hat{\kappa}_{\text{ROT}} \), implement a series regression of order \((m + 1)\) to obtain derivative estimates for every \( u \in \Lambda_m \). Denote this preliminary estimate by \( \hat{\partial}^u \mu_{\text{pre}}(\cdot) \).

- **Bias constant.** Construct an estimate \( \hat{\mathbb{L}}_{m,q}(\cdot) \) of the leading error \( \mathbb{L}_{m,q}(\cdot) \) simply by replacing \( \partial^u \mu(\cdot) \) by \( \hat{\partial}^u \mu_{\text{pre}}(\cdot) \). \( \hat{\mathbb{L}}_{m,0}(\cdot) \) can be obtained similarly. Then we use the pre-asymptotic
version of conditional bias to estimate the bias constant:

\[
\hat{B}_{\kappa, q} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{B}_{m, q}(x_i) - \hat{\gamma}_{q, 0}(x) \hat{\Sigma}_0 \hat{\gamma}_{q, 0}(x) \right)^2.
\]

- **Variance constant.** Implement a series regression of order \(m\) using \(\hat{\kappa}_{\text{rot}}\), and then use the pre-asymptotic version of conditional variance to obtain an estimate of variance constant. Specifically, we have

\[
\hat{V}_{\kappa, q} = \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_{q, 0}(x) \hat{\Sigma}_0 \hat{\gamma}_{q, 0}(x), \quad \hat{\Sigma}_0 = \mathbb{E}_n \left[ \Pi_0(x_i) \Pi_0(x_i)' \hat{\epsilon}_{i, 0}^2 \right]
\]

where \(\hat{\epsilon}_{i, 0}\)'s are regression residuals. Different weighting schemes for residuals may be used, leading to various “heteroskedasticity-consistent” variance estimates.

- **Direct plug-in \(\hat{\kappa}\).** Collecting all these results, a direct plug-in choice of \(\kappa\) is

\[
\hat{\kappa}_{\text{DPI}} = \left[ \frac{2(m - [q]) \hat{\kappa}_{\text{ROT}}^{2(m-[q])} \hat{B}_{\kappa, q}}{(d + 2[q]) \hat{\kappa}_{\text{ROT}}^{-(d+2[q])} \hat{V}_{\kappa, q}} \right]^{\frac{1}{2m+d}} n^{\frac{1}{2m+d}}.
\]
SA-13  Simulations

In this section, we present detailed simulation results. We consider the following four regression functions:

Model 1: $\mu(x) = \sin\left(\frac{\pi}{2}(2x - 1)\right)/(1 + 2(2x - 1)^2(\text{sign}(2x - 1) + 1))$

Model 2: $\mu(x) = \sin\left(\frac{3\pi}{2}(2x - 1)\right)/(1 + 18(2x - 1)^2(\text{sign}(2x - 1) + 1))$

Model 3: $\mu(x) = 2x - 1 + 5\phi(20x - 10)$

Model 4: $\mu(x) = \sin(5x_1)\sin(10x_2)$

where $\text{sign}(x) = -1, 0, 1$ if $x < 0, x = 0, x > 0$ respectively, and $\phi(\cdot)$ is the standard normal density function. Models 1-3 are one-dimensional functions used in many previous studies, e.g., [36]. Model 4 is a two-dimensional regression function also used in, for example, [12].

Given each regression function, we generate $(y_i, x_i)_{i=1,\ldots,n}$ by $y_i = \mu(x_i) + \varepsilon_i$ where $x_i \sim i.i.d. U[0,1]^d$ and $\varepsilon_i \sim i.i.d. N(0,1)$. We generate 5000 simulated datasets with $n = 1000$. For each model, we use linear splines ($m = 2$) to form the classical point estimates. For robust inference, we use quadratic splines ($\tilde{m} = 3$) to implement bias correction. Both evenly-spaced and quantile-spaced knot placements are considered, and for simplicity point estimators and bias correction employ the same knot placements ($\Delta = \tilde{\Delta}$).

For each model, we present three sets of simulation evidence. First, for three fixed points, we calculate the (simulated) root mean squared error (RMSE), coverage rate (CR) and confidence interval length (IL). The nominal coverage is set to be 95%. Second, to evaluate the performance of our rule-of-thumb (ROT) and direct plug-in (DPI) knot selection procedures, we show some basic summary statistics (mean, median, standard deviation, etc.) of the selected number of knots. Third, for uniform confidence bands, we mimic [36] to calculate three measures: the proportion of values covered with probability at least 95% (CP), average coverage errors (ACE), and average width of confidence band (AW). In addition, uniform coverage rate (UCR), as a more stringent criterion, is also presented. The quantile estimation based on the plug-in and bootstrap methods uses 1000 random draws conditional on the data for each simulated dataset.
SA-14  Proofs: Supplemental Appendix

SA-14.1  Proof of Lemma SA-8.1

Proof. We first prove the boundedness of eigenvalues of $Q_m$. We use the fact that $\lambda_{\text{max}}(Q_m) = \max_{a' = 1} a'Q_m a$ and $\lambda_{\text{min}}(Q_m) = \min_{a' = 1} a'Q_m a$. By definition of $Q_m$,

$$a'Q_m a = a'\mathbb{E}[p(x_i)p(x_j)]a = \int_X \left( \sum_{k=1}^K a_k p_k(x) \right)^2 f(x) \, dx =: \|s(x)\|^2_{F,L_2(X)}$$

where $s(x) = \sum_{k=1}^K a_k p_k(x)$. Then it follows from Assumption SA-1 that

$$\|s(x)\|^2_{L_2(X)} \lesssim \|s(x)\|^2_{F,L_2(X)} \lesssim \|s(x)\|^2_{L_2(X)}.$$ 

By Assumption SA-3(a), the number of basis functions in $p(\cdot)$ which are active on a generic cell $\delta_l$, $l = 1, \ldots, \bar{r}$, is bounded by a constant. Denoted them by $(\bar{p}_{l,1}, \ldots, \bar{p}_{l,M_l})'$, where $M_l$ may vary across $l$. It follows from Assumption SA-3(c) that

$$s(x)^2 = \left( \sum_{k=1}^{M_l} a_k \bar{p}_{l,k}(x) \right)^2 \lesssim \sum_{k=1}^{M_l} a_k^2$$

for all $x \in \delta_l$.

Taking integral and summing over all $\delta_l$, we have $\|s(x)\|^2_{L_2(X)} \lesssim h^d$, and the upper bound on $\lambda_{\text{max}}(Q_m)$ follows. On the other hand, it follows from Assumption SA-3(b) that $\|s(x)\|^2_{L_2(H_k)} \gtrsim a_k^2 h^d$. Then taking sum over all $H_k$, we have $\|s(x)\|^2_{L_2(X)} \gtrsim h^d$, and the lower bound on $\lambda_{\text{min}}(Q_m)$ follows.

To derive the convergence rate of $\hat{Q}_m$, let $\alpha_{k,l} = \frac{1}{n} \sum_{i=1}^n \alpha_{k,l}(i)$ be the $(k,l)$th element of $(\hat{Q}_m - Q_m)$ where $\alpha_{k,l}(i) := p_k(x_i)p_l(x_i) - \mathbb{E}[p_k(x_i)p_l(x_i)]$. It follows from Assumption SA-1 and SA-3 that $\alpha_{k,l}$ is the sum of $n$ independent random variables with zero means, $|\alpha_{k,l}(i)| \lesssim 1$ uniformly over $i, k$ and $l$, and thus $\mathbb{E}[|\alpha_{k,l}(i)|] \lesssim h^d/n$. By Bernstein’s inequality, for every $\vartheta > 0$,

$$\mathbb{P}(|\alpha_{k,l}| > \vartheta) \leq 2 \exp \left( -\frac{\vartheta^2/2}{C_1 h^d/n + C_2 \vartheta/(3n)} \right).$$

By Assumption SA-3, $(\hat{Q}_m - Q_m)$ only has a finite number of nonzeros on any row or column, and
thus for every $\vartheta > 0$,
\[
P(\max_{k,l} |\alpha_{k,l}| > \vartheta) \leq 2CK \exp\left(-\frac{\vartheta^2/2}{C_1h^d/n + C_2\vartheta/(3n)}\right).
\]

Then it follows that $\max_{k,l} |\alpha_{k,l}| \lesssim P h^d \sqrt{\log n/(nh^d)}$, which suffices to show that
\[
\|\hat{Q}_m - Q_m\|_{\infty} \lesssim_P h^d \sqrt{\log n/(nh^d)} \quad \text{and} \quad \|\hat{Q}_m - Q_m\|_1 \lesssim_P h^d \sqrt{\log n/(nh^d)}.
\]

By the relation between induced operator norms,
\[
\|\hat{Q}_m - Q_m\| \lesssim_P h^d \sqrt{\log n/(nh^d)}.
\]

Hence when $\log n/(nh^d) = o(1)$, $\|\hat{Q}_m - Q_m\| = o_P(h^d)$. Notice that for any vector $a \in \mathbb{R}^K$ such that $a^\prime a = 1$,
\[
\|\hat{Q}_m - Q_m\| \geq |a^\prime(\hat{Q}_m - Q_m)a| = |a^\prime\hat{Q}_m a - a^\prime Q_m a|.
\]

Since $h^d \lesssim a^\prime Q_m a \lesssim h^d$, this suffices to show that $\|\hat{Q}_m\| \lesssim_P h^d$, and with probability going to 1, $\lambda_{\min}(\hat{Q}_m) \gtrsim h^d$.

Next, we show $\|\hat{Q}_m^{-1}\|_{\infty} \lesssim_P h^{-d}$. In the univariate case, it is easy to see that by Assumption SA-3, $\hat{Q}_m$ is a banded matrix with a finite bandwidth (independent of $n$). In the multidimensional case, $\hat{Q}_m$ is a block banded matrix ($A_{kl}$) with $A_{kl} = 0$ if $|k - l| > L$ for some integer $L > 0$ where each block $A_{kl}$ is banded or takes further block banded structure if we appropriately arrange the ordering of basis functions. Specifically, we “rectangularize” the partition $\Delta$. By Assumption SA-2, $\Delta$ is quasi-uniform and there exists a universal measure of mesh size $h$. Construct an initial rectangular partition covering $\mathcal{X}$, which is formed as tensor products of intervals of length $h$. A generic cell in this partition is indexed by a $d$ tuple $d$-tuple $(l_1,\ldots,l_d)$. Then for each cell in this partition, take its intersection with $\mathcal{X}$ and exclude all cells outside of $\mathcal{X}$. Thus we construct a “trimmed” rectangular partition $\Delta^{\text{rec}} = \{\delta^{\text{rec}}_{l_1,\ldots,l_d}\}$. Clearly, each element $\delta \in \Delta$ is covered by a finite number of cells in $\Delta^{\text{rec}}$. On the other hand, each $\delta^{\text{rec}} \in \Delta^{\text{rec}}$ is also overlapped with a finite number of cells in $\Delta$ (if not, cells in $\Delta$ overlapping with $\delta^{\text{rec}}$ cannot be covered by a ball of radius $2h$). Arrange these cells by first increasing $l_d$ with other $l_\ell$’s fixed at the lowest values and then
increasing \( l_\ell \)'s sequentially. Then arrange the basis functions in \( p(\cdot) \) according to their supports. Specifically, start with basis functions which are active on the first cell, and then arrange those functions which are active on the second and have not been included yet. Continue this procedure until the functions active on the last cell have been included. According to this particular ordering, the Gram of this basis has the same banded structure as that of tensor-product local basis on rectangular partitions. The nested banded structure involves at most \( d \) layers. The bandwidth at each layer may be different, but is bounded by a universal constant \( \bar{L} \) by Assumption SA-3.

In the one-dimensional case, it directly follows from [26, Theorem 2.2] that \( \| \hat{Q}_m^{-1} \|_\infty \lesssim h^{-1} \). In other cases when \( \hat{Q}_m \) involves multilayer banded structures, we need to slightly modify their original proof. We only prove for the case when \( \hat{Q}_m \) is block banded with banded blocks (two layers of banded structures). The general case follows similarly.

For some universal constants \( C_1 \) and \( C_2 \), with probability approaching one, \( \lambda_{\min}(\hat{Q}_m) \geq C_1 h^d \) and \( \lambda_{\max}(\hat{Q}_m) \leq C_2 h^d \). Hence for any vector \( a \in \mathbb{R}^K \) such that \( a^t a = 1 \), there are some constants \( C_3, C_4 \) and \( C_5 \) such that with probability approaching one,

\[
0 < C_3 \leq \frac{a^t \hat{Q}_m a}{C_5 h^d} \leq C_4 < 1.
\]

Hence \( \Psi := I_K - \hat{Q}_m/(C_5 h^d) \) is a block banded matrix with banded blocks satisfying \( \| \Psi \| < 1 \) with probability going to 1. Therefore, we can write \( C_5 h^d \hat{Q}_m^{-1} = \sum_{l=1}^\infty \Psi^l \). The \((s,t)\)th entry of \( \hat{Q}_m^{-1} \), denoted by \( \alpha_{s,t} \), is an element of \( \sum_{l=1}^\infty \Psi^l \), and hence

\[
|\alpha_{s,t}| \leq \sum_{l=\chi(s,t,L)}^\infty \| \Psi^l \| = \frac{\| \Psi \| \chi(s,t,L)}{1 - \| \Psi \|}
\]

where \( \chi(s,t,\bar{L}) \) is a number depending on the row index \( s \), column index \( t \) and the upper bound on bandwidths \( \bar{L} \). We further denote the block row index and column index of \( \alpha_{s,t} \) as \( r_s \) and \( r_t \), and the row index and column index within the block containing it as \( \iota_s \) and \( \iota_t \). \( |r_s - r_t| \) and \( |\iota_s - \iota_t| \) measure how far away \( \alpha_{s,t} \) is from the diagonals of the entire matrix and the block it belongs to.

As in the one-dimensional case, the first few products of \( \Psi \) do not contribute to off-diagonal blocks of the inverse matrix. As \( |r_s - r_t| \) increases, \( \chi(s,t,\bar{L}) \) also gets larger. Meanwhile, since each block of \( \hat{Q}_m \) is also banded, \( \chi(s,t,\bar{L}) \) also gets larger when \( |\iota_s - \iota_t| \) increases. By the same argument in
\[|\alpha_{s,t}| \leq (1 - \|\Psi\|)^{-1}\|\Psi\| |r_s - r_t|/C_6 + |\iota_s - \iota_t|/C_6 \]

where \(C_6\) is some constant depending on \(\bar{L}\). Thus \(\alpha_{s,t}\) exponentially decays when \(|r_s - r_t|\) or \(|\iota_s - \iota_t|\) becomes large. Then it directly follows from the convergence of geometric series that \(\|\hat{Q}_m^{-1}\|_\infty \lesssim \bar{h}^{-d}\).

It follows from exactly the same argument that \(\|Q_m^{-1}\|_\infty \lesssim \bar{h}^{-d}\), and the proof is complete. \(\square\)

### SA-14.2 Proof of Lemma SA-8.2

**Proof.** It directly follows from Lemma SA-8.1 that

\[
\|\hat{Q}_m^{-1} - Q_m^{-1}\|_\infty \leq \|\hat{Q}_m^{-1}\|_\infty \|Q_m^{-1}\|_\infty \lesssim \bar{h}^{-d} \sqrt{\log n/(nh^d)}
\]

and

\[
\|\hat{Q}_m^{-1} - Q_m^{-1}\| \leq \|\hat{Q}_m^{-1}\| \|Q_m^{-1}\| \lesssim \bar{h}^{-d} \sqrt{\log n/(nh^d)}.
\]

Since \(\sigma^2(x) \lesssim 1\) uniformly over \(x \in \mathcal{X}\), the bound on \(\|\mathbb{E}_n[p(x_i) \bar{p}(x_i) \sigma^2(x_i)] - \mathbb{E}[p(x_i) \bar{p}(x_i)]\|\)

immediately follows from the same argument as that for \(\|\hat{Q}_m - Q_m\|\) in Lemma SA-8.1. \(\square\)

### SA-14.3 Proof of Lemma SA-8.3

**Proof.** We first prove the uniform bound on the conditional bias. By Assumption SA-4 we can find \(s^* \in S_{\Delta,m}\) such that \(\sup_{x \in \mathcal{X}} |\partial^q \mu(x) - \partial^q s^*(x)| \lesssim h^{-m-|q|}\). Since

\[
\mathbb{E}[\partial^q \mu_0(x) | \mathcal{X}] = \partial^q p(x) \bar{Q}_m^{-1} \mathbb{E}_n[p(x_i) \mu(x_i)] - \partial^q \mu(x)
\]

\[
= \partial^q p(x) \bar{Q}_m^{-1} \mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i))] + \partial^q s^*(x) - \partial^q \mu(x),
\]

we only need to show that the first term in the second line is properly bounded uniformly over \(x \in \mathcal{X}\). It follows from Lemma SA-8.2 and Assumption SA-3 that

\[
\sup_{x \in \mathcal{X}} \left| \partial^q p(x) \bar{Q}_m^{-1} \mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i))] \right|
\]

\[
\lesssim h^{-|q|} \|\bar{Q}_m^{-1}\|_\infty \|\mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i))]\|_\infty
\]

\[
\lesssim \bar{P} h^{-|q|-d} \|\mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i))]\|_\infty
\]

72
By Assumption SA-3 and SA-4, \( \max_{1 \leq k \leq K} \| \mathbb{E}[p_k(x_i)(\mu(x_i) - s^*(x_i))] \| \lesssim h^{m+d} \), \( \| \mathbb{E}[p_k(x)(\mu(x) - s^*(x))] \| \lesssim h^m \). By Bernstein’s inequality, for any \( \theta > 0 \),

\[
P \left( \max_{1 \leq k \leq K} \left| \mathbb{E}_n[p_k(x_i)(\mu(x_i) - s^*(x_i))] - \mathbb{E}[p_k(x_i)(\mu(x_i) - s^*(x_i))] \right| > \theta \right) 
\leq \sum_{k=1}^{K} P \left( \left| \mathbb{E}_n[p_k(x_i)(\mu(x_i) - s^*(x_i))] - \mathbb{E}[p_k(x_i)(\mu(x_i) - s^*(x_i))] \right| > \theta \right) 
\leq 2K \exp \left( \frac{-\theta^2/2}{C_1h^m \theta/(3n) + C_2h^{2m+d}/n} \right).
\]

Therefore,

\[
\max_{1 \leq k \leq K} \left| \mathbb{E}_n[p_k(x_i)(\mu(x_i) - s^*(x_i))] - \mathbb{E}[p_k(x_i)(\mu(x_i) - s^*(x_i))] \right| \lesssim h^{m+d} \sqrt{\frac{\log n}{nh^d}},
\]

which suffices to prove the desired uniform bound by the assumption that \( \frac{\log n}{nh^d} = o(1) \).

Next, we prove the leading bias expansion. By Assumption SA-4, we can find \( s^* \in \mathcal{S}_{\Delta,m} \) such that

\[
\sup_{x \in \mathcal{X}} | \partial^q \mu(x) - \partial^q s^*(x) + \mathcal{B}_{m,q}(x) | \lesssim h^{m+e-[q]}.
\]

Then conditional bias can be expanded as follows:

\[
\mathbb{E}[\hat{\partial}^q \mu_0(x)|X] - \partial^q \mu(x) = \partial^q p(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[p(x_i)\mu(x_i)] - \partial^q \mu(x) \\
= \partial^q p(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[p(x_i)\mu(x_i)] - \partial^q s^*(x) + \mathcal{B}_{m,q}(x) + O(h^{m+e-[q]}) \\
= \mathcal{B}_{m,q}(x) + \partial^q p(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[p(x_i)\mu(x_i) - s^*(x_i)] + O(h^{m+e-[q]}).
\]

The second term in the last line can be further written as

\[
\partial^q p(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i))]
= - \partial^q p(x)' \tilde{Q}_m^{-1} \mathbb{E}_n [p(x_i) \mathcal{B}_{m,0}(x_i)] + \partial^q p(x)' \tilde{Q}_m^{-1} \mathbb{E}_n \left[ p(x_i)(\mu(x_i) - s^*(x_i)) + \mathcal{B}_{m,0}(x_i) \right].
\]

It follows from the same argument used to bound \( \| \mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i))] \|_{\infty} \) that

\[
\max_{1 \leq k \leq K} \left| p_k(x_i)(\mu(x_i) - s^*(x_i)) + \mathcal{B}_{m,0}(x_i) \right| \lesssim h^{m+e+d}.
\]
Then by Assumption SA-3 and Lemma SA-8.2,

\[
\|\partial^q p(x)\hat{Q}_m^{-1}\mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i) + \Phi_{m,0}(x_i))]\|_{L_\infty(X)} \lesssim \epsilon^m + \epsilon^{-|q|},
\]

which suffices to show the desired bias expansion.

Suppose that Equation (SA-6) holds. Then by Lemma SA-8.2 and Assumption SA-3, the second term in the bias expansion is \(o(h^{m-|q|})\). Thus the leading bias is further reduced to \(\Phi_{m,q}(x)\). The proof of Lemma SA-8.3 is complete. \(\square\)

### SA-14.4 Proof of Lemma SA-8.4

**Proof.** For \(j = 0, 1\), the results immediately follow from Assumption SA-3, SA-5, Lemma SA-8.1 and SA-8.2. For \(j = 2\),

\[
\sup_{x \in \mathcal{X}} \|\gamma_{q,2}(x)^t\|_\infty \leq \sup_{x \in \mathcal{X}} \|\gamma_{q,0}(x)^t\|_\infty + \sup_{x \in \mathcal{X}} \|\gamma_{q,0}(x)^tQ_m\hat{Q}_m^{-1}\|_\infty + \sup_{x \in \mathcal{X}} \|\gamma_{q,1}(x)^t\|_\infty
\]

\[
\lesssim h^{-d-|q|} + h^{-d-|q|}\|Q_m\hat{Q}_m^{-1}\|_\infty\|Q_m^{-1}\|_\infty.
\]

By Assumptions SA-5, both \(p(\cdot)\) and \(\bar{p}(\cdot)\) are local bases, and hence \(Q_m\hat{Q}_m^{-1}\) has a finite number of nonzero elements on any row or column, and all element in \(Q_m\hat{Q}_m^{-1}\) are bounded by \(Ch^d\) for some universal constant \(C\). Thus \(\|Q_m\hat{Q}_m^{-1}\|_\infty \lesssim h^d\), \(\|Q_m\hat{Q}_m^{-1}\|_1 \lesssim h^d\) and \(\|Q_m\hat{Q}_m^{-1}\|_\infty \lesssim h^d\). Therefore we conclude \(\|\gamma_{q,2}(x)^t\|_\infty \lesssim h^{-d-|q|}\). It follows similarly that \(\|\gamma_{q,2}(x)^t\|_\infty \lesssim h^{-d-|q|}\). The lower bound on the \(L_2\)-norm follows from Assumption SA-3.

Moreover, note that

\[
\|\tilde{\gamma}_{q,2}(x)^t - \gamma_{q,2}(x)^t\|_\infty
\]

\[
\leq \|\tilde{\gamma}_{q,0}(x)^t - \gamma_{q,0}(x)^t\|_\infty + \|\tilde{\gamma}_{q,0}(x)^t\hat{Q}_m\hat{Q}_m^{-1} - \gamma_{q,0}Q_m\hat{Q}_m^{-1}\|_\infty
\]

\[
+ \|\tilde{\gamma}_{q,1}(x)^t - \gamma_{q,1}(x)^t\|_\infty
\]

\[
\leq \|(\tilde{\gamma}_{q,0}(x)^t - \gamma_{q,0}(x)^t)\hat{Q}_m\hat{Q}_m^{-1}\|_\infty + \|\gamma_{q,0}(x)^t(\hat{Q}_m - Q_m)\hat{Q}_m^{-1}\|_\infty
\]

\[
+ \|\gamma_{q,0}(x)^tQ_m\hat{Q}_m^{-1}(Q_m^{-1} - \hat{Q}_m^{-1})\|_\infty + h^{-d-|q|}\sqrt{\log n/(nh^d)}
\]

where the last line uses the results for \(j = 0, 1\). Using the sparsity of \((\hat{Q}_m - Q_m)\) and the same
argument for Lemma \textbf{SA-8.1}, \[ \| \hat{Q}_{m, \hat{m}} - Q_{m, \hat{m}} \|_{\infty} \lesssim h^d \sqrt{\log n / (nh^d)}. \] Then the desired uniform bound follows from Lemma \textbf{SA-8.1} and Assumption \textbf{SA-3}. The $L_2$-bound follows similarly.

For $j = 3$, notice that
\[
\| \gamma_{q,3}(x) \|_{\infty} \leq \| \gamma_{q,0}(x) \|_{\infty} + \sum_{u \in \Lambda_m} \| \gamma_{u,1}(x)^j h_{x}^{m-\|q\|} B_{u,q}(x) \|_{\infty} + \sum_{u \in \Lambda_m} \| \gamma_{q,0}(x)^j E[p(x_i)h_{x}^{m} B_{u,0}(x_i) \partial^u \tilde{\rho}(x_i)^j] Q^{-1}_{\hat{m}} \|_{\infty}.
\]

By Assumption \textbf{SA-4} and \textbf{SA-5}, both $p(\cdot)$ and $\tilde{\rho}(\cdot)$ are locally supported and all elements in $p(x_i)h_{x}^{m} B_{u,0}(x_i) \partial^u \tilde{\rho}(x_i)^j$ is bounded by a universal constant. Then it follows from the argument given in Lemma \textbf{SA-8.1} that
\[
\| E[p(x_i)h_{x}^{m} B_{u,0}(x_i) \partial^u \tilde{\rho}(x_i)^j] \|_{\infty} \lesssim h^d \quad \text{and}
\]
\[
\| E[p(x_i)h_{x}^{m} B_{u,0}(x_i) \partial^u \tilde{\rho}(x_i)^j - E[p(x_i)h_{x}^{m} B_{u,0}(x_i) \partial^u \tilde{\rho}(x_i)^j]] \|_{\infty} \lesssim h^d \sqrt{\log n / nh^d}.
\]

Then the desired results follow from Assumption \textbf{SA-3}, Lemma \textbf{SA-8.1} and \textbf{SA-8.2}. \hfill \Box

\textbf{SA-14.5 Proof of Lemma \textbf{SA-8.5}}

\textit{Proof.} For $j = 0, 1$, the results directly follow from Assumption \textbf{SA-1}, \textbf{SA-3} and Lemma \textbf{SA-8.1}. The proof for $j = 2, 3$ is divided into several steps.

**Step 1:** We first establish the upper bounds on $\Omega_j(x)$. By Assumption \textbf{SA-1},
\[
\Sigma_j = E[\varepsilon_j^2 \Pi_j(x_i)] \Pi_j(x_i)^j \lesssim E[\Pi_j(x_i)] \Pi_j(x_i)^j.
\]

To bound $\| \Sigma_j \|$, it suffices to give an upper bound on $E[\Pi_j(x_i) \Pi_j(x_i)^j]$. It follows from Assumption \textbf{SA-5} and the same proof strategy for Lemma \textbf{SA-8.1} that $\| E[\Pi_j(x_i) \Pi_j(x_i)] \| \lesssim h^d$. By Lemma \textbf{SA-8.4}, $\sup_{x \in \mathcal{X}} \| \gamma_{q,j}(x) \| \lesssim h^{-d-\|q\|}$. Thus $\sup_{x \in \mathcal{X}} \Omega_j(x) \lesssim h^{-d-2\|q\|}$.

**Step 2:** Next, we show the lower bound on $\Omega_2$. Since $\sigma^2(x) \gtrsim 1$ uniformly over $x \in \mathcal{X}$, we have
\[
\Omega_2(x) \gtrsim \gamma_{q,2}(x)^j E[\Pi_2(x_i) \Pi_2(x_i)^j] \gamma_{q,2}(x).
\]
Expanding the expression on the RHS of this inequality, we have a trivial lower bound:

$$\gamma_{q,2}(x)\mathbb{E}[\Pi_2(x_i)\Pi_2(x_i)]\gamma_{q,2}(x)$$

$$= \partial^q p(x)'Q_m^{-1}\partial^q p(x) + (\gamma_{q,0}(x)'Q_m\tilde{m} - \partial^q\tilde{p}(x)')Q_m^{-1}(\gamma_{q,0}(x)'Q_m\tilde{m} - \partial^q\tilde{p}(x)')'$$

$$- 2\partial^q p(x)'Q_m^{-1}Q_m\tilde{m}Q_m^{-1}(\gamma_{q,0}(x)'Q_m\tilde{m} - \partial^q\tilde{p}(x)')'$$

$$= \partial^q \tilde{p}(x)'Q_m^{-1}\partial^q \tilde{p}(x) + \left[\partial^q p(x)'Q_m^{-1}(Q_m - Q_m\tilde{m}Q_m^{-1}Q_m\tilde{m})Q_m^{-1}\partial^q p(x)\right].$$

By properties of projection operator, \((Q_m - Q_m\tilde{m}Q_m^{-1}Q_m\tilde{m})\) is positive semidefinite. By Assumption SA-5 and Lemma SA-8.1, \(\partial^q \tilde{p}(x)'Q_m^{-1}\partial^q \tilde{p}(x) \geq h^{-d-2[q]}\), and thus the desired lower bound is obtained.

**Step 3.1:** Now let us bound \(\Omega_3(x)\) from below. Suppose condition (i) in Assumption SA-5 holds. Then there exists a linear map \(\Upsilon\) such that \(\Pi_3(\cdot) = \Upsilon \tilde{p}(\cdot)\). By Lemma SA-8.1 and Assumption SA-1,

$$\Omega_3(x) \geq \gamma_{q,3}(x)'\Upsilon \mathbb{E}[\tilde{p}(x_i)\tilde{p}(x_i)]'\Upsilon'\gamma_{q,3}(x) \geq h^d\gamma_{q,3}(x)'\Upsilon \Upsilon'\gamma_{q,3}(x).$$

Define \(v(x)' := \gamma_{q,3}(x)'\Upsilon\). Notice that for any function \(s(x) \in \text{span}(p(\cdot))\), there exists some \(c \in \mathbb{R}^\tilde{k}\) such that \(s(x) = \tilde{p}(x)'c\). It follows that \(v(x)'\mathbb{E}[\tilde{p}(x_i)s(x_i)] = \partial^q s(x)\). Then we have

$$\|v(x)\| \geq \frac{\|v(x)'\mathbb{E}[\tilde{p}(x_i)s(x_i)]\|}{\|\mathbb{E}[\tilde{p}(x_i)s(x_i)]\|} = \frac{\|\partial^q s(x)\|}{\|\mathbb{E}[\tilde{p}(x_i)s(x_i)]\|}.$$

Since the choice of \(s(x)\) is arbitrary within the span of \(p\), by Assumption SA-3(c) we can take a function in \(p(x)\) to be \(s(x)\) such that \(\|\partial^q s(x)\| \geq Ch^{-[q]}\) where \(C\) is a constant independent of \(x\) and \(n\). Also, \(\|\mathbb{E}[\tilde{p}(x_i)s(x_i)]\| \lesssim h^d\). Hence \(h^{-[q]-d} \lesssim \inf_{x \in X} \|v(x)\|\). The desired bound follows.

**Step 3.2:** Now suppose condition (ii) in Assumption SA-5 is satisfied. Again, since \(\sigma^2(x) \geq 1\) uniformly over \(x \in X\),

$$\Omega_3(x) \geq \gamma_{q,3}(x)'\mathbb{E}[\Pi_3(x_i)\Pi_3(x_i)]'\gamma_{q,3}(x).$$

Define \(\mathcal{X}_h(x, x_1) := \gamma_{q,3}(x)'\Pi_3(x_1)\). Then it suffices to bound \(\mathbb{E}_{x_1}\mathcal{X}_h(x, x_1)^2\) where \(\mathbb{E}_{x_1}\) denotes the expectation with respect to the distribution of \(x_1\). We write

$$\langle g_1, g_2 \rangle_U := \int_U g_1(x_1)g_2(x_1)f(x_1)dx_1$$
for the inner product of \( g_1(\cdot) \) and \( g_2(\cdot) \) with respect to the probability measure of \( x_1 \) on \( U \subset X \).

Clearly, \( \langle g_1, g_2 \rangle_X = E[g_1(x_1)g_2(x_1)] \). By Cauchy-Schwartz inequality, for \( g \in L^2(U) \),

\[
\langle \mathcal{H}(x, x_1), g(x_1) \rangle_U^2 \leq \langle \mathcal{H}(x, x_1), \mathcal{H}(x, x_1) \rangle_U \cdot \langle g(x_1), g(x_1) \rangle_U.
\]

Given an evaluating point \( x \in X \), choose a polynomial

\[
\varphi_h(x_1; x) := \frac{(x_1 - x)^q}{h^q}.
\]

Clearly, \( \partial^q \varphi_h(x_1; x) = h^{-q} \). By Assumption SA-5, the operator \( \langle \mathcal{H}(x, x_1), \cdot \rangle_X \) reproduces the \( q \)th derivative of \( \varphi_h(x_1; x) \) at \( x \), i.e. \( \langle \mathcal{H}(x, x_1), \varphi_h(x_1; x) \rangle_X = h^{-q} \).

In addition, we rectangularize \( \Delta \) as described in the proof of Lemma SA-8.1. Then for each \( x_1 \in X \), we can choose a rectangular region that contains \( x \) and consists of a fixed number of subrectangles in \( \Delta_{\text{rec}} \). Thus, the size of this region shrinks as \( n \to \infty \). Specifically, let

\[
\mathcal{W}_x := \{ \bar{x} : t_{\ell, x} - L \leq \bar{x}_j \leq t_{\ell, x} + L, \ell = 1, \ldots, d \}
\]

where \( t_{\ell, x} \) is the closest point in \( \Delta_{\text{rec}} \) which is no greater than \( x_\ell \) and \( L \) is some fixed number to be determined which only depends on \( d, m \) and \( \tilde{m} \). If such a region spans outside of \( X \), take its intersection with \( X \). Then it follows that \( \langle \varphi_h(x_1; x), \varphi_h(x_1; x) \rangle_{\mathcal{W}_x} \lesssim h^d \), and thus

\[
\langle \mathcal{H}(x, x_1), \mathcal{H}(x, x_1) \rangle_X \geq \langle \mathcal{H}(x, x_1), \mathcal{H}(x, x_1) \rangle_{\mathcal{W}_x} \geq h^{-d} \langle \mathcal{H}(x, x_1), \varphi_h(x_1; x) \rangle_{\mathcal{W}_x}^2.
\]

It suffices to show \( |\langle \mathcal{H}(x, x_1), \varphi_h(x_1; x) \rangle_X \rangle_{\mathcal{W}_x} \) can be made sufficiently small such that

\[
|\langle \mathcal{H}(x, x_1), \varphi_h(x_1; x) \rangle_{\mathcal{W}_x} | \geq h^{-q}.
\]

By Lemma SA-8.1, the elements of \( h^d Q^{-1}_m \) and \( h^d Q^{-1}_{\tilde{m}} \) exponentially decays when they get far away from the (block) diagonals. In view of Assumption SA-5, with \( x \) fixed, \( \mathcal{H}(x, x_1) \) also expo-
nentially decays as \( \|x_1 - x\| \) increases. Formally, we can write

\[
\mathcal{K}(x, x_1) = \partial^q p(x)' Q^{-1}_m p(x_1) + \sum_{u \in \Lambda_m} \partial^u \tilde{p}(x)' h_x^{m - |q|} B_{u,q}(x) Q^{-1}_m \tilde{p}(x_1) \\
- \partial^q p(x)' Q^{-1}_m \mathbb{E} \left[ p(x_i) h_x^u B_{u,0}(x_i) \partial^u \tilde{p}(x_i) \right] Q^{-1}_m \tilde{p}(x_1). 
\] (SA-1)

We temporarily change the meaning of subscripts of subrectangles in \( \Delta^{\text{rec}} \): let

\[
\delta^{\text{rec}}_{t_1 \ldots t_d} := \{ \hat{x} : t_{\ell, t_\ell} \leq \hat{x}_\ell \leq t_{\ell, t_\ell + 1}, \ell = 1, \ldots, d \}
\]

with \( \delta^{\text{rec}}_0 \) denoting the subrectangle where \( x \) is located, and we index other subrectangles with \( \delta^{\text{rec}}_0 \) as the “origin”. First notice that for any given point \( x_0 \in \mathcal{X} \), \( p(x_0) \) and \( \tilde{p}(x_0) \) are two vectors containing a fixed number of nonzeros and all their elements are bounded by some universal constant. Their nonzero elements are obtained by evaluating those basis functions with local supports covering \( x_0 \). Moreover, \( h^{-d} \mathbb{E}[p(x_i) h_x^m B_{u,0}(x_i) \partial^u \tilde{p}(x_i)] \) also admits a block banded structure in the sense that only the products of basis functions with overlapping supports are nonzero, and all elements in this matrix is bounded by some universal constant. Hence for

\[
r_1(x)' = h^{[q]} \partial^q p(x)' Q^{-1}_m \mathbb{E} \left[ \sum_{u \in \Lambda_m} p(x_i) B_{u,0}(x_i) \partial^u \tilde{p}(x_i) \right] \quad \text{and} \quad r_2(x_1)' = h^{d} \tilde{p}(x_1)' Q^{-1}_m,
\]

we can find another two vectors \( \tilde{r}_1(x) \) and \( \tilde{r}_2(x_1) \) with strictly positive elements which are greater than the absolute values of the corresponding elements in \( r_1(x) \) and \( r_2(x_1) \), i.e., \( \tilde{r}_1(x) \) and \( \tilde{r}_2(x_1) \) are bounds on the magnitudes of \( r_1(x) \) and \( r_2(x_1) \). Moreover, the elements of \( \tilde{r}_1(x) \) and \( \tilde{r}_2(x_1) \) are decaying exponentially with some rates \( \vartheta_1, \vartheta_2 \in (0, 1) \) when they get far away from the positions of those basis functions whose supports are around \( x \) and \( x_1 \) in \( p(\cdot) \) and \( \tilde{p}(\cdot) \) respectively. Notice that for some constant \( \vartheta \in (0, 1) \), \( \{ \sum_{l=i}^{\infty} \vartheta^l \}_{i=1}^{\infty} \) is also a geometric sequence. Therefore, the inner product between \( \tilde{r}_1(x) \) and \( \tilde{r}_2(x_1) \) (the third term in (SA-1)) exponentially decays when \( \|x_1 - x\| \) increases. Similarly, the inner products between \( \partial^q p(x) \) and \( Q^{-1}_m p(x_1) \), \( \sum_{u \in \Lambda_m} \partial^u \tilde{p}(x) h_x^{m - |q|} B_{u,q}(x) \) and \( Q^{-1}_m \tilde{p}(x_1) \) have the same property.

Given these results, we have for some \( \vartheta \in (0, 1) \)

\[
\|h^{[q]} + d, \mathcal{K}_h(x, \cdot)\|_{L_\infty(\delta_{t_1 \ldots t_d})} \leq C \vartheta^{\sum_{\ell=1}^d |l|}.
\]

78
Meanwhile,
\[ \| \varphi_h(\cdot; \mathbf{x}) \|_{L^\infty(\mathcal{G}^\text{rec}_{[1,...,d]})} \lesssim (|l_1| + 1)^{q_1} \cdots (|l_d| + 1)^{q_d}. \]

Let \( \mathbf{l} = (l_1, \cdots, l_d) \) and \( \mathbf{1} = (1, \cdots, 1) \). Denote \( |\mathbf{l}| = (|l_1|, \ldots, |l_d|) \). The above results imply
\[ |\langle \mathcal{X}_h(\mathbf{x}, \mathbf{x}_1), \varphi_h(\mathbf{x}_1; \mathbf{x}) \rangle_{\mathcal{G}^\text{rec}_{[1,...,d]}^1} | \lesssim h^{-[q]}(|\mathbf{l}| + 1)^{q_d[|\mathbf{l}|]}. \]

Then the desired result follows from the fact that
\[ \sum_{|\mathbf{l}|=0}^{\infty} (|\mathbf{l}| + 1)^{q_d[|\mathbf{l}|]} \]
exists. Therefore, we can choose \( L \) large enough which is independent of \( \mathbf{x} \) and \( n \) such that
\[ |\langle \mathcal{X}_h(\mathbf{x}, \mathbf{x}_1), \varphi_h(\mathbf{x}_1; \mathbf{x}) \rangle_{\mathcal{W}_k} | \gtrsim h^{-[q]}. \] Then the proof is complete.

\[ \square \]

**SA-14.6 Proof of Lemma SA-9.1**

*Proof.* For \( j = 0, 1 \), the results directly follow from Assumption SA-5, Lemma SA-8.3 and [3, Lemma 4.1].

For \( j = 2, 3 \), conditional on \( \mathbf{X} \), \( R_{1n,q}(\mathbf{x}) \) has mean zero, and its variance can be bounded as follows:
\[
\forall [R_{1n,q}(\mathbf{x})|\mathbf{X}] \lesssim \frac{1}{n} \left[ \gamma_{q,j}(\mathbf{x})' - \gamma_{q,1}(\mathbf{x})' \right] \mathbb{E}[\Pi_j(\mathbf{x}_i)\Pi_j(\mathbf{x}_i)'] \left[ \tilde{\gamma}_{q,j}(\mathbf{x}) - \gamma_{q,j}(\mathbf{x}) \right] \\
\lesssim \frac{1}{n} \left| \gamma_{q,j}(\mathbf{x})' - \gamma_{q,j}(\mathbf{x})' \right|^2 \mathbb{E}[\Pi_j(\mathbf{x}_i)\Pi_j(\mathbf{x}_i)'] \\
\lesssim \frac{1}{n} h^{-2[q]-2d} \frac{\log n}{nh^d} h^d \\
\lesssim \frac{\log n}{n^2 h^{2d+2[q]}} \\
\] where the third line follows from the Lemma SA-8.4 and the fact that \( \mathbb{E}[\Pi_j(\mathbf{x}_i)\Pi_j(\mathbf{x}_i)'] \lesssim h^d \) shown in the proof of Lemma SA-8.5. Then by Chebyshev’s inequality we conclude that
\[ R_{1n,q}(\mathbf{x}) \lesssim \mathbb{P}\frac{\sqrt{\log n}}{nh^{d+2[q]}}. \]
For the conditional bias $R_{2n,q}(x)$, we analyze the least-squares bias correction and plug-in bias correction separately. For $j = 2$, by construction,

$$
\mathbb{E}[\hat{\partial}_q^2 \mu_2(x)|X] - \partial_q \mu(x)
$$

$$
= \left( \mathbb{E}[\hat{\partial}_q \mu_1(x)|X] - \partial_q \mu(x) \right) - \partial_q p(x)' \hat{Q}_m^{-1} \mathbb{E}_n[p(x_i) \mathcal{B}_{\hat{m},0}(x_i)]
$$

$$
= O_P(h^{m+e-|q|}) - \partial_q p(x)' \hat{Q}_m^{-1} \mathbb{E}_n[p(x_i) \mathcal{B}_{\hat{m},0}(x_i)]
$$

where $\mathcal{B}_{\hat{m},0}(x_i) = \mathbb{E}[\hat{\mu}_1(x_i)|X] - \mu(x_i)$ is the conditional bias of $\hat{\mu}_1(x_i)$ and the last line follows from Lemma SA-8.3. Since

$$
\| \mathbb{E}_n[p(x_i) \mathcal{B}_{\hat{m},0}(x_i)] \|_{\infty} \leq \sup_{x \in \chi} |\mathcal{B}_{\hat{m},0}(x)| \| \mathbb{E}_n[|p(x_i)|] \|_{\infty},
$$

using the same proof strategy as that for Lemma SA-8.3, we have $\| \mathbb{E}_n[|p(x_i)|] \|_{\infty} \lesssim_P h^d$. Also, by Lemma SA-8.3, $\sup_{x \in \chi} |\mathcal{B}_{\hat{m},0}(x)| \lesssim_P h^{m+e}$. Then it follows from Lemma SA-8.4 and that the conditional bias of $\hat{\partial}_q \mu_2(x)$ is $O_P(h^{m+e-|q|})$.

Next, for $j = 3$, using Lemma SA-8.3, we have

$$
\mathbb{E}[\hat{\partial}_q^3 \mu_3(x)|X] - \partial_q \mu(x)
$$

$$
= \mathbb{E} \left[ \hat{\partial}_q^3 \mu_3(x) + \sum_{u \in \Lambda_m} \hat{\partial}_u^3 \mu_1(x) h_x^{m-|q|} B_{u,q}(x) + \partial_q p(x)' \hat{Q}_m^{-1} \mathbb{E}_n[p(x_i) \mathcal{B}_{\hat{m},0}(x_i)|X] \right] - \partial_q \mu(x)
$$

$$
= \sum_{u \in \Lambda_m} h_x^{m-|q|} B_{u,q}(x) \mathbb{E} \left[ \hat{\partial}_u^3 \mu_1(x) - \partial_u^3 \mu(x)|X \right] + \partial_q p(x)' \hat{Q}_m^{-1} \mathbb{E}_n \left[ p(x_i) \left( \mathbb{E}[\mathcal{B}_{m,0}(x_i)|X] - \mathcal{B}_{m,0}(x_i) \right) \right] + O_P(h^{m+e-|q|})
$$

$$
= \partial_q p(x)' \hat{Q}_m^{-1} \mathbb{E}_n \left[ p(x_i) \left( \mathbb{E}[\mathcal{B}_{m,0}(x_i)|X] - \mathcal{B}_{m,0}(x_i) \right) \right] + O_P(h^{m+e-|q|})
$$

where $\mathcal{B}_{m,0}(x) = -\sum_{u \in \Lambda_m} \left( \hat{\partial}_u^3 \mu_1(x) \right) h_x^{m} B_{u,q}(x)$, and the last line follows from Assumption SA-4, SA-5 and Lemma SA-8.3. Also by Lemma SA-8.3 and the fact that $B_{u,0}(\cdot)$ is bounded, $\sup_{x \in \chi} |\mathbb{E}[\mathcal{B}_{m,0}(x)|X] - \mathcal{B}_{m,0}(x)| \lesssim_P h^{m+e}$. The desired result immediately follows by using the similar argument for $j = 2$. □
Proof. (a) This is a direct application of [3, Theorem 4.2] combined with our Lemma SA-9.1.

(b) By Lemma SA-8.5 and Lemma SA-9.1, it suffices to show that \( \gamma_{q,j}(x) \mathbb{G}_n[\Pi_j(x_i)\varepsilon_i] / \sqrt{\Omega_j(x)} \) weakly converges to a standard normal, for \( j = 1, 2, 3 \). First, by construction

\[
\forall \left[ \frac{\gamma_{q,j}(x)^\prime}{\sqrt{\Omega_j(x)}} \mathbb{G}_n[\Pi_j(x_i)\varepsilon_i] \right] = 1.
\]

Next, we write \( a_{ni} := \frac{\gamma(x)_{ni}^{\prime} \Pi_j(x_i)}{\sqrt{\Omega_j(x)}} \). For all \( \vartheta > 0 \),

\[
\sum_{i=1}^{n} E \left[ a_{ni}^2 \varepsilon_i^2 1 \{ |a_{ni}\varepsilon_i / \sqrt{n} > \vartheta \} \right] \leq \left[ E \left[ a_{ni}^2 \varepsilon_i^2 1 \{ |\varepsilon_i| > \vartheta \sqrt{n} / |a_{ni}| \} |x_i\right] \right]
\]

\[
\leq E[a_{ni}^2] \cdot \sup_{x \in \mathcal{X}} E \left[ \varepsilon_i^2 1 \{ |\varepsilon_i| > \vartheta \sqrt{n} / |a_{ni}| \} |x_i = x\right] \]

\[
\leq h^{-2d}/(h^{-d} / \log n)
\]

for all \( \vartheta > 0 \), where the third line follows from Lemma SA-8.4 and Lemma SA-8.5. Since \( |a_{ni}| \lesssim h^{-d/2} \) and \( \log n/(nh^d) = o(1) \), it follows that \( \vartheta \sqrt{n} / |a_{ni}| \to \infty \) as \( n \to \infty \). By the moment condition in the theorem, the upper bound in Eq. (SA-2) goes to 0 as \( n \) goes to infinity, which completes the proof of part (b).

\[\square\]

Proof of Lemma SA-10.1

Proof. The proof is divided into two steps.

**Step 1:** We first bound \( \sup_{x \in \mathcal{X}} |R_{1n,q}(x)| \) for \( j = 0, 1, 2, 3 \). We truncate the errors by an increasing sequence of constants \( \vartheta_n : n \geq 1 \) such that \( \vartheta_n \propto \sqrt{nh^d / \log n} \). To simplify notation, we write \( \Pi_j(x_i) = (\pi_1(x_i), \ldots, \pi_{K_j}(x_i))' \) where \( K_j = \dim(\Pi_j(\cdot)) \). Then let \( H_{ik} = \pi_k(x_i) (\varepsilon_i 1 \{ |\varepsilon_i| \leq \vartheta_n \}) - E[\varepsilon_i 1 \{ |\varepsilon_i| \leq \vartheta_n |x_i\}] \) and \( T_{ik} = \pi_k(x_i) (\varepsilon_i 1 \{ |\varepsilon_i| > \vartheta_n \}) - E[\varepsilon_i 1 \{ |\varepsilon_i| > \vartheta_n |x_i\}] \). Regarding the truncated term, it follows from the truncation strategy, Assumption SA-3 and SA-5 that \( |H_{ik}| \leq \vartheta_n \)
and $\mathbb{E}[H_{ik}^2] \lesssim h^d$. By Bernstein’s inequality, for $t > 0$,

$$
\mathbb{P}\left( \max_{1 \leq k \leq K_j} |\mathbb{E}_n[H_{ik}]| > h^d \sqrt{\log n/(nh^d)t} \right) 
\leq 2 \sum_{k=1}^{K_j} \exp \left\{ -\frac{n^2h^{2d}h^{-d} \log n t^2/n}{C_1nh^d + C_2\vartheta_n nh^d \sqrt{\log n/(nh^d)t}} \right\} 
\leq C \exp \left\{ \log n \left( 1 - \frac{t^2}{C_1 + C_2\vartheta_n \sqrt{\log n/(nh^d)t}} \right) \right\} \quad (SA-3)
$$

which is arbitrarily small for $t$ large enough by the truncation strategy. It immediately follows from Lemma SA-8.4 that

$$
\sup_{x \in \mathcal{X}} \left| \tilde{\gamma}_{q,j}(x) - \gamma_{q,j}(x) \right| \lesssim h^{-[q-d]} \sqrt{\log n/(nh^d)}h^d \sqrt{\log n/(nh^d)} = h^{-[q]} \log n/(nh^d).
$$

Regarding the tails, let $\mathcal{X}_{j,i}(x) := (\tilde{\gamma}_{q,j}(x) - \gamma_{q,j}(x))\Pi_j(x_i)$. By Lemma SA-8.4 and Assumption SA-3, we have

$$
\sup_{x \in \mathcal{X}} |\mathcal{X}_{j,i}(x)| \lesssim h^{-[q]} \sqrt{\log n/(nh^d)}.
$$

Let $\mathcal{A}_n(M)$ denote the event on which $\sup_{x \in \mathcal{X}} |\mathcal{X}_{j,i}(x)| \leq Mh^{-[q]} \sqrt{\log n/(nh^d)}$ for some $M > 0$, and $\mathbb{I}_{\mathcal{A}_n(M)}$ be an indicator function of $\mathcal{A}_n(M)$. Then by Markov’s inequality, for $t > 0$,

$$
\mathbb{P}\left( \sup_{x \in \mathcal{X}} \left| \mathbb{E}_n[\mathbb{I}_{\mathcal{A}_n(M)}\mathcal{X}_{j,i}(x) \mathbb{I}\{|\xi_i| > \vartheta_n\} - \mathbb{E}^{\mathbb{I}_{\mathcal{A}_n(M)}}[\mathbb{I}\{|\xi_i| > \vartheta_n\}] \right| > th^{-[q]} \log n/(nh^d) \right) 
\lesssim \frac{Mh^{-[q]} \sqrt{\log n/(nh^d)} \mathbb{E}[\mathbb{I}_{\{\mathbb{I}\{|\xi_i| > \vartheta_n\}\}}]}{th^{-[q]} \log n/(nh^d)} 
\leq \frac{M\sqrt{n}}{t \sqrt{h^d \log n}} \frac{\mathbb{E}[\mathbb{I}_{\{\mathbb{I}\{|\xi_i| > \vartheta_n\}\}}]}{\vartheta_n^{1+\nu}} \quad (SA-4)
$$

which is arbitrarily small for $t/M$ large enough by the additional moment condition specified in the lemma and the rate restriction. Since $\mathbb{P}(\mathcal{A}_n(M)^C) = o(1)$ as $M \to \infty$, simply let $t = M^2$ and $M \to \infty$, then the desired conclusion immediately follows.

**Step 2**: Next, we bound $\sup_{x \in \mathcal{X}} |R_{2n,q}(x)|$. For $j = 0, 1$, the result directly follows from Lemma
SA-8.3. For $j = 2$, notice that the proof of Lemma SA-9.1 essentially establishes a bound on the uniform norm of $\partial^q p(x)Q_1^{\partial^q \mu(x)}[p(x_i)\mathcal{B}_{m,0}(x_i)]$. The bound on $\mathbb{E}[\hat{\mathcal{H}}^q(x)] - \partial^q \mu(x)$ follows from Lemma SA-8.3. Then the desired bound on $R_{2n,q}(x)$ is obtained.

For $j = 3$, notice that by Assumption SA-4 we can write $R_{2n,q}(x)$ explicitly as

$$R_{2n,q}(x) = \hat{\gamma}_{q,0}(x) \left\{ \mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i))] - \mathbb{E}_n \left[ \mathbb{E}_m B_{u,0}(x_i) \mathbb{E} \left[ \partial^u \mu_1(x_i) | X \right] \right] \right\}$$

$$+ \partial^q s^*(x) - \partial^q \mu(x) - \mathcal{B}_{m,q}(x)$$

$$+ \mathcal{B}_{m,q}(x) + \sum_{u \in \Lambda_m} h_x^{m-q} B_{u,q}(x) \partial^u \tilde{p}(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[\tilde{p}(x_i) \mu(x_i)].$$

The second line is uniformly bounded by Assumption SA-4: $\sup_{x \in X} |\partial^q s^*(x) - \partial^q \mu(x) - \mathcal{B}_{m,q}(x)| \lesssim h^{m-\epsilon-[q]}$.

Next, the third line can be written as

$$\mathcal{B}_{m,q}(x) + \sum_{u \in \Lambda_m} h_x^{m-q} B_{u,q}(x) \partial^u \tilde{p}(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[\tilde{p}(x_i) \mu(x_i)]$$

$$= \sum_{u \in \Lambda_m} h_x^{m-q} B_{u,q}(x) \left( \partial^u \tilde{p}(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[\tilde{p}(x_i) \mu(x_i)] - \partial^u \mu(x) \right)$$

It follows from Assumption SA-4 that $\sup_{x \in X} |B_{u,q}(x)| \lesssim 1$. Moreover, as we have shown in the proof of Lemma SA-8.3,

$$\sup_{x \in X} |\partial^u \tilde{p}(x)' \tilde{Q}_m^{-1} \mathbb{E}_n[\tilde{p}(x_i) \mu(x_i)] - \partial^u \mu(x)| \lesssim h^{m-\epsilon-[q]}$$

which suffices to show that the third line is $O_{\mathbb{P}}(h^{m+\epsilon-[q]}).

Finally, it also follows from the proof of Lemma SA-8.3 that

$$\sup_{x \in X} \left| \hat{\gamma}_{q,0}(x)' \mathbb{E}_n[p(x_i)(\mu(x_i) - s^*(x_i) + \mathcal{B}_{m,0}(x_i))] \right| \lesssim_{\mathbb{P}} h^{m+\epsilon-[q]},$$

and

$$\sup_{x \in X} \left| \hat{\gamma}_{q,0}(x)' \mathbb{E}_n[p(x_i) h_x^m B_{u,0}(x_i) \left( \mathbb{E}[\partial^u \mu_1(x_i) | X] - \partial^u \mu(x_i) \right) \right| \lesssim_{\mathbb{P}} h^{m+\epsilon-[q]}.$$ 

Then the proof is complete. \hfill \square
SA-14.9 Proof of Lemma SA-10.2

Proof. It suffices to adjust the proof for $R_{1n,q}(x)$. We still use the same proof strategy, but let $\vartheta_n = \log n$. For the truncated term, it can be seen from the Bernstein’s inequality (Equation SA-3) that the upper bound can be made arbitrarily small for $t$ large enough when $(\log n)^3/(nh^d) \lesssim 1$. On the other hand, when applying Markov’s inequality to control the tail (Equation (SA-4)), we employ the exponential moment condition:

$$
P\left(\sup_{x \in X} \left| E_n[\mathbb{1}_{A_n(M)}\mathcal{X}_{ji}(\varepsilon_i \mathbb{1}\{|\varepsilon_i| > \vartheta_n\} - E[\varepsilon_i \mathbb{1}\{|\varepsilon_i| > \vartheta_n\}])]\right| > th^{-[q]}\log n/(nh^d)\right) \lesssim M \frac{\sqrt{\log n}}{t \sqrt{h^d}} E[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] \\
\lesssim M \frac{\sqrt{\log n}}{t \sqrt{h^d}} \vartheta_n^2 \exp(\vartheta_n) \\
\leq \frac{M}{t (\log n)^{5/2} h^d} E[|\varepsilon_i|^3 \exp(|\varepsilon_i|)]$$

which is arbitrarily small for $t/M$ large enough. Thus the same bound on $R_{1n,q}$ is established. □

SA-14.10 Proof of Theorem SA-10.1

Proof. Consider the case when the conditions of Lemma SA-10.1 hold. We use the same truncation strategy. Specifically, separate $\varepsilon_i$ into

$$\varepsilon_i \mathbb{1}\{|\varepsilon_i| \leq \vartheta_n\} - E[\varepsilon_i \mathbb{1}\{|\varepsilon_i| \leq \vartheta_n\}|x_i] \quad \text{and} \quad \varepsilon_i \mathbb{1}\{|\varepsilon_i| > \vartheta_n\} - E[\varepsilon_i \mathbb{1}\{|\varepsilon_i| > \vartheta_n\}|x_i]$$

where $\vartheta_n \asymp \sqrt{nh^d/\log n}$. By Lemmas SA-8.4,

$$\sup_{x \in X} |\gamma_{q,j}(x)^\top \Pi_j(x_i)| \lesssim h^{-d-[q]}.$$

Then repeating the argument given in the proof of Lemma SA-10.1 for the truncated and tails respectively, we have

$$\sup_{x \in X} |\gamma_{q,j}(x)^\top E_n[\Pi_j(x_i)\varepsilon_i]| \lesssim h^{-[q]} \sqrt{\log n/(nh^d)}.$$
Moreover, \( \tilde{R}_{1n,q} = o(h^{-d/2-[q]} \sqrt{\log n/n}) \) since \( \log n/(nh^d) = o(1) \). Combining these bounds with the results in Lemma SA-10.1, we obtain the desired rate of uniform convergence.

Finally, the same results can be proved under conditions of Lemma SA-10.2 if we let \( \vartheta_n = \log n \) and assume \( (\log n)^3/(nh^d) \lesssim 1 \). \(\square\)

**SA-14.11 Proof of Theorem SA-10.2**

*Proof.* Notice that for \( j = 0, 1, 2, 3 \),

\[
\tilde{\Sigma}_j - \Sigma_j = \mathbb{E}_n[(\tilde{\varepsilon}_j^2 - \varepsilon_j^2)\Pi_j(x_i)\Pi_j(x_i')] + \left( \mathbb{E}_n[\varepsilon_j^2\Pi_j(x_i)\Pi_j(x_i')] - \Sigma_j \right). \tag{SA-5}
\]

We then divide our proof into several steps.

**Step 1:** For \( \|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i')]\| \), by Assumption SA-3, SA-5 and the same argument as that for \( \|\hat{Q}_m - Q_m\| \),

\[
\|\mathbb{E}[\Pi_j(x_i)'\Pi_j(x_i)']\| \lesssim h^d \quad \text{and} \quad \|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i')] - \mathbb{E}[\Pi_j(x_i)\Pi_j(x_i)']\| \lesssim h^d \sqrt{\log n/(nh^d)}.
\]

By the triangle inequality and the fact that \( \log n/nh^d = o(1) \), \( \|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i')]\| \lesssim h^d \).

**Step 2:** Next, we control the magnitude of the second term in Equation (SA-5). To simplify our notations, let \( L_j(x_i) := W_j^{-1/2}\Pi_j(x_i) \) be the normalized basis where \( W_j = Q_m \) for \( j = 0 \), \( W_j = Q_{\tilde{m}} \) for \( j = 1 \) and \( W = \text{diag}(Q_m, Q_{\tilde{m}}) \) for \( j = 2, 3 \). Introduce a sequence of positive numbers:

\[
M_n^2 \asymp \frac{K^{1+1/\nu}n^{1/(2+\nu)}}{(\log n)^{1/(2+\nu)}}.
\]

Then we write

\[
H_j(x_i) = \varepsilon_j^2L_j(x_i)L_j(x_i)' \mathbb{I}\{\|\varepsilon_j^2L_j(x_i)L_j(x_i)\| \leq M_n^2\}, \quad \text{and}
\]

\[
T_j(x_i) = \varepsilon_j^2L_j(x_i)L_j(x_i)' \mathbb{I}\{\|\varepsilon_j^2L_j(x_i)L_j(x_i)\|> M_n^2\}.
\]

Clearly,

\[
\mathbb{E}_n[L_j(x_i)L_j(x_i)'\varepsilon_i^2] - \mathbb{E}[L_j(x_i)L_j(x_i)'\varepsilon_i^2] = \frac{1}{n} \sum_{i=1}^{n} (H_j(x_i) - \mathbb{E}[H_j(x_i)]) + \frac{1}{n} \sum_{i=1}^{n} (T_j(x_i) - \mathbb{E}[T_j(x_i)]).
\]
For the truncated terms, by definition, \( \|H_j(x_i)\| \leq M_n^2 \). It follows from the triangle inequality and Jensen’s inequality that \( \|H_j(x_i) - \mathbb{E}[H_j(x_i)]\| \leq 2M_n^2 \). In addition, by Assumption SA-1,

\[
\mathbb{E}[(H_j(x_i) - \mathbb{E}[H_j(x_i)])^2] \leq \mathbb{E}[\varepsilon_i^4 \|L_j(x_i)\|^2 L_j(x_i) L_j(x_i)'] I \{\|\varepsilon_i^2 L_j(x_i) L_j(x_i)'\| \leq M_n^2\}
\]

\[
\leq M_n^2 \mathbb{E}[\varepsilon_i^2 L_j(x_i) L_j(x_i)'] I \{\|\varepsilon_i^2 L_j(x_i) L_j(x_i)'\| \leq M_n^2\}
\]

\[
\lesssim M_n^2 \mathbb{E}[L_j(x_i) L_j(x_i)']
\]

where the inequalities are in the sense of semi-definite matrices. Hence \( \|\mathbb{E}[(H_j(x_i) - \mathbb{E}[H_j(x_i)])^2]\| \lesssim M_n^2 \). Let

\[
\vartheta_n = \sqrt{\frac{(\log n)^{2+\nu}}{n^{2+\nu} h^d}}.
\]

By an inequality of [58] for independent matrices, we have for all \( t > 0 \),

\[
\mathbb{P} \left( \left\| \frac{1}{n} \sum_{i=1}^{n} (H_j(x_i) - \mathbb{E}[H_j(x_i)]) \right\| > \vartheta_n t \right) \leq \exp \left( \log n - \frac{\vartheta_n^2 t^2/2}{M_n^2/n + M_n^2 \vartheta_n t/(3n)} \right)
\]

\[
\leq \exp \left\{ \log n \left( 1 - \frac{t^2/2}{M_n^2 \log n \vartheta_n^{-2} n^{-1} (1 + \vartheta_n t/3)} \right) \right\}
\]

where \( M_n^2 \log n \vartheta_n^{-2} n^{-1} \asymp (\log n)^{1/(2+\nu)} / (n^{1/(2+\nu)} h^{d/\nu}) = o(1) \) and \( \vartheta_n = o(1) \). Hence we have

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} (H_j(x_i) - \mathbb{E}[H_j(x_i)]) \right\| \lesssim \mathbb{P} \vartheta_n = o(1).
\]

Regarding the tails, it directly follows from Lemma SA-8.1 that

\[
\|T_j(x_i)\| \lesssim h^{-d} \varepsilon_i^2 \mathbb{1} \{ \varepsilon_i^2 \gtrsim M_n^2 h^d \}.
\]

Then by the triangle inequality, Jensen’s inequality and the assumption that \( (2 + \nu) \)th moment of \( \varepsilon_i \) is bounded, we have

\[
\mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} (T_j(x_i) - \mathbb{E}[T_j(x_i)]) \right\| \right] \lesssim 2h^{-d} \mathbb{E}[\varepsilon_i^2 \mathbb{1} \{ |\varepsilon_i| \gtrsim M_n \sqrt{h^d} \}]
\]

\[
\lesssim \frac{2h^{-d(1+\nu)/2} \mathbb{E}[\varepsilon_i^{2+\nu} \mathbb{1} \{ |\varepsilon_i| \gtrsim M_n \sqrt{h^d} \}]}{M_n^\nu}
\]

\[
\lesssim \vartheta_n = o(1).
\]
By Markov's inequality, we have \( \frac{1}{n} \sum_{i=1}^{n} (T_j(x_i) - \mathbb{E}[T_j(x_i)]) \| \leq_{P} \vartheta_n \). Since \( \|W_j^{1/2}\| \leq h^{d/2} \) and \( \|W_j^{-1/2}\| \leq h^{-d/2} \), we conclude that

\[
\|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i)'\varepsilon_i^2] - \Sigma_j\| \leq_{P} h^d \vartheta_n = o_P(h^d).
\]

**Step 3:** The first term in Equation (SA-5) satisfies

\[
\|\mathbb{E}_n[(\tilde{\varepsilon}_{i,j}^2 - \varepsilon_{i,j}^2)\Pi_j(x_i)\Pi_j(x_i)']\|
\leq \|\mathbb{E}_n[(\mu(x_i) - \tilde{\mu}_j(x_i))^2\Pi_j(x_i)\Pi_j(x_i)']\| + 2\|\mathbb{E}_n[(\mu(x_i) - \tilde{\mu}_j(x_i))\varepsilon_i\Pi_j(x_i)\Pi_j(x_i)']\|
\leq \max_{1 \leq i \leq n} |\mu(x_i) - \tilde{\mu}_j(x_i)|^2 \|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i)']\| + 
\max_{1 \leq i \leq n} |\mu(x_i) - \tilde{\mu}_j(x_i)(\|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i)']\| + \|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i)']\|^2 |)
\]

where the last line follows from the fact that \( 2|a| \leq 1 + a^2 \). By Theorem SA-10.1 and the results proved in Step 1 and 2, we have \( \max_{1 \leq i \leq n} |\mu(x_i) - \tilde{\mu}_j(x_i)| = R_{0,j}^{ac} = o_P(1) \), \( \|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i)']\| \leq_{P} h^d \) and \( \|\mathbb{E}_n[\Pi_j(x_i)\Pi_j(x_i)']\|^2 \| \leq_{P} h^d \). Hence we conclude that

\[
\|\tilde{\Sigma}_j - \Sigma_j\| \leq_{P} h^d (R_{0,j}^{ac} + \vartheta_n) = o_P(h^d)
\]  

**Step 4:** Using all above results, we have

\[
|\tilde{\Omega}_j(x) - \Omega_j(x)|
= |\tilde{\gamma}_{q,j}(x)'\tilde{\Sigma}_j\tilde{\gamma}_{q,j}(x) - \gamma_{q,j}(x)'\Sigma_j\gamma_{q,j}(x)|
\leq |(\tilde{\gamma}_j(x)' - \gamma_{q,j}(x))'\Sigma_j\tilde{\gamma}_{q,j}(x)| + |\gamma_{q,j}(x)'(\tilde{\Sigma}_j - \Sigma_j)\tilde{\gamma}_{q,j}(x)|
+ |\gamma_{q,j}(x)'\Sigma_j(\tilde{\gamma}_{q,j}(x) - \gamma_{q,j}(x))|
\]

By Lemma SA-8.4 and Equation (SA-6), we have

\[
\sup_{x \in X} |\tilde{\Omega}_j(x) - \Omega_j(x)| \leq_{P} h^{-d-2|q|(R_{0,j}^{ac} + \vartheta_n)} = o_P(h^{-d-2|q|}).
\]

Then the proof is complete. \(\square\)
SA-14.12 Proof of Theorem SA-10.3

Proof. We only need to adjust Step 2 in the proof of Theorem SA-10.2. Apply the same proof strategy with \( M_n = \sqrt{Ch^d \log n} \) and \( \vartheta_n = \sqrt{\frac{(\log n)^3}{nh^d}} \). For the truncated term, since \( M_n^2 \log n/\vartheta_n^2 n \leq 1 \), by Bernstein’s inequality, \( \| \frac{1}{n} \sum_{i=1}^{n} \mathbf{H}_j(x_i) - \mathbb{E}[\mathbf{H}_j(x_i)] \| \lesssim_P \vartheta_n = o_P(1) \). On the other hand, when applying Markov’s inequality to bound the tail, we employ the stronger moment condition:

\[
\mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} (T_j(x_i) - \mathbb{E}[T(x_i)]) \right\| \right] \lesssim 2h^{-d} \mathbb{E}[\varepsilon_i^3 \mathbb{I}\{|\varepsilon_i| \geq M_n/\sqrt{Ch^d}\}]
\]

\[
\lesssim \frac{h^{-3d/2}}{M_n \exp(M_n/\sqrt{Ch^d})} \mathbb{E}[(|\varepsilon_i|^3 \exp(|\varepsilon_i|))]
\]

\[
\lesssim \vartheta_n = o(1).
\]

Then the results follow from the same argument as shown in the proof of Theorem SA-10.2. \( \square \)

SA-14.13 Proof of Lemma SA-10.3

Proof. By Lemma SA-8.5 and Theorem SA-10.2, we have

\[
\sup_{x \in \mathcal{X}} \left| \frac{\hat{\Omega}_j^{1/2} - \Omega_j^{1/2}}{\hat{\Omega}_j^{1/2}} \right| \leq \sup_{x \in \mathcal{X}} \left| \frac{\hat{\Omega}_j(x) - \Omega_j(x)}{\hat{\Omega}_j(x)^{1/2} + \Omega_j(x)^{1/2}} \right| \frac{1}{\hat{\Omega}_j(x)^{1/2}}
\]

\[
\lesssim_P R_{\theta,0} + \frac{n^{\frac{1}{2+\nu}} (\log n)^{\frac{\nu}{2+\nu}}}{\sqrt{nh^d}}
\]

\[
\lesssim_P n^{\frac{1}{2+\nu}} \sqrt{(\log n)^{\frac{\nu}{2+\nu}}/(nh^d)} + h^{m+\nu}.
\] (SA-7)

Then it follows from Theorem SA-10.1 and Lemma SA-8.5 that

\[
\sup_{x \in \mathcal{X}} \left| \frac{\partial^q \mu_j(x) - \partial^q \mu(x)}{\sqrt{\hat{\Omega}_j(x)/n}} \right| - \frac{\partial^q \mu_j(x) - \partial^q \mu(x)}{\sqrt{\hat{\Omega}_j(x)/n}}
\]

\[
\leq \sup_{x \in \mathcal{X}} \left| \frac{\partial^q \mu_j(x) - \partial^q \mu(x)}{\sqrt{\hat{\Omega}_j(x)/n}} \right| \sup_{x \in \mathcal{X}} \left| \frac{\hat{\Omega}_j(x)^{1/2} - \Omega_j(x)^{1/2}}{\hat{\Omega}_j(x)^{1/2}} \right|
\]

\[
\lesssim_P \left( \sqrt{\log n} + \sqrt{n h^m + d/2} \right) \left( n^{\frac{1}{2+\nu}} \sqrt{(\log n)^{\frac{\nu}{2+\nu}}/(nh^d)} + h^{m+\nu} \right)
\]

\[
= o_P(r_n^{-1})
\]

88
where the last line follows from the additional rate restrictions given in the lemma.

SA-14.14 Proof of Lemma SA-10.4

Proof. The proof is exactly the same as that for Lemma SA-10.3 except that the exponential moment condition is employed. Using Theorem SA-10.1 and SA-10.3, the same result is true if \((\log n)^2/\sqrt{nhd} = o(r_n^{-1})\).

SA-14.15 Proof of Theorem SA-10.5

We employ the following lemma in the proof, conditional on \(X\), which is a direct application of [53].

Lemma SA-14.1. Let \(\{\varepsilon_{i,n} : 1 \leq i \leq n\}\) be a triangular array of independent mean-zero random variables with \(\sigma^2_{i,n} := \mathbb{V}[\varepsilon_{i,n}]\) satisfying \(0 < \sigma^2_{i,n} < \infty\).

(a) If for some \(\bar{\nu} > 0\),
\[
\sup_{n \geq 1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[|\varepsilon_{i,n}|^{2+\bar{\nu}}] < \infty,
\]
then there exists a triangular array of standard normal random variables \(\zeta_{i,n}, i = 1, 2, \ldots, n\), row-wise independent, such that
\[
\max_{1 \leq l \leq n} \left| \sum_{i=1}^{l} (\varepsilon_{i,n} - \sigma_{i,n} \zeta_{i,n}) \right| \lesssim_{\mathbb{P}} n^{\frac{1}{2+\bar{\nu}}};
\]

(b) If there exists a constant \(C > 0\) such that for all \(n \geq 1, i = 1, \ldots, n,\)
\[
C \mathbb{E}[|\varepsilon_{i,n}|^{3 \exp(|\varepsilon_{i,n}|)}] \leq \mathbb{E}[\varepsilon_{i,n}^2] \quad \text{and} \quad \sup_{n \geq 1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[|\varepsilon_{i,n}|^{2}] < \infty,
\]
then there exists a triangular array of standard normal random variables \(\zeta_{i,n}, i = 1, 2, \ldots, n,\) row-wise independent, such that
\[
\max_{1 \leq l \leq n} \left| \sum_{i=1}^{l} (\varepsilon_{i,n} - \sigma_{i,n} \zeta_{i,n}) \right| \lesssim_{\mathbb{P}} \log n.
\]
Proof. (a) We rearrange the ordering of \( \{x_i : 1 \leq i \leq n\} \) according to the values of \( \{\mathcal{K}(x, x_i) : 1 \leq i \leq n\} \). Specifically, the rearranged sequence \( \{x_{i,n} : 1 \leq i \leq n\} \) satisfies that \( \mathcal{K}(x, x_{1,n}) \leq \mathcal{K}(x, x_{2,n}) \leq \cdots \leq \mathcal{K}(x, x_{n,n}) \). Accordingly, \( \{(\sigma_i, \epsilon_i, \zeta_i) : 1 \leq i \leq n\} \) can be rearranged in the order induced by \( \{x_{i,n} : 1 \leq i \leq n\} \) to obtain \( \{(\sigma_{i,n}, \epsilon_{i,n}, \zeta_{i,n}) : 1 \leq i \leq n\} \), where \( \sigma_i = \sigma(x_i) \) and \( \sigma_{i,n} = \sigma(x_{i,n}) \). Define

\[
S_{l,n} = \sum_{i=1}^{l} \epsilon_{i,n} - \sum_{i=1}^{l} \sigma_{i,n} \zeta_{i,n}.
\]

Using summation by parts after rearrangement, we have

\[
\left| \sum_{i=1}^{n} \mathcal{K}(x, x_{i,n})(\epsilon_{i,n} - \sigma_{i,n} \zeta_{i,n}) \right|
= \left| \mathcal{K}(x, x_{n,n})S_{n,n} - \sum_{i=1}^{n-1} S_{i,n} (\mathcal{K}(x, x_{i+1,n}) - \mathcal{K}(x, x_{i,n})) \right|
\leq \left( \max_{1 \leq i \leq n} |\mathcal{K}(x, x_i)| + \sum_{i=1}^{n-1} |\mathcal{K}(x, x_{i+1,n}) - \mathcal{K}(x, x_{i,n})| \right) \max_{1 \leq l \leq n} |S_{l,n}|.
\]

After the rearrangement, \( \{x_{i,n}, \epsilon_{i,n}\} \) is no longer i.i.d, but importantly, conditional on \( X \), \( \epsilon_{i,n}'s \) are still independent. Then, we apply Lemma SA-14.1(a) conditional on \( X \), to construct a sequence of i.i.d standard normal random variables \( \zeta_{i,n} \) such that \( \max_{1 \leq l \leq n} |S_{l,n}| \leq P \log n \). Moreover, by construction of the rearranged sequence,

\[
\sup_{x \in \mathcal{X}} \sum_{i=1}^{n-1} |\mathcal{K}(x, x_{i+1,n}) - \mathcal{K}(x, x_{i,n})| = \sup_{x \in \mathcal{X}} (\mathcal{K}(x, x_{n,n}) - \mathcal{K}(x, x_{1,n}))
\leq \sup_{x \in \mathcal{X}} \max_{1 \leq i \leq n} 2|\mathcal{K}(x, x_i)|.
\]

Thus, by the assumption imposed,

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{K}(x, x_i) \epsilon_i = d \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{K}(x, x_i) \sigma_{i,n} \zeta_i + O_P(r_n^{-1}).
\]

(b) The proof is exactly the same as part (a) except that under the stronger moment condition, we have \( \max_{1 \leq l \leq n} |S_{l,n}| \leq P \log n \) by employing Lemma SA-14.1(b). \( \square \)
**SA-14.16 Proof of Theorem SA-10.4**

*Proof.* By Lemma SA-10.3, the remainders in the linearization of $\tilde{T}_j(\cdot)$ is $o_p(r_n^{-1})$ in $L^\infty(\mathcal{X})$, and hence we only need to show $z_j(\cdot)$, as a conditional Gaussian process, approximates $t_j(\cdot)$. Define $\mathcal{H}(x, x_i) := \gamma_{q_j}(x)^{y_j} \Pi_j(x_i) \sigma(x_i) \zeta_i$ for each $j = 0, 1, 2, 3$. It follows Lemma SA-8.4, SA-8.5 and Assumption SA-3 that

$$\sup_{x \in \mathcal{X}} \sup_{u \in \mathcal{X}} |\mathcal{H}(x, u)| \lesssim h^{d/2}.$$ 

By Lemma SA-10.5(a) and the rate restriction given in the theorem, the desired result follows. \Box

**SA-14.17 Proof of Theorem SA-10.5**

*Proof.* The proof is exactly the same as that for Theorem SA-10.4 except that the stronger moment condition allows us to apply Lemma SA-10.5(b). \Box

**SA-14.18 Proof of Theorem SA-10.6**

*Proof.* By Theorem SA-10.4 or Theorem SA-10.5, for $j = 0, 1, 2, 3$, conditional on $X$,

$$z_j(x) = \gamma_{q_j}(x)^{y_j} \Pi_j(x) \sigma(x) \zeta_i$$

$$= d_{x} \gamma_{q_j}(x)^{y_j} \Sigma_{j}^{1/2} N_{K_j}$$

$$= \gamma_{q_j}(x)^{y_j} \Sigma_{j}^{1/2} N_{K_j} + \gamma_{q_j}(x)^{y_j} \left( \Sigma_{j}^{1/2} - \Sigma_{j}^{1/2} \right) N_{K_j}$$

where $\Sigma_{j} := \mathbb{E} \left[ \Pi_j(x_i) \Sigma_j \Pi_j(x_i) \right]$, $N_{K_j}$ 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 Lemma SA-8.4 and SA-8.5,

$$\sup_{x \in \mathcal{X}} \left\| \gamma_{q_j}(x)^{y_j} \right\|_{\infty} \lesssim h^{d/2}.$$ 

In addition, by Gaussian Maximal Inequality (see [21, Lemma 13]),

$$\mathbb{E} \left[ \left\| \left( \Sigma_{j}^{1/2} - \Sigma_{j}^{1/2} \right) N_{K_j} \right\|_{\infty} \mid X \right] \lesssim \sqrt{\log n} \left\| \Sigma_{j}^{1/2} - \Sigma_{j}^{1/2} \right\|.$$ 

91
By the same argument given for Theorem SA-10.2,

$$\|\hat{\Sigma}_j - \Sigma_j\| \lesssim_P h^d \sqrt{\log n/(nh^d)}.$$ 

Then it follows from [4, Theorem X.1.1] that

$$\|\Sigma_j^{1/2} - \Sigma_j^{1/2}\| \leq h^d/2 \left( \frac{\log n}{nh^d} \right)^{1/4}.$$ 

For \(j = 0, 1\), a sharper bound is available: by Theorem X.3.8 of [4] and Lemma SA-8.1,

$$\|\Sigma_j^{1/2} - \Sigma_j^{1/2}\| \leq \frac{1}{\lambda(\Sigma_j)} \|\hat{\Sigma}_j - \Sigma_j\| \lesssim_P h^d/2 \sqrt{\log n/(nh^d)}.$$ 

Thus combining all these results, we have

$$\mathbb{E} \left[ \sup_{x \in X} \left| \gamma_{q,j}^{(x)'} \sqrt{\Omega_j(x)} \left( \Sigma_j^{1/2} - \Sigma_j^{1/2} \right) N_{K_j} \right| \bigg| X \right] \lesssim_P h^{-d/2} \sqrt{\log n} \|\Sigma_j^{1/2} - \Sigma_j^{1/2}\| = o_P(r_n^{-1}).$$ 

where the last equality follows from the additional rate restriction given in the theorem (for \(j = 0, 1\), no additional restriction is needed). By Markov inequality, this suffices to show that for any \(\vartheta > 0\),

$$\mathbb{P} \left( \sup_{x \in X'} \left| z_j(x) - Z_j(x) \right| > \vartheta \bigg| X \right) = o_P(r_n^{-1}).$$ 

Since the conditional probability is bounded, by dominated convergence theorem, the desired result immediately follows. \(\square\)

SA-14.19 Proof of Theorem SA-10.7

Proof. Notice that for \(j = 0, 1, 2, 3\),

$$\tilde{Z}_j(x) - Z_j(x) = \left( \hat{\gamma}_{q,j}^{(x)'} - \gamma_{q,j}^{(x)'} \right) \sqrt{\Omega_j(x)} \Sigma_j^{1/2} N_{K_j}$$

$$+ \gamma_{q,j}^{(x)'} \left( \Omega_j^{-1/2}(x) - \Omega_j^{-1/2}(x) \right) \Sigma_j^{1/2} N_{K_j}$$
\[
\gamma_{q,j}(x)\sqrt{\Omega_j(x)} \left( \hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2} \right) N_{K_j}.
\]

Conditional on the data, each term on the RHS is a mean-zero Gaussian process. Applying Gaussian maximal inequality to each term as in the proof of Theorem SA-10.6 and using Lemma SA-8.4 and Theorem SA-10.2, we have

\[
\mathbb{E} \left[ \sup_{x \in X} \left| \hat{Z}_j(x) - Z_j(x) \right| \right] \lesssim \mathbb{P} \sqrt{\log n} \left( \sqrt{\frac{\log n}{nh^d}} + \left( h^{m+\nu} + \frac{n^{1/2} \sqrt{\log n}}{\sqrt{nh^d}} \right)^{1/2} \right) = o_P(r_n^{-1}).
\]

by the rate restriction given in the theorem. Then the desired result follows by Markov inequality.

For \( j = 0, 1 \), as in the proof of Theorem SA-10.6, a sharper bound on \( \| \hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2} \| \) is available. Thus the last step of the above proof is slightly changed:

\[
\mathbb{E} \left[ \sup_{x \in X} \left| \hat{Z}_j(x) - Z_j(x) \right| \right] \lesssim \mathbb{P} \sqrt{\log n} \left( \sqrt{\frac{\log n}{nh^d}} + a_n + \frac{n^{1/2} (\log n)^{4+2\nu}}{\sqrt{nh^d}} \right) = o_P(r_n^{-1}),
\]

where \( a_n = h^m \) for \( j = 0 \) and \( a_n = h^{m+\nu} \) for \( j = 1 \). The last equality holds without the additional rate restriction.

**SA-14.20 Proof of Theorem SA-10.8**

*Proof.* The proof is exactly the same as that for Theorem SA-10.7 except that Theorem SA-10.3 gives a sharper bound on \( \| \hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2} \| \). □

**SA-14.21 Proof of Theorem SA-10.9**

*Proof.* First notice that for \( j = 0, 1, 2, 3 \),

\[
\sup_{x \in \mathcal{X}} |\hat{z}_j(x) - z_j(x)| = \sup_{x \in \mathcal{X}} \left| \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \mathbb{E} \left[ \Pi_j(x_i) \hat{g}_i(x_i) \right] - \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \mathbb{E} \left[ \Pi_j(x_i) g_i(x_i) \right] \right|
\]

\[
= \sup_{x \in \mathcal{X}} |D_n(x)|
\]

Conditional on the data, \( \{D_n(x), x \in \mathcal{X}\} \) is a Gaussian process with zero means. Let \( \mathbb{E}^{*} \) denote the expectation with respect to the distribution of \( \{ \zeta_i \} \). For notational simplicity, we define a norm
\[ \| \cdot \|_{n,2}^2 = n^{-1} \sum_{i=1}^{n} a_i^2 \text{ for } a \in \mathbb{R}^n. \]

Then by Cauchy-Schwarz inequality,

\[ \mathbb{E}^*[D_n(x)^2] = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\gamma_{q,j}(x)'(x)}{\Omega_j(x)^{1/2}} \Pi_j(x_i) \tilde{\sigma}(x_i) - \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \Pi_j(x_i) \sigma(x_i) \right]^2 \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \Pi_j(x_i) \tilde{\sigma}(x_i) + \gamma_{q,j}(x)'(\Omega_j(x)^{-1/2} - \Omega_j(x)^{-1/2}) \Pi_j(x_i) \tilde{\sigma}(x_i) + \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \Pi_j(x_i) (\tilde{\sigma}(x_i) - \sigma(x_i)) \right\}^2 \]

\[ \lesssim \left\| \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \Pi_j(x_i) \tilde{\sigma}(x_i) \right\|_{n,2}^2 + \left\| \gamma_{q,j}(x)'(\frac{1}{\Omega_j(x)^{1/2}} - \frac{1}{\Omega_j(x)^{1/2}}) \Pi_j(x_i) \tilde{\sigma}(x_i) \right\|_{n,2}^2 + \left\| \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \Pi_j(x_i) (\tilde{\sigma}(x_i) - \sigma(x_i)) \right\|_{n,2}^2. \]

Using Lemma SA-8.4, SA-8.5, Equation (SA-7) and the assumption that \( \max_{1 \leq i \leq n} |\tilde{\sigma}(x_i) - \sigma(x_i)| = o_P(1/(r_n \sqrt{\log n})) \), we have

\[ \sup_{x \in \mathcal{X}} \mathbb{E}^*[D_n(x)^2] \lesssim \frac{\log n}{nh^{d}} + \left( \frac{n^{2/3} (\log n)^{2/3}}{nh^{d}} + a_n \right) + (\max_{1 \leq i \leq n} |\tilde{\sigma}(x_i) - \sigma(x_i)|)^2 \lesssim \frac{1}{r_n^2 \log n}, \]

where \( a_n = h^{2m} \) for \( j = 0 \) and \( a_n = h^{2m+2\varphi} \) for \( j = 1, 2, 3 \). Moreover, using the fact that for \( x, \tilde{x} \in \mathcal{X} \),

\[ (\mathbb{E}^*[\|D_n(x) - D_n(\tilde{x})\|^2])^{1/2} \]

\[ \lesssim \left\| \left( \frac{\gamma_{q,j}(x)_{q,j}(x)'}{\Omega_j(x)^{1/2}} - \frac{\gamma_{q,j}(\tilde{x})'}{\Omega_j(\tilde{x})^{1/2}} \right) \Pi_j(x_i) \tilde{\sigma}(x_i) \right\|_{n,2} + \left\| \left( \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} - \frac{\gamma_{q,j}(\tilde{x})'}{\Omega_j(\tilde{x})^{1/2}} \right) \Pi_j(x_i) \sigma(x_i) \right\|_{n,2}, \]

we will show later in the proof that \( \sup_{\delta \in \Delta} \sup_{x, \tilde{x} \in \text{clo}(\delta)} (\mathbb{E}^*[\|D_n(x) - D_n(\tilde{x})\|^2])^{1/2} \lesssim \mathbb{P}^{-1} \|x - \tilde{x}\| \).

Since \( \sup_{\delta \in \Delta} \text{vol}({\delta}) \lesssim h^d \), and the number of cells in \( \Delta \) is \( k \sim h^{-d} \), we apply Dudley’s inequality (see the proof of Lemma 4.2 of [3] for details of this inequality) to obtain

\[ \mathbb{E}^*\left[ \sup_{x \in \mathcal{X}} |D_n(x)| \right] = o_P(r_n^{-1}) \]

which suffices to show for any \( t > 0 \), \( \mathbb{P}^*(\sup_{x \in \mathcal{X}} |D_n| > tr_n^{-1}) = o_P(1) \) by Markov inequality. Then
the desired result follows from the same argument for Theorem SA-10.6.

In the end, we establish the desired Lipschitz bound on the basis and variance. For any $x, \tilde{x} \in \delta$, by Assumption SA-3, there are only a finite number of basis functions in $p$ and $\tilde{p}$ which are active on $\delta$, and hence it follows from Lemma SA-8.1 that there exists some universal constant $C_1 > 0$ such that $\|\gamma_{q,j}(x) - \gamma_{q,j}(\tilde{x})\| \leq C_1 h^{-[q] - 1} h^{-d} \|x - \tilde{x}\|$. Similarly, by Lemma SA-8.4 and Theorem SA-10.2, $\|\hat{\gamma}_{q,j}(x) - \hat{\gamma}_{q,j}(\tilde{x})\| \lesssim_p h^{-[q] - 1} h^{-d} \|x - \tilde{x}\|$, $|\gamma_{j}(x) - \gamma_{j}(\tilde{x})| \lesssim h^{-[q] - 1} h^{-d} \|x - \tilde{x}\|$, and $|\tilde{\gamma}_{j}(x) - \tilde{\gamma}_{j}(\tilde{x})| \lesssim_p h^{-[q] - 1} h^{-d} \|x - \tilde{x}\|$. Then the proof is complete. 

\[ \square \]

SA-14.22 Proof of Theorem SA-10.10

Proof. Denote by $E^*$ the expectation conditional on the data. By construction, $E^*[\partial^{\mu}_{q,j}(x)] = \hat{\mu}_{q,j}(x)$. Therefore,

\[ \frac{\partial^{\mu}_{q,j}(x) - \partial^{\mu}_{q,j}(\tilde{x})}{\sqrt{\Omega_j(x)/n}} \approx \frac{\hat{\gamma}_{q,j}(x)^{p}}{\Omega_j(x)^{1/2}} G_n[\Pi_j(x_i)\omega_i \hat{\epsilon}_{i,j}]. \]

Note that $\hat{\epsilon}_{i,j} = y_i - \hat{\mu}_j(x_i) = \mu(x_i) - \tilde{\mu}_j(x_i) + \epsilon_i$. Then

\[ \frac{\partial^{\mu}_{q,j}(x) - \partial^{\mu}_{q,j}(\tilde{x})}{\sqrt{\Omega_j(x)/n}} \approx \frac{\gamma_{q,j}(x)^{p}}{\Omega_j(x)^{1/2}} G_n[\Pi_j(x_i)\epsilon_{i,j} \omega_i] + \hat{\gamma}_{q,j}(x)^{p} \approx \frac{\gamma_{q,j}(x)^{p}}{\Omega_j(x)^{1/2}} G_n[\Pi_j(x_i)\epsilon_{i,j} \omega_i] + o_p(r_n^{-1}) \]

where the first and second equalities follow from a similar argument for Theorem SA-10.9 and the last line uses the rate of uniform convergence shown in Theorem SA-10.1. Note that we employ different rate restrictions given in the theorem: for $j = 1, 2, 3$, $h^{m+\vartheta} \log n \lesssim \sqrt{n} h^{d/2} = o(r_n^{-1})$, whereas for $j = 0$, $h^{m} \log n \lesssim \sqrt{n} h^{d/2} = o(r_n^{-1})$. Moreover, $\log n \lesssim \frac{n}{\sqrt{n} h^{d/2}} = o(r_n^{-1})$. Simply notice that $M_n \leq 1$ means $P(|M_n| > \vartheta_n) = o(1)$ for any $\vartheta_n \to \infty$. By Markov inequality, it immediately follows that $P^*(|M_n| > \vartheta_n) = o_p(1)$. Therefore, the above derivation still holds in $P$-probability if we replace $P$ by $P^*$.

Then we only need to construct strong approximation to the leading term in the above derivation. Since we assume $\omega$ is independent of the data and bounded, the moment condition in Theorem SA-10.5 is trivially satisfied. Repeat the argument for Theorem SA-10.5 conditional on the data.

95
Under the additional rate restriction, the wild bootstrap process is approximated by

\[ \frac{\gamma_{q,j}(x)'}{\Omega_j(x)^{1/2}} \left( \mathbb{E}_n[ \Pi_j(x_i) \Pi_j(x_i)' \varepsilon_i^2] \right)^{1/2} N_{K_j}, \]

where \( N_{K_j} \) is a \( K_j \)-dimensional standard normal vector, and the error rate is \( O_P(n^{-1/2} \log n/\sqrt{nhd}) = o_P(r_n^{-1}) \) by the rate restriction in the theorem. Then the desired result follows from the same argument for Theorem SA-10.6.

\[ \square \]

**SA-14.23 Proof of Theorem SA-10.11**

**Proof.** The proof is similar to that for Theorem SA-10.10. The differences are: (i) we employ Theorem SA-10.3 to replace \( \hat{\Omega}_j(x) \) by \( \Omega_j(x) \); (ii) when we construct the conditional Gaussian process that approximates the leading term in (SA-8) using the strategy given in Theorem SA-10.5, due to the stronger moment condition on \( \varepsilon_i \), the error rate is \( O_P((\log n)^2/\sqrt{nhd}) = o(r_n^{-1}) \); (iii) Theorem SA-10.3 leads to a sharper bound on \( \| (\mathbb{E}_n[ \Pi_j(x_i) \Pi_j(x_i)' \varepsilon_i^2])^{1/2} - \Sigma^{1/2} \| \) which is used in the unconditional coupling step.

\[ \square \]

**SA-14.24 Proof of Theorem SA-10.12**

**Proof.** In view of Theorem SA-10.6, there exists a sequence of constants \( \eta_n \) such that \( \eta_n = o(1) \) and

\[ \mathbb{P} \left( \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| - \sup_{x \in \mathcal{X}} |Z_j(x)| > \eta_n/r_n \right) = o(1). \]

Therefore, for any \( u \in \mathbb{R} \),

\[
\begin{align*}
\mathbb{P} & \left[ \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| \leq u \right] \\
& \leq \mathbb{P} \left[ \left\{ \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| \leq u \right\} \cap \left\{ \left| \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| - \sup_{x \in \mathcal{X}} |Z_j(x)| \right| \leq \eta_n/r_n \right\} \right] + \\
& \quad + \mathbb{P} \left[ \left| \sup_{x \in \mathcal{X}} |\hat{T}_j(x)| - \sup_{x \in \mathcal{X}} |Z_j(x)| \right| > \eta_n/r_n \right] \\
& \leq \mathbb{P} \left[ \sup_{x \in \mathcal{X}} |Z_j(x)| \leq u + \eta_n/r_n \right] + o(1) \\
& \leq \mathbb{P} \left[ \sup_{x \in \mathcal{X}} |Z_j(x)| \leq u \right] + Cr_n^{-1} \eta_n \mathbb{E} \left[ \left( \sup_{x \in \mathcal{X}} |Z_j(x)| \right)^{1/2} \right] + o(1)
\end{align*}
\]
where the last line follows from the Anti-Concentration Inequality given by [17]. By Gaussian maximal inequality, we have $\mathbb{E}[\sup_{x \in \mathcal{X}} |Z_j(x)|] \lesssim \sqrt{\log n}$. Since we assume $r_n = \sqrt{\log n}$, the two terms on the far right of the last line is $o(1)$ and do not depend on $u$. The reverse of the inequality can be established by a similar argument. Hence we conclude that

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

On the other hand, by Theorem SA-10.7 (or SA-10.8), Theorem SA-10.9 and Theorem SA-10.10 (or SA-10.11), under corresponding conditions $\tilde{Z}_j(\cdot)$, $\hat{Z}_j(\cdot)$ and $\hat{Z}_j^*(\cdot)$ are approximated by the same Gaussian process conditional on the data. Thus by the same argument given above, the desired result follows.

SA-14.25 Proof of Lemma SA-11.1

Proof. For the upper bound on the maximum eigenvalue, simply notice that all elements in $\mathbf{R}$ is bounded by some constant $C$ and the number of nonzeros on any row or column of $\mathbf{R}$ is bounded by some constant $L$. Then for any $\alpha \in \mathbb{R}^d$ such that $\|\alpha\| = 1$, $\alpha' \mathbf{R} \mathbf{R}' \alpha \leq L^2 C^2 \|\alpha\|^2 \lesssim 1$.

For the other side of the bound, since $\mathbf{R} \mathbf{R}'$ is a symmetric block Toeplitz matrix, Szegő’s theorem and its extensions state that the asymptotic behavior of Toeplitz or block Toeplitz matrices is characterized by the corresponding Fourier transformation of their entries. See [34] for more details. Specifically, $\mathbf{R} \mathbf{R}'$ is transformed into the following matrix

$$
\mathcal{F}_i(\omega) = \begin{bmatrix}
  r_{11}' r_{11} + 2 r_{11}' r_{21} \cos \omega & r_{11}' r_{12} + (r_{11}' r_{22} + r_{12}' r_{21}) \cos \omega & \cdots & r_{11}' r_{12} + (r_{11}' r_{21} + r_{12}' r_{22}) \cos \omega \\
  \vdots & \ddots & \ddots & \vdots \\
  r_{11}' r_{11} + (r_{11}' r_{21} + r_{11}' r_{22}) \cos \omega & \cdots & r_{11}' r_{12} + 2 r_{11}' r_{22} \cos \omega
\end{bmatrix},
$$

Using the representation of $\mathbf{R} \mathbf{R}'$ given in the discussion after the lemma, we can concisely write $\mathcal{F}_i(\omega) = \mathbf{A} + (\mathbf{B} + \mathbf{B}') \cos \omega$. By Equation (7) in [31], we have

$$
\lambda_{\min}(\mathbf{R} \mathbf{R}') \to \min_{\omega \in [0, 2\pi]} \lambda_{\min}(\mathcal{F}_i(\omega)) \quad \text{as} \quad \kappa \to \infty.
$$

The minimum of the minimum eigenvalue function of $\mathcal{F}_i(\omega)$ is attainable since each entry of
\( \mathcal{F}_i(\omega) \) is a linear function of \( \cos \omega \), and thus each coefficient of the corresponding characteristic polynomial is a continuous function of \( \cos \omega \). By Theorem 3.9.1 of [60] there exist \( \bar{i} \) continuous functions of \( \cos \omega \) such that they are the roots of the characteristic polynomial, and thus the minimum eigenvalue is a continuous function of \( \cos \omega \). In addition, since \( \mathbf{R} \mathbf{R}' \) is positive semi-definite, this function is nonnegative over \([-1, 1]\).

By construction, \( \mathbf{R} \mathbf{R}' \) is a real symmetric positive semi-definite matrix, and thus its eigenvalues are real and nonnegative. Moreover, given any fixed \( \kappa \), \( \mathbf{R} \mathbf{R}' \) is positive definite since the restrictions specified in \( \mathbf{R} \) are non-redundant. Therefore, it suffices to show that the limit of the minimum eigenvalue sequence is bounded away from zero. The original problem is transformed into showing that the minimum eigenvalue of a finite-dimensional matrix \( \mathcal{F}_i(\omega) \) is strictly positive for any \( \omega \in [0, 2\pi] \).

The next critical fact we employ is that the smallest eigenvalue as a function of a real symmetric matrix is concave (see Property 2.1 in [39]). In our case, each entry is a linear function of \( \cos \omega \), and thus \( \lambda_{\min}(\mathcal{F}_i(\omega)) \) is concave with respect to \( \cos \omega \). Therefore, the minimum of the smallest eigenvalue function can only be attained at two endpoints, i.e., when \( \cos \omega = 1 \) or \( \cos \omega = -1 \).

We start with the case in which \( (m - 1) \) continuity constraints are imposed at each knot, i.e., \( \bar{i} = m - 1 \). Since each knot is treated the same way, the restriction matrix \( \mathbf{R} \) can be fully characterized by \( \bar{i} \) row vectors. A typical restriction that the \( \varsigma \)th derivative \( (0 \leq \varsigma \leq \bar{i} - 1) \) is continuous at a knot can be represented by the following vector

\[
\left( \overbrace{\bar{P}_0^{(\varsigma)}(1), \ldots, \bar{P}_{m-1}^{(\varsigma)}(1)}^{\text{left interval}}, \overbrace{-\bar{P}_0^{(\varsigma)}(-1), \ldots, -\bar{P}_{m-1}^{(\varsigma)}(-1)}^{\text{right interval}} \right)
\]

where we omit all zero entries and \( \bar{P}_l^{(\varsigma)}(x) \), \( 0 \leq l \leq m - 1 \) denotes the \( \varsigma \)th derivative of the normalized Legendre polynomial of degree \( l \). Generally, Legendre polynomial \( P_l(x) \) can be written as

\[
P_l(x) = \frac{1}{2^l l!} \sum_{i=0}^{l} \binom{l}{i}^2 (x - 1)^{l-i} (x + 1)^i,
\]

and they have the following properties: for any \( l, l' \in \mathbb{Z}_+ \)

\[
P_l(1) = 1, \quad P_l(-x) = (-1)^l P_l(x), \quad \int_{-1}^{1} P_l(x) P_{l'}(x) \, dx = \frac{2}{2l + 1} \delta_{ll'}
\]

98
where $\delta_{ll'}$ is the Kronecker delta. Therefore, $\bar{P}_l(x) = \frac{\sqrt{2l+1}}{\sqrt{2}} P_l(x)$. Using these formulas, 

$$P_l^{(c)}(1) = 2^{-\varsigma!} \binom{l}{\varsigma} \frac{(l+\varsigma)!}{\varsigma!} = 2^{-\varsigma!} \frac{(l+\varsigma)(l+\varsigma-1) \cdots (l-\varsigma+1)}{\varsigma!}.$$  

Therefore, 

$$\bar{P}_l^{(c)}(1) = \frac{\sqrt{2l+1}}{\sqrt{2}} 2^{-\varsigma!} \binom{l}{\varsigma} \frac{(l+\varsigma)(l+\varsigma-1) \cdots (l-\varsigma+1)}{\varsigma!}.$$  

In addition, since Legendre polynomials are symmetric or antisymmetric, we have 

$$\bar{P}_l^{(c)}(-1) = (-1)^{l+\varsigma} \bar{P}_l^{(c)}(1)$$  

Thus we obtain an explicit expression for $R$. 

Next, $RR'$ can be fully characterized by two matrices $A$ and $B$ given in (SA-3). In what follows we use $A[\varsigma, \ell]$ to denote the $(\varsigma, \ell)$th element of $A$. The same notation is used for $B$ and $\mathcal{F}_l(\omega)$. If we arrange restrictions by increasing $\varsigma$ (here we allow row and column indices to start from 0), then we have 

$$A[\varsigma, \varsigma] = \sum_{u=\varsigma}^{\ell} (2u+1) 2^{-2\varsigma} \left[ \varsigma! \binom{u}{\varsigma} \left( \binom{u+\varsigma}{\varsigma} \right)^2 \right],$$  

and for $\varsigma > \ell$ 

$$A[\varsigma, \ell] = \begin{cases} 
0 & \varsigma + \ell \text{ is odd} \\
\sum_{u=\varsigma}^{\ell} (2u+1) 2^{-\varsigma-\ell} \varsigma! \binom{u+\varsigma}{\varsigma} \binom{u+\ell}{\ell} & \varsigma + \ell \text{ is even} 
\end{cases}. $$  

$B$ can be expressed explicitly as well: 

$$B[\varsigma, \varsigma] = \sum_{u=\varsigma}^{\ell} (-1)^{u+\varsigma+1} \frac{2u+1}{2} 2^{-2\varsigma} \left[ \varsigma! \binom{u}{\varsigma} \left( \binom{u+\varsigma}{\varsigma} \right)^2 \right]$$  

and 

$$B[\varsigma, \ell] = \begin{cases} 
\sum_{u=\varsigma}^{\ell} (-1)^{u+\varsigma+1} \frac{2u+1}{2} 2^{-\varsigma-\ell} \varsigma! \binom{u+\varsigma}{\varsigma} \binom{u+\ell}{\ell} & \varsigma > \ell \\
(-1)^{\ell+\varsigma} B[\ell, \varsigma] & \varsigma < \ell 
\end{cases}. $$
Therefore, \((\varsigma, \ell)\)th element of \(\mathcal{F}_\ell(\omega)\) is

\[
\mathcal{F}_\ell(\omega)[\varsigma, \ell] = \begin{cases} 
0 & \ell + \varsigma \text{ is odd} \\
A[\varsigma, \ell] + 2B[\varsigma, \ell] \cos \omega & \ell + \varsigma \text{ is even}
\end{cases}
\]

When \(\ell + \varsigma\) is even, the corresponding entry of \(\mathcal{F}_\ell(\omega)\) is nonzero. In addition, the summands in \(A[\varsigma, \ell]\) and \(2B[\varsigma, \ell]\) are the same in terms of absolute values and only differ in signs. Consider the case when \(\cos \omega = 1\) and \(\varsigma \geq \ell\). There are several cases:

(i) \(\bar{i}\) is even, \(\varsigma\) is even

\[
\mathcal{F}_\ell(\omega)[\varsigma, \ell] = \sum_{u=\varsigma/2}^{(i-2)/2} \left(2(2u+1)+1\right)2^{\varsigma-\ell+1}!\ell!\left(\begin{array}{c} 2u+1 \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u+1 + \varsigma \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u+1 \\ \ell \end{array}\right)\left(\begin{array}{c} 2u+1 + \ell \\ \ell \end{array}\right);
\]

(ii) \(\bar{i}\) is even, \(\varsigma\) is odd

\[
\mathcal{F}_\ell(\omega)[\varsigma, \ell] = \sum_{u=\varsigma+1/2}^{i/2} \left(2(2u)+1\right)2^{\varsigma-\ell+1}!\ell!\left(\begin{array}{c} 2u \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u + \varsigma \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u \\ \ell \end{array}\right)\left(\begin{array}{c} 2u + \ell \\ \ell \end{array}\right);
\]

(iii) \(\bar{i}\) is odd, \(\varsigma\) is even

\[
\mathcal{F}_\ell(\omega)[\varsigma, \ell] = \sum_{u=\varsigma/2}^{(i-1)/2} \left(2(2u+1)+1\right)2^{\varsigma-\ell+1}!\ell!\left(\begin{array}{c} 2u+1 \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u+1 + \varsigma \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u+1 \\ \ell \end{array}\right)\left(\begin{array}{c} 2u+1 + \ell \\ \ell \end{array}\right);
\]

(iv) \(\bar{i}\) is odd, \(\varsigma\) is odd

\[
\mathcal{F}_\ell(\omega)[\varsigma, \ell] = \sum_{u=\varsigma+1/2}^{(i-1)/2} \left(2(2u)+1\right)2^{\varsigma-\ell+1}!\ell!\left(\begin{array}{c} 2u \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u + \varsigma \\ \varsigma \end{array}\right)\left(\begin{array}{c} 2u \\ \ell \end{array}\right)\left(\begin{array}{c} 2u + \ell \\ \ell \end{array}\right).
\]
When $\bar{\iota}$ is odd, $\mathcal{F}_{\bar{\iota}}(\omega)$ can be written as a Gram matrix $\mathcal{F}_{\bar{\iota}}(\omega) = GG'$ where

$$
G = \begin{bmatrix}
\bar{P}_1(1) & 0 & \bar{P}_3(1) & \cdots & 0 & \bar{P}_{m-1}(1) \\
0 & \bar{P}_2^{(1)}(1) & 0 & \cdots & \bar{P}_m^{(1)}(1) & 0 \\
0 & 0 & \bar{P}_3^{(2)}(1) & \cdots & 0 & \bar{P}_m^{(2)}(1) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & \bar{P}_m^{(m-2)}(1)
\end{bmatrix}.
$$

Clearly it is a row echelon matrix and has full row rank. Thus, the minimum eigenvalue of $\mathcal{F}_{\bar{\iota}}(\omega)$ is strictly positive.

When $\bar{\iota}$ is even, $\mathcal{F}_{\bar{\iota}}(\omega)$ can be written as $GG'$ where

$$
G = \begin{bmatrix}
\bar{P}_1(1) & 0 & \bar{P}_3(1) & \cdots & \bar{P}_{m-2}(1) & 0 \\
0 & \bar{P}_2^{(1)}(1) & 0 & \cdots & 0 & \bar{P}_m^{(1)}(1) \\
0 & 0 & \bar{P}_3^{(2)}(1) & \cdots & \bar{P}_{m-2}^{(2)}(1) & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & \bar{P}_m^{(m-2)}(1)
\end{bmatrix}.
$$

The case of $\cos \omega = -1$ can be proved the same way. This suffices to show that the minimum eigenvalue of $\mathcal{F}_{\bar{\iota}}(\omega)$ as a function of $\cos \omega$ is strictly positive at two endpoints 1 and -1, thus completing the proof for $\bar{\iota} = m - 1$.

To complete the proof of the lemma, it remains to extend this result to the case in which fewer constraints are imposed. Compared with the case when $(m - 1)$ constraints are imposed, some rows in the bigger restriction matrix are removed, and accordingly $RR'$ is a principle submatrix of the original one. By Cauchy Interlacing Theorem, the smallest eigenvalue of the principle submatrix must be no less than the smallest eigenvalue of the original matrix. Combining this fact with the results proved for $\bar{\iota} = m - 1$, we have the minimum eigenvalue of $RR'$ uniformly bounded away from 0 when fewer restrictions are imposed, and then the proof is complete. \qed
SA-15  Proofs: Main Paper

This section gives proofs of the theoretical results in the main paper. For precise statements see the text there.

SA-15.1  Proof of Lemma 2.1

Proof. The result follows from the proof of Lemma SA-8.3. □

SA-15.2  Proof of Theorem 3.1

Proof. The result follows from the proof of Theorem SA-10.6. □

SA-15.3  Proof of Theorem 3.2

Proof. The result follows from the proof of Theorem SA-10.7 and SA-10.8. □

SA-15.4  Proof of Theorem 3.3

Proof. The result follows from the proof of Theorem SA-10.10 and SA-10.11. □

SA-15.5  Proof of Theorem 3.4

Proof. The result follows from the proof of Theorem SA-10.12. □

SA-15.6  Proof of Theorem 4.1

Proof. First, using Lemma SA-8.2 and SA-8.4, the integrated conditional variance can be written as

\[
\int_X \nabla [\widehat{\partial q} \mu_0(x) | X] w(x) \, dx = \frac{1}{n} \text{trace} \left[ \Sigma_0 \int X \gamma_{q,0}(x) \gamma_{q,0}(x) w(x) \, dx \right] + o_P(n^{-1}h^{-d-2|q|}).
\]
Moreover,

\[
\frac{1}{n} \text{trace} \left[ \Sigma_0 \int_X \gamma_{q,0}(x) \gamma_{q,0}(x)' w(x) \, dx \right] \\
\leq \frac{1}{n} \lambda_{\max} \left( Q_m^{-1}\mathbb{E}[p(x_i)p(x_i)'\sigma(x_i)^2] Q_m^{-1} \right) \text{trace} \left[ \int_X \partial^q p(x) \partial^q p(x)' w(x) \, dx \right] \\
\leq \frac{h^{-d}}{n} \text{trace} \left[ \int_X \partial^q p(x) \partial^q p(x)' w(x) \, dx \right] \\
\leq h^{-d-2[q]} / n
\]

where the second line follows from Trace Inequality, the third line from Lemma SA-8.1 and SA-8.2, and the last line from the continuity of \( w(\cdot) \) and Assumption SA-3. For the other side of the bound, since \( \sigma(\cdot)^2 \) is bounded away from zero, we have \( \lambda_{\min}(Q_m^{-1}\mathbb{E}[p(x_i)p(x_i)'\sigma(x_i)^2] Q_m^{-1}) \gtrsim h^{-d} \), and then the lower bound follows from Trace inequality, Assumption SA-3 and the condition that \( w(\cdot) \) is bounded away from 0.

For the integrated squared bias, we have

\[
\begin{align*}
\int_X \left( \mathbb{E}[\partial^q \mu_0(x) | X] - \partial^q \mu(x) \right)^2 w(x) \, dx &= \int_X \left( \mathcal{B}_{m,q}(x) - \partial^q p(x)' Q_m^{-1} \mathbb{E}[p(x_i) \mathcal{B}_{m,0}(x_i)] \right)^2 w(x) \, dx + o_F(h^{2m-2[q]}) \\
&= \int_X \mathcal{B}_{m,q}(x)^2 w(x) \, dx + \int_X \left( \partial^q p(x)' Q_m^{-1} \mathbb{E}[p(x_i) \mathcal{B}_{m,0}(x_i)] \right)^2 w(x) \, dx \\
&\quad - 2 \int_X \mathcal{B}_{m,q}(x) \partial^q p(x)' Q_m^{-1} \mathbb{E}[p(x_i) \mathcal{B}_{m,0}(x_i)] w(x) \, dx + o_F(h^{2m-2[q]}) \\
&= : B_1 + B_2 - 2B_3 + o_F(h^{2m-2[q]}). \quad \text{(SA-1)}
\end{align*}
\]

First, let \( h_\delta \) be the diameter of \( \delta \) and \( t_\delta^* \) be an arbitrary point in \( \delta \). Then

\[
\begin{align*}
B_1 &= \sum_{u_1, u_2 \in \Lambda_m} \int_X \left[ \partial^{u_1} \mu(x) \partial^{u_2} \mu(x) h_x^{2m-2[q]} B_{u_1,q}(x) B_{u_2,q}(x) \right] w(x) \, dx \\
&= \sum_{u_1, u_2 \in \Lambda_m} \sum_{\delta \in \Delta} \int_{h_\delta^{2m-2[q]}} \left[ \partial^{u_1} \mu(t_\delta^*) \partial^{u_2} \mu(t_\delta^*) B_{u_1,q}(x) B_{u_2,q}(x) \right] w(t_\delta^*) \, dx + o(h^{2m-2[q]}) \\
&= \sum_{u_1, u_2 \in \Lambda_m} \sum_{\delta \in \Delta} \left\{ h_\delta^{2m-2[q]} \partial^{u_1} \mu(t_\delta^*) \partial^{u_2} \mu(t_\delta^*) w(t_\delta^*) \int_{h_\delta} B_{u_1,q}(x) B_{u_2,q}(x) \, dx \right\} + o(h^{2m-2[q]}) \\
&\lesssim h^{2m-2[q]}
\end{align*}
\]
where the third line holds by the continuity of $\partial^{u_1}\mu$, $\partial^{u_2}\mu$ and $w(x)$ and the last by Assumption SA-2 and SA-4.

Second,

$$B_2 = \text{trace} \left[ Q_m^{-1}E[p(x_i)\mathcal{B}_{m,0}(x_i)]E[p(x_i)\mathcal{B}_{m,0}(x_i)]Q_m^{-1} \int_X \partial^q p(x) \partial^q p(x)' w(x) dx \right]$$

$$\leq \lambda_{\text{max}} \left( Q_m^{-1} \int_X \partial^q p(x) \partial^q p(x)' w(x) dx Q_m^{-1} \right) \text{trace} \left[ E[p(x_i)\mathcal{B}_{m,0}(x_i)]E[p(x_i)\mathcal{B}_{m,0}(x_i)] \right]$$

$$\lesssim h^{-d-2\lceil q \rceil} \text{trace} \left[ E[p(x_i)\mathcal{B}_{m,0}(x_i)]E[p(x_i)\mathcal{B}_{m,0}(x_i)] \right]$$

where the second line holds by Trace Inequality, the third by Lemma SA-8.1 and Assumption SA-3 and the last by Assumption SA-3 and SA-4.

Finally,

$$|B_3| = \left| \int_X \mathcal{B}_{m,q}(x) \partial^q p(x)' w(x) dx \right| Q_m^{-1}E[p(x_i)\mathcal{B}_{m,0}(x_i)]$$

$$\leq \left\| \int_X \mathcal{B}_{m,q}(x) \partial^q p(x)' w(x) dx \right\|_\infty \left\| Q_m^{-1}E[p(x_i)\mathcal{B}_{m,0}(x_i)] \right\|_\infty$$

$$\lesssim K h^{m-2\lceil q \rceil} h^{-d} h^{m+d}$$

$$\lesssim h^{2m-2\lceil q \rceil}$$

where the third line follows from Assumption SA-3 and Lemma SA-8.1. Then the proof is complete.

□

**SA-15.7 Proof of Lemma A.1**

*Proof.* (a) Assumption SA-3(a) directly follows from the construction of tensor-product $B$-splines. Assumption SA-3(b) follows from [54, Theorem 12.5]. To prove Assumption SA-3(c), notice that given univariate $B$-splines $\{p_l(x_i)\}_{i=1}^{K_l}$, there exists a universal constant $C > 0$ such that for any $\varsigma_l \leq m$, $\delta \in \Delta$,

$$\left\| \frac{d^{\varsigma_l} p_l(x_i)}{dx_i^{\varsigma_l}} \right\|_{L_\infty(\text{clo}(\delta))} \lesssim h^{-\varsigma_l}.$$
Since there are only a fixed number of nonzero elements in \( p \), we have for \( \varsigma \leq m \),

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

To derive the other side of the bound, notice that the proof of \([66, \text{Lemma 5.4}]\) shows that for a univariate \( B \)-spline basis \( \tilde{p}_\ell(x) := (\tilde{p}_1(x), \cdots, \tilde{p}_{K_\ell}(x))' \), for any \( \varsigma_\ell \leq m - 1 \), \( x_\ell \in X_\ell \),

\[
\left\| \frac{d^\varsigma \tilde{p}_\ell(x_\ell)}{dx_\ell^\varsigma} \right\| \gtrsim h^{-\varsigma_\ell}.
\]

Since for any \( x_\ell \), there are only \( m \) nonzero elements in \( \tilde{p}_\ell(x_\ell) \), this suffices to show that for any \( x_\ell \in X_\ell \), there exists some \( \tilde{q}_\ell(x_\ell) \) such that

\[
\left| \frac{d^\varsigma \tilde{q}_\ell(x_\ell)}{dx_\ell^\varsigma} \right| \gtrsim h^{-\varsigma_\ell}.
\]

Then the direct lower bound directly follows from the construction of tensor-product \( B \)-splines.

(b): The proof of orthogonality between the constructed leading error and \( B \)-splines can be found in \([2]\). We first consider \( \varsigma = 0 \). Noticing that \( B_{m,0}(x) = -\sum_{\ell=1}^d \frac{\partial^m \mu(t_\ell^L)}{\partial x_\ell^m} \frac{h_{t_{\ell,\ell}}^m}{m!} B_m \left( \frac{x_\ell - t_{\ell,\ell}}{b_{\ell,\ell}} \right) + O(h^{m+\phi}) \) for \( x \in \delta_{l_1 \cdots l_d} \), where \( B_m(\cdot) \) is the \( m \)th Bernoulli polynomial, we only need to work with the first term on the RHS, denoted by \( \tilde{B}_m(x) \). By construction, \( \tilde{B}_m(x) \) is continuous on the interior of each subrectangle \( \delta_{l_1 \cdots l_d} \), and the discontinuity only takes place at boundaries of subrectangles. Let \( J_0 \) denote the magnitude of jump of \( \tilde{B}_m(x) \). By Assumption SA-1, \( J_0 \) is also the jump of \( \tilde{\mu} := \mu + \tilde{B}_m \). We first check the magnitude of the jump as \([2]\) did in their proof. We introduce the following notation:

i. \( \tau := (\tau_1, \cdots, \tau_d) \) is a point on the boundary of a generic rectangle \( \delta_{l_1 \cdots l_d} \);

ii. \( \tau^- := (\tau^-_1, \cdots, \tau^-_d) \) and \( \tau^+ := (\tau^+_1, \cdots, \tau^+_d) \) are two points close to \( \tau \) but belong to two different subrectangles \( \delta^-_{l_1 \cdots l_d} := \{x : t^-_{\ell,\ell} \leq x_\ell < t^-_{\ell,\ell+1}\} \) and \( \delta^+_{l_1 \cdots l_d} := \{x : t^+_{\ell,\ell} \leq x_\ell < t^+_{\ell,\ell+1}\} \);

iii. \( t_{L}^- \) and \( t_{L}^+ \) are the starting points of \( \delta^-_{l_1 \cdots l_d} \) and \( \delta^+_{l_1 \cdots l_d} \);
iv. \((b_{1,-}, \cdots, b_{d,-})\) and \((b_{1,+}, \cdots, b_{d,+})\) are the corresponding mesh widths of \(\delta_{i_1 \ldots i_d}^-\) and \(\delta_{i_1 \ldots i_d}^+\);

v. \(\Xi := \{ \ell: \tau_{\ell}^- - \tau_{\ell}^+ \text{ and } \tau_{\ell}^+ - \tau_{\ell}^- \text{ differ in signs} \}\).

In words, the index set \(\Xi\) indicates the directions in which we cross boundaries when we move from \(\tau_{\ell}^-\) to \(\tau_{\ell}^+\). To further simplify notation, we write \(\tilde{\mu}(\tau^-) := \lim_{x \to \tau, \mu \in \delta_{i_1 \ldots i_d}^-} \tilde{\mu}(x)\) and \(\tilde{\mu}(\tau^+) := \lim_{x \to \tau, \mu \in \delta_{i_1 \ldots i_d}^+} \tilde{\mu}(x)\). Then we have

\[
J_0 = |\tilde{\mu}(\tau^+) - \tilde{\mu}(\tau^-)| = \left| \tilde{\mathcal{B}}_m(\tau^+) - \tilde{\mathcal{B}}_m(\tau^-) \right|
\]

\[
= \sum_{\ell \in \Xi} \left( \frac{B_m(0)}{|m!|} \left| \frac{\partial^m \mu(t^+)_{\ell}}{\partial x_{\ell}^m} b_{\ell,+} - \frac{\partial^m g(t^-_{\ell})}{\partial x_{\ell}^m} b^m_{\ell,-} \right| \right)
\]

\[
= \sum_{\ell \in \Xi} \left( \frac{B_m(0)}{|m!|} \left( \frac{\partial^m g(t^+)_{\ell}}{\partial x_{\ell}^m} - \frac{\partial^m g(t^-_{\ell})}{\partial x_{\ell}^m} \right) b_{\ell,+} + \frac{\partial^m g(t^-_{\ell})}{\partial x_{\ell}^m} (b_{\ell,+} - b_{\ell,-}) \right)
\]

\[
\leq \sum_{\ell \in \Xi} \left( \frac{B_m(0)}{|m!|} \left[ O(h^{m+\varrho}) + Ch^{m-1}|b_{\ell,+} - b_{\ell,-}| \right] \right)
\]

\[
\leq \sum_{\ell \in \Xi} \left( \frac{B_m(0)}{|m!|} \left[ O(h^{m+\varrho}) + Ch^{m-1}O(h^{1+\varrho}) \right] \right)
\]

where the fourth line follows from Assumption \(\text{SA-1}\) and the last line follows from the stronger quasi-uniformity condition given in Lemma A.1. This suffices to show that \(J_0\) is \(O(h^{m+\varrho})\).

Then we mimic the proof strategy used in [54, Theorem 12.7]. By [54, Theorem 12.6], we can construct a bounded linear operator \(\mathcal{L}[\cdot]\) mapping \(L_1(\mathcal{X})\) onto \(S_{\Delta,m}\) with \(\mathcal{L}[s] = s\) for all \(s \in S_{\Delta,m}\). Specifically, \(\mathcal{L}[\cdot]\) is defined as

\[
\mathcal{L}[\mu](x) := \sum_{l_1=1}^{K_1} \cdots \sum_{l_d=1}^{K_d} (\psi_{l_1 \ldots l_d} \mu)p_{l_1 \ldots l_d}(x)
\]

where \(\{\psi_{l_1 \ldots l_d}\}_{l_1=1, \ldots, l_d=1}^{K_1, \ldots, K_d}\) is a dual basis defined as [54, Equation (12.24)]. By multi-dimensional Taylor expansion, there exists a polynomial \(\varphi_{l_1 \ldots l_d}\) such that \(\|\tilde{\mu} - \varphi_{l_1 \ldots l_d}\|_{L_\infty(\delta_{l_1 \ldots l_d})} \lesssim h^{m+\varrho}\), and the degree of \(\varphi_{l_1 \ldots l_d}\) is no greater than \(m - 1\). Since \(\mathcal{L}\) reproduces polynomials, we have

\[
\|\tilde{\mu} - \mathcal{L}[\tilde{\mu}]\|_{L_\infty(\delta_{l_1 \ldots l_d})} \leq \|\tilde{\mu} - \varphi_{l_1 \ldots l_d}\|_{L_\infty(\delta_{l_1 \ldots l_d})} + \|\mathcal{L}[\tilde{\mu} - \varphi_{l_1 \ldots l_d}]\|_{L_\infty(\delta_{l_1 \ldots l_d})}
\]

\[
\leq C\|\tilde{\mu} - \varphi_{l_1 \ldots l_d}\|_{L_\infty(\delta_{l_1 \ldots l_d})} \lesssim h^{m+\varrho}.
\]
Taking account of the jump of $\bar{\mu}$ along boundaries, the approximation error of $L[\bar{\mu}]$ is still $O(h^{m+e})$. Evaluate the $L_\infty$ norm on all subrectangles and then we conclude that there exists some $s^* \in S_{\Delta,m}$ such that $\|\mu + \mathcal{R}_m - s^*\|_{L_\infty(X)} \lesssim h^{m+e}$.

For other $\varsigma$, we only need to show that the desired result holds for $s^* = L[\bar{\mu}]$. By construction of $L$,

$$\|\partial^s (L[\bar{\mu}])\| \leq \sum_{l_1=1}^{m+\kappa_1} \cdots \sum_{l_d=1}^{m+\kappa_d} |\psi_{l_1\ldots l_d}\bar{\mu}| \|\partial^s p_{l_1\ldots l_d}(x)\| \leq Ch^{-|\varsigma|}\|ar{\mu}\|_{L_\infty(\delta_{1\ldots d})} \tag{SA-2}$$

where the last line follows from [54, Theorem 12.5]. Then we have

$$\|\partial^s \mu + \partial^s \mathcal{R}_m - \partial^s (L[\bar{\mu}])\|_{L_\infty(\delta_{1\ldots d})} \leq \|\partial^s \mu + \partial^s \mathcal{R}_m - \partial^s \varphi_{l_1\ldots l_d}\|_{L_\infty(\delta_{1\ldots d})} + \|\partial^s (L[\bar{\mu}] - \varphi_{l_1\ldots l_d})\|_{L_\infty(\delta_{1\ldots d})} \leq O(h^{m+e-|\varsigma|}) + Ch^{-|\varsigma|}\|ar{\mu} - \varphi_{l_1\ldots l_d}\|_{L_\infty(\delta_{1\ldots d})} \lesssim h^{m+e-|\varsigma|}$$

where the second inequality follows from Taylor expansion and Equation (SA-2). Moreover, By the similar argument for $J_0$, the jump of $\partial^s \mathcal{R}_m$ is $O(h^{m+e-|\varsigma|})$.

(c): By construction of $\tilde{p}$, $\rho = 1$. It follows from the same argument in part (a) and (b) that $\tilde{p}$ satisfies Assumption SA-3 and SA-4. Finally, by definition of tensor-product splines, both $p$ and $\tilde{p}$ reproduce polynomials of degree no greater than $m - 1$. Then the proof is complete. 

SA-15.8 Proof of Lemma A.2

Proof. (a): Assumption SA-3(a) directly follows from the fact that the father wavelet is compactly supported and $\{\phi_{s\ell}\}$ is generated by translation and dilation. Assumption SA-1(b) follows from the fact that $\{\phi_{s\ell}\}$ is an orthonormal basis with respect to the Lebesgue measure. For Assumption SA-3(c), notice that

$$\frac{d^s \phi(2^s x_{\ell} - l_{\ell})}{dx^s_{\ell}} = 2^{s_{\ell}} \frac{d^s \phi(z)}{dz^s} \bigg|_{z=2^s x_{\ell} - l_{\ell}} = b^{-s_{\ell}} \frac{d^s \phi(z)}{dz^s} \bigg|_{z=2^s x_{\ell} - l_{\ell}}.$$
Since the wavelet basis reproduces polynomials of degree no greater than \( m - 1 \) and \( \phi \) is assumed to have \( q + 1 \) continuous derivatives, the desired bounds follow.

(b): We follow the same strategy used in [56], but extend their proof to the multidimensional case. First, we denote by \( V^\ell_s \) the closure of the level-\( s \) subspace spanned by \( \{ \phi_{sl}(x_\ell) \} \) and \( W^\ell_s \) the orthogonal complement of \( V^\ell_s \) in \( V_{s+1}^\ell \). Then we write \( V_s := \bigotimes_{\ell=1}^d V^\ell_s \) for the space spanned by the tensor-product level-\( s \) father wavelets, and \( W_s \) as the orthogonal complement of \( V_s \) in \( V_{s+1} \). We use the following fact:

\[
W_s = \bigoplus_{i=1}^{2^d-1} W_{s,i}
\]

where \( \oplus \) denotes “direct sum”, and each \( W_{s,i} \) takes the following form:

\[
W_{s,i} = \bigotimes_{\ell=1}^d Z^\ell_{s,i}.
\]

Each \( Z^\ell_{s,i} \) is either \( V^\ell_s \) or \( W^\ell_s \), but \( \{ Z^\ell_{s,i} \}_{i=1}^{d} \) cannot be identical to \( \{ V^\ell_s \}_{i=1}^{d} \). There are in total \((2^d - 1)\) such subspaces. Accordingly, a typical element in a basis vector of \( W_s \) can be written as

\[
\bar{\psi}_{s \alpha l}(x) = \prod_{\ell=1}^d [\alpha_\ell \phi_{sl}(x_\ell) + (1 - \alpha_\ell) \psi_{sl}(x_\ell)]
\]

where \( l = (l_1, \ldots, l_d) \) and \( \alpha_\ell = 0 \) or \( 1 \), but \( \alpha = (\alpha_1, \ldots, \alpha_d) \neq (1, \ldots, 1) \). Then it directly follows from the properties of wavelet basis that for \( \bar{\psi}_{s \alpha l} \), \( s \geq m \),

\[
\langle x^\varsigma, \bar{\psi}_{s \alpha l}(x) \rangle := \int_\mathcal{X} x^\varsigma \bar{\psi}_{s \alpha l}(x) dx = 0, \quad \text{for } \varsigma \text{ such that } [\varsigma] \leq m, \quad \text{and } \varsigma_\ell \neq m \forall j. \quad \text{(SA-3)}
\]

Denote by \( L_s[ \cdot ] \) the orthogonal projection operator onto \( W_s \). Then the approximation error of the tensor-product wavelet space \( V_{s,n} \) can be written as

\[
\sum_{s=s_n}^{\infty} L_s[\mu](x) = \sum_{s=s_n}^{\infty} \sum_{\alpha} \sum_{l=1}^d \langle \mu(\tilde{x}), \bar{\psi}_{s \alpha l}(\tilde{x}) \rangle \bar{\psi}_{s \alpha l}(x)
\]

\[
= \sum_{s=s_n}^{\infty} \sum_{\alpha} \sum_{l=1}^d \left( \sum_{[\varsigma] \leq m} \partial^\varsigma \mu(x) \frac{(\tilde{x} - x)^\varsigma}{\varsigma!} + \vartheta_n(\tilde{x}, x, \bar{\psi}_{s \alpha l}(\tilde{x})) \right) \bar{\psi}_{s \alpha l}(x)
\]

where \( \vartheta_n(\tilde{x}, x) \lesssim \| \tilde{x} - x \|^{n+\rho} \), and the inner product in the above equations are taken with respect to \( \tilde{x} \) in terms of Lebesgue measure. The index sets where \( \alpha \) and \( l \) live are described in Appendix
A and the proof, and omitted in the above derivation for simplicity. It follows from Assumption SA-1 and Assumption SA-3 that

\[
\sup_{x \in X} \left| \sum_{s=s_n}^{\infty} \sum_{\alpha} \sum_{l} \left\langle \vartheta_n(\tilde{x}, x), \tilde{\psi}_{s\alpha}(\tilde{x}) \right\rangle \tilde{\psi}_{s\alpha}(x) \right| \\
= \sup_{x \in X} \left| \sum_{s=s_n}^{\infty} \left( \frac{b}{2^{s-s_n}} \right)^{m+\varrho} \sum_{\alpha} \sum_{l} \left\langle \vartheta_n(\tilde{x}, x)2^{s(m+\varrho)}, \tilde{\psi}_{s\alpha}(\tilde{x}) \right\rangle \tilde{\psi}_{s\alpha}(x) \right| \\
\lesssim b^{m+\varrho}.
\]

Recall that \( b = 2^{-s_n} \).

Regarding the leading terms

\[
\sum_{s=s_n}^{\infty} \sum_{\alpha} \sum_{l} \left\langle \left( \sum_{s=0}^{m} \partial^s \mu(x) \frac{(\tilde{x} - x)\varsigma}{\varsigma!}, \tilde{\psi}_{s\alpha}(\tilde{x}) \right) \tilde{\psi}_{s\alpha}(x), \right\rangle,
\]

it is clear that the coefficients of the wavelet basis can be viewed as a linear combination of the inner products of monomials and the mother wavelets themselves, and thus by Equation (SA-3) the leading error is of order \( b^m \) and can be characterized as

\[
\mathcal{B}_{m,0}(x) = - \sum_{u \in \Lambda_m} \frac{b^m}{u!} \partial^u \mu(x) B^u_{0,0}(x/b).
\]

\( B^u_{0,0} \) is referred to as “monowavelet” in [56]. Here we extend it to the multidimensional case. Specifically, define a mapping

\[
\varphi : \Lambda_m \to \{1, \ldots, d\} \\
\quad u \mapsto \ell
\]

such that \( \varphi(u) \)th element of \( u \) is nonzero. We denote \( 1_{-\ell} := (l_{1}, \ldots, l_{\ell-1}, l_{\ell+1}, \ldots, l_d) \) and

\[
\mathcal{L}_{s}^{-\ell} := \left\{ 1_{-\ell} : l_{\ell'} \in \mathcal{L}_s, j' = \{1, \cdots, d\} \setminus \{\ell\} \right\}.
\]
Then define

\[ \varpi_{u,s}(x) = \sum_{l \in \mathcal{L}_s} \sum_{\varphi \in \mathcal{L}_s} c_m \psi(2^s x \varphi(u) - l \varphi(u)) \prod_{\ell = 1}^d \phi(2^s x_\ell - l_\ell) \]

where \( c_m := \int_0^1 x^m \psi(x) \, dx \). Then \( B_{u,0}^\psi(\cdot) \) can be expressed as

\[ B_{u,0}^\psi(x) = \sum_{s=0}^\infty 2^{-sm} \varpi_{u,s}(x) = \sum_{s=0}^\infty \xi_{u,s}(x). \quad (SA-4) \]

Moreover, since the series in Eq. \((SA-4)\) converges uniformly and for \( s \geq s_n \), \( \varpi_{u,s}(x) \) is orthogonal to the tensor-product wavelet basis \( \mathbf{p} \) with respect to the Lebesgue measure, it follows from Dominated Convergence Theorem that the approximate orthogonality condition holds.

For other \( \varsigma \), let

\[ \mathcal{B}_{m,\varsigma}(x) = - \sum_{u \in \Lambda_m} b^{m-\varsigma} \frac{\partial^u \mu(x) B_{u,\varsigma}^\varpi(x/h)}{u!} \]

where \( B_{u,\varsigma}^\varpi(x) = \partial^\varsigma B_{u,0}^\varpi(x) \). By assumption that for \( \varsigma \) such that \([\varsigma] \leq \varsigma\), \( \partial^\varsigma \phi \) and \( \partial^\varsigma \psi \) are continuously differentiable, we have \( \sum_{s=0}^\infty 2^{-sm} \partial^\varsigma \varpi_{u,s}(x) \) converge uniformly, and hence we can interchange the differentiation and infinite summation. Therefore, \( B_{u,\varsigma}^\varpi(\cdot) \) is well defined and continuously differentiable. Then the lipschitz condition on \( B_{u,\varsigma}^\varpi(\cdot) \) in Assumption \text{SA-4} holds.

Let \( s^* \) be the orthogonal projection of \( \mu \) onto \( \mathcal{V}_{s_n} \). To complete the proof of part (b), it suffices to show \( \| \partial^\varsigma \mu - \partial^\varsigma s^* + \mathcal{B}_{m,\varsigma} \|_{L_\infty(X)} \lesssim b^{m+\rho-\varsigma} \). Note that for a given \( s_n \), we have

\[
\sum_{s=s_n}^\infty \sum_{\varsigma=1}^{[\varsigma]} \langle \mu(x), \tilde{\psi}_{s\varsigma\alpha}(x) \rangle \partial^\varsigma \tilde{\psi}_{s\varsigma\alpha}(x)
\]

\[
= \sum_{s=s_n}^\infty \sum_{\varsigma=1}^{[\varsigma]} \left( \sum_{u \leq m} \partial^u \mu(x) \frac{(x-x)^u}{u!} + \theta_n(x, x), \tilde{\psi}_{s\varsigma\alpha}(x) \right) \partial^\varsigma \tilde{\psi}_{s\varsigma\alpha}(x)
\]

\[
= b^{m-\varsigma} \sum_{s=s_n}^{[\varsigma]} 2^{s_2(s-s_n)} \sum_{s=s_n}^{[\varsigma]} \sum_{\varsigma=1}^{2^s} 2^{-sd} \sum_{u \leq m} \partial^u \mu(x) \frac{(x-x)^u}{u!} + \theta_n(x, x), \tilde{\psi}_{s\varsigma\alpha}(x)
\]

By changing variables, the vanishing moments of the wavelet function and the fact that geometric series converges, the last line uniformly converges to the \( \varsigma \)th derivative of the approximation error.
of $V_{sn}$, $\mathcal{B}_{m,\varsigma}(\cdot)$ is the leading error and the remainder behaves like $O(b^{m+e[-\varsigma]})$.

(c): By construction of $\hat{p}$, $\rho = 1$. It follows from the same argument as that for part (a) and (b) that $\hat{p}$ satisfies Assumption SA-3 and SA-4. Finally, both $p$ and $\hat{p}$ reproduce polynomials of degree no greater than $m - 1$. Thus Assumption SA-5 holds. The proof is complete. \hfill \square

**SA-15.9 Proof of Lemma A.3**

**Proof.** (a): By construction, each basis function $p_k(x)$ is supported by only one subrectangle, and there are only a fixed number of $p_k(x)$’s which are not identically zero on each subrectangle. Thus Assumption SA-3(a) is satisfied. In addition, given one particular subrectangle $\delta_{l_1...l_d}$, store all basis functions supported on $\delta_{l_1...l_d}$ in a vector $p_{l_1...l_d}$. By [12, Lemma A.3],

$$Q_{l_1...l_d} := \mathbb{E}[p_{l_1...l_d}(x_i)p_{l_1...l_d}(x_i)'] \approx I_{\text{dim}(\mathcal{R}(\cdot))}$$

where $I_{\text{dim}(\mathcal{R}(\cdot))}$ is an identity matrix of size $\text{dim}(\mathcal{R}(\cdot))$. In fact, $\int_{\delta_{l_1...l_d}} p_{l_1...l_d}(x)p_{l_1...l_d}(x)' \, dx$ is a finite-dimensional matrix with the minimum eigenvalue bounded from below by $Ch^d$ for some $C > 0$. Hence for any $a \in \mathbb{R}^{\text{dim}(\mathcal{R}(\cdot))}$,

$$a' \int_{\delta_{l_1...l_d}} p_{l_1...l_d}(x)p_{l_1...l_d}(x)' \, dx \, a \geq Ch^d a' a$$

which suffices to show Assumption SA-3(b).

To show Assumption SA-3(c), simply notice that given any $x \in \mathcal{X}$, there are only a fixed number of nonzero elements in $\partial^q p(x)$, and for any $k = 1, \ldots, K$,

$$\sup_{\delta \in \Delta} \sup_{x \in \text{clo}(\delta)} |\partial^\varsigma p_k(x)| \lesssim h^{-\varsigma} \max_{|\alpha| = m-1} \frac{\alpha!}{(\alpha - \varsigma)!}$$

Moreover, for any $x \in \mathcal{X}$, there exists some $p_k$ in $p$ such that for $|\varsigma| \leq m - 1$, $|\partial^\varsigma p_k(x)| \gtrsim h^{-\varsigma}$.

(b): The result directly follows from the proofs of Lemma A.2 and Theorem 3 in [12]. The only difference here is that we use shifted Legendre polynomials to re-express the approximating function $s^*(x) = p(x)'\beta^*$ and the leading error. Clearly, $\beta^*$ is just a linear combination of coefficients of power series basis defined in their paper. The orthogonality between approximating basis and leading error directly follows from the property of Legendre polynomials and the fact that every
basis function is locally supported on only one cell.

(c): By construction of \( \tilde{p} \), \( \rho = 1 \). It follows from the same argument as that for part (a) and (b) that \( \tilde{p} \) satisfies Assumption SA-3 and SA-4. Finally, when the degree of piecewise polynomials is increased, \( \tilde{p} \) spans a larger space containing the span of \( p \), and both bases reproduce polynomials of degree no greater than \( m - 1 \). Thus Assumption SA-5 holds. \( \square \)

SA-15.10 Proof of Lemma A.4

Proof. Assumption SA-3(a), SA-3(c) and SA-4 directly follow from the construction of this basis and Taylor expansion restricted to a particular cell. For Assumption SA-3(b), given a generic cell \( \delta \), by Assumption SA-2, we can find an inscribed ball with diameter \( L_1 \) which is proportional to \( h_x \). Thus we can further find an inscribed rectangle with lengths equal to \( L_2 \) which is proportional to \( h_x \) as well. Thus, by changing variables and the same argument as that for the basis defined on rectangular cells, we have Assumption SA-3(b) holds. The properties of \( \tilde{p} \) follow similarly as in Lemma A.3. \( \square \)

SA-15.11 Proof of Theorem B.1

Proof. We divide the proof into two steps.

Step 1: For the integrated variance, first define operators \( \mathcal{M}(\cdot) \) and \( \mathcal{M}_q(\cdot) \) as follows:

\[
\mathcal{M}(\phi) := \int_X p(x)p(x)'\phi(x)dx,
\]

\[
\mathcal{M}_q(\phi) := \int_X \partial^q p(x)\partial^q p(x)'\phi(x)dx.
\]

Then

\[
\int_X \nabla[I[\partial^q \mu_q(x)]X]w(x)dx = \frac{1}{n} \text{trace} \left[ Q_m^{-1} \mathbb{E}[p(x_i)p(x_i)']\sigma^2(x_i^2)Q_m^{-1} \int_X \partial^q p(x)\partial^q p(x)'w(x)dx \right] + o_P(n^{-1}h^{-d})
\]

\[
= \frac{1}{n} \text{trace} \left[ \mathcal{M}(f)^{-1}\mathcal{M}(\sigma^2 f)\mathcal{M}(f)^{-1}\mathcal{M}_q(w) \right] + o_P(n^{-1}h^{-d-2}[q]).
\]

Moreover, define another operator generating \( K \times K \) diagonal matrix:

\[
\mathcal{D}(\phi) := \text{diag}\{\phi(\tau_1), \phi(\tau_2), \cdots, \phi(\tau_K)\}.
\]
Recall that $\tau_k$ is an arbitrary point in $\text{supp}(p_k)$, for $k = 1, \ldots, K$. Then we can write

$$\mathcal{M}(\phi) = \mathcal{M}(1) \mathcal{D}(\phi) - \mathcal{E}(\phi) \quad \text{(SA-5)}$$

where $\mathcal{E}(\phi)$ can be viewed as errors defined by Eq. (SA-5). Similarly, write

$$\mathcal{M}_q(\phi) = \mathcal{M}_q(1) \mathcal{D}(\phi) - \mathcal{E}_q(\phi).$$

Then it directly follows that

$$\mathcal{M}(f)^{-1} \mathcal{M}(\sigma^2 f) = [I - \mathcal{U}(f)]^{-1} [\mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) - \mathcal{L}(f, \sigma^2 f)] \quad \text{and}$$

$$\mathcal{M}(f)^{-1} \mathcal{M}_q(w) = [I - \mathcal{U}(f)]^{-1} [\mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) - \mathcal{L}_q(f, w)]$$

where

$$\mathcal{U}(\phi) := \mathcal{D}(\phi)^{-1} \mathcal{M}(1)^{-1} \mathcal{E}(\phi),$$

$$\mathcal{L}(\phi, \varphi) := \mathcal{D}(\phi)^{-1} \mathcal{M}(1)^{-1} \mathcal{E}(\phi)$$

and

$$\mathcal{L}_q(\phi, \varphi) := \mathcal{D}(\phi)^{-1} \mathcal{M}(1)^{-1} \mathcal{E}_q(\phi).$$

The number of nonzeros on any row or any column of $\mathcal{E}(\phi)$ (or $\mathcal{E}_q(\phi)$) is bounded by some constant. In fact, as explained in the proof of Lemma SA-8.1, it may take a multi-layer banded structure when we rectangularize the partition and arrange the ordering of basis functions properly. If $\text{supp}(p_k) \cap \text{supp}(p_l) \neq \emptyset$, then by Assumption SA-3 and the continuity of $f$, the $(k, l)$th element of $\mathcal{M}(f)$ can be approximated as follows:

$$\int_{\mathcal{X}} p_k(x)p_l(x)f(x)dx = f(\tau_k) \int_{\mathcal{X}} p_k(x)p_l(x)dx + o(h^d) = f(\tau_l) \int_{\mathcal{X}} p_k(x)p_l(x)dx + o(h^d). \quad \text{(SA-6)}$$

Moreover, since $\mathcal{X}$ is compact, $f$ is uniformly continuous. Thus we have $\|\mathcal{E}(f)\|_1 = o(h^d)$, $\|\mathcal{E}(f)\|_\infty = o(h^d)$, and then $\|\mathcal{E}(f)\| = o(h^d)$. Since $\|\mathcal{D}(f)^{-1}\| \lesssim 1$ and $\|\mathcal{M}(1)^{-1}\| \lesssim h^{-d}$, we
conclude \( \| \mathcal{U}(f) \| = o(1) \). For \( K \) large enough, we can make \( \| \mathcal{U}(f) \| < 1 \), and thus
\[
[I - \mathcal{U}(f)]^{-1} = I + \mathcal{U}(f) + \mathcal{U}(f)^2 + \cdots = I + \mathcal{W}(f)
\]
where \( \mathcal{W}(f) := \sum_{l=1}^{\infty} \mathcal{U}(f)^l \). Now we can write
\[
\text{trace} \left[ \mathcal{M}(f)^{-1} \mathcal{M}(\sigma^2 f) \mathcal{M}(f)^{-1} \mathcal{M}_q(w) \right] = \text{trace} \left[ \left( I + \mathcal{W}(f) \right) \left( \mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) - \mathcal{L}(f, \sigma^2 f) \right) \left( I + \mathcal{W}(f) \right) \right. \\
\times \left( \mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) - \mathcal{L}_q(f, w) \right) \right] = \text{trace} \left[ \left( \mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) + \mathcal{E}_1 \right) \left( \mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) + \mathcal{E}_2 \right) \right] = \text{trace} \left[ \mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) \mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) + \mathcal{E}_1 \mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) + \mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) \mathcal{E}_2 + \mathcal{E}_1 \mathcal{E}_2 \right]
\]
where
\[
\mathcal{E}_1 = -\mathcal{L}(f, \sigma^2 f) + \mathcal{W}(f) \mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) - \mathcal{W}(f) \mathcal{L}(f, \sigma^2 f),
\]
\[
\mathcal{E}_2 = -\mathcal{L}_q(f, w) + \mathcal{W}(f) \mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) - \mathcal{W}(f) \mathcal{L}_q(f, w).
\]

By assumptions in the theorem, \( \text{vol}(\delta_x) = \prod_{\ell=1}^d b_{x,\ell} = \prod_{\ell=1}^d \kappa_{\ell}^{-1} g_\ell(\kappa)^{-1} + o(h^d) \). Hence
\[
\text{trace} \left[ \mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) \mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) \right] = \prod_{\ell=1}^d \kappa_{\ell} \text{trace} \left[ \mathcal{D}(f)^{-1} \mathcal{D}(\sigma^2 f) \mathcal{D}(f)^{-1} \mathcal{M}(1)^{-1} \mathcal{M}_q(1) \mathcal{D}(w) \mathcal{D} \left( \prod_{\ell=1}^d g_\ell \right) \mathcal{D} \left( \prod_{\ell=1}^d g_\ell \right)^{-1} \prod_{\ell=1}^d \kappa_{\ell}^{-1} \right] = \kappa^{1+2q} \left( \sum_{k=1}^K \sum_{\ell=1}^d \frac{\sigma^2(\tau_k)w(\tau_k)}{f(\tau_k)} \prod_{\ell=1}^d g_\ell(\tau_k) e_k(1)^{-1} \kappa^{-2q} \mathcal{M}_q(1) e_k(\delta_{\tau_k}) \right) + o(\kappa^{1+2q})
\]

By Assumptions SA-2, SA-3 and Lemma SA-8.1, the summation in parenthesis is bounded from above and below. It remains to show all other terms are of smaller order. It directly follows from the same argument as that in the proof of [1, Theorem 6.1] that the trace of the remaining terms is \( o(\kappa^{1+2q}) \).

**Step 2:** We derive the integrated squared bias by analyzing the three components in (SA-1)
respectively. For $B_1$, notice that by assumption of the theorem

$$\mathcal{B}_{m,q}(x) = - \sum_{u \in \Lambda_m} \partial^u \mu(x) \left( \prod_{\ell=1}^d \kappa_{u-\kappa \ell}^{-u_\ell+q_\ell} g_\ell(x)^{-u_\ell+q_\ell} \right) b_x^{-u+q} h_x^{m-|q|} B_{m,q}(x) + o(h^{m-|q|}).$$

Recall that $\kappa = (\kappa_1, \ldots, \kappa_d)$ and define $g(x) := (g_1(x), \ldots, g_d(x))$. Given the above fact and using the same notation as in the proof of Theorem 4.1, we have:

$$B_1 = \sum_{u_1, u_2 \in \Lambda_m} \int_\mathcal{X} \left[ \partial^{u_1} \mu(x) \partial^{u_2} \mu(x) \frac{h_x^{2m-2|q|} B_{u_1,q}(x) B_{u_2,q}(x)}{\kappa_{u_1+u_2-2q} g(x)^{u_1+u_2-2q} b_x^{u_1+u_2-2q}} \right] w(x) dx + o(h^{2m-2|q|})$$

$$= \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1+u_2-2q)} \left( \sum_{\delta \in \Delta} \left[ \frac{\partial^{u_1} \mu(x) \partial^{u_2} \mu(x) w(x)}{g(x)^{u_1+u_2-2q}} \right] \frac{h_x^{2m-2|q|}}{b_x^{u_1+u_2-2q}} \int_\mathcal{X} B_{u_1,q}(x) B_{u_2,q}(x) dx \right) + o(h^{2m-2|q|})$$

$$= \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1+u_2-2q)} \eta_{u_1, u_2, q} \int_\mathcal{X} \frac{\partial^{u_1} \mu(x) \partial^{u_2} \mu(x) w(x)}{g(x)^{u_1+u_2-2q}} dx + o(h^{2m-2|q|})$$

where the last line follows from the integrability of $\partial^{u_1} \mu(x) \partial^{u_2} \mu(x) w(x)/g(x)^{u_1+u_2-2q}$ over $\mathcal{X}$.

For $B_2$, first notice that

$$\left\| E[p(x) \mathcal{B}_{m,0}(x)] + \sum_{u \in \Lambda_m} \kappa^{-u} \int_\mathcal{X} \left( p(x) \partial^u \mu(x) g(x)^{-u} b_x^{-u} h_x^m B_{m,0}(x) f(x) \right) dx \right\|_\infty = o(h^{m+d})$$

which implies that the errors given rise to by approximating $b_x$ are of smaller order. The integral in this approximation is a vector with typical elements given by

$$\int_\mathcal{X} \left( p_k(x) \partial^u \mu(x) g(x)^{-u} b_x^{-u} h_x^m B_{m,0}(x) f(x) \right) dx$$

$$= \partial^u \mu(\tau_k) g(\tau_k)^{-u} f(\tau_k) \int_\mathcal{X} p_k(x) b_x^{-u} h_x^m B_{m,0}(x) dx + o(h^d).$$

Recall that $v_{u,q}$ defined in the theorem has the $k$th element equal to

$$\frac{\partial^u \mu(\tau_k) \sqrt{v(\tau_k)}}{\kappa^q g(\tau_k)^{u-q}} \int_\mathcal{X} \partial^q p_k(x) b_x^{-(u-q)} h_x^{m-|q|} B_{u,q}(x) dx.$$
Given these results and Lemmas SA-8.1 and SA-8.2, it follows that

\[
B_2 = \text{trace} \left[ Q_m^{-1} \mathbb{E}[\mathcal{P}(x_i) \mathcal{B}_{m,0}(x_i)] \mathbb{E}[\mathcal{P}(x_i)'] \mathcal{B}_{m,0}(x_i)] Q^{-1} \int_X \partial^q \mathcal{P}(x) \partial^q \mathcal{P}(x)' w(x) dx \right] \\
= \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1 + u_2 - 2q)} \text{trace} \left( Q_m^{-1} \mathcal{D}(f/\sqrt{w}) v_{u_2,0} v'_{u_1,0} \mathcal{D}(f/\sqrt{w}) Q_m^{-1} \kappa^{-2q} \mathcal{M}_q(w) \right) + o(h^{2m-2|q|}) \\
= \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1 + u_2 - 2q)} \left( v'_{u_1,0} \mathcal{D}(1/\sqrt{w}) (\mathcal{M}^{-1} \mathcal{M}_q^{-1} \kappa^{-2q} \mathcal{M}_q(w)) \mathcal{D}(1/\sqrt{w}) v_{u_2,0} \right) + o(h^{2m-2|q|}) \\
= \sum_{u_1, u_2 \in \Lambda_m} \kappa^{-(u_1 + u_2 - 2q)} \left( v'_{u_1,0} \mathcal{D}(1/\sqrt{w}) \mathcal{M}^{-1} \mathcal{M}_q^{-1} \kappa^{-2q} \mathcal{M}_q(w) \mathcal{D}(1/\sqrt{w}) v_{u_2,0} \right) + o(h^{2m-2|q|})
\]

It should be noted that (SA-6) implies that the approximation given by (SA-5) is still valid if \( \mathcal{M}(1) \) is pre-multiplied by \( \mathcal{D}(\phi) \) instead of being post-multiplied if \( \phi(\cdot) \) is continuous. Therefore, we can use the argument in Step 1 again and further write the term in parenthesis in the last line as:

\[
v'_{u_1,0} \mathcal{D}(1/\sqrt{w}) \mathcal{M}^{-1} \mathcal{M}_q^{-1} \kappa^{-2q} \mathcal{M}_q(w) \mathcal{D}(1/\sqrt{w}) v_{u_2,0} \\
= v'_{u_1,0} \mathcal{D}(1/\sqrt{w}) (\mathcal{M}^{-1} \mathcal{M}_q^{-1} \kappa^{-2q} \mathcal{M}_q(w)) \mathcal{D}(1/\sqrt{w}) v_{u_2,0} + o(1)
\]

where \( |v'_{u_1,0} H_0^{-1} H_q H_0^{-1} v_{u_2,0}| \lesssim 1 \) by Assumption SA-2, SA-3 and Lemma SA-8.1.

Finally, for \( B_3 \), notice that

\[
\left\| \int_X \mathcal{B}_{m,q}(x) \partial^q \mathcal{P}(x)' w(x) dx + \sum_{u \in \Lambda_m} \int_X \left( \frac{\partial^u \mu(x) w(x)}{\kappa^{u-q} g(x)^{u-q} b_{u-q}^{1/2}} \right) \right\|_{\infty} \\
= o(h^{m+d-2|q|}).
\]
Thus repeating the argument for $B_2$, we have

$$B_3 = \left( \int_{\mathcal{X}} B_{m,q}(x) \partial^q p(x)' w(x) dx \right) Q_m^{-1} \mathbb{E}[p(x_i) B_{m,0}(x_i)]$$

$$= \sum_{u_1,u_2 \in \Lambda_m} \kappa^{-(u_1+u_2-2q)} v_{u_1,q} (\sqrt{w}) H_0^{-1} D(f)^{-1} D(f/\sqrt{w}) v_{u_2,0} + O(h^{2m-2|q|})$$

where $\left| v_{u_1,q} H_0^{-1} v_{u_2,0} \right| \lesssim 1$ by Assumption SA-2, SA-3 and Lemma SA-8.1. Then the proof is complete.

SA-15.12 Proof of Corollary 1

Proof. For the integrated variance, when $q = 0$ and $p$ generates $J$ complete covers, $\mathcal{M}(1)^{-1} \mathcal{M}_q(1) = I_K$ and hence

$$\text{trace} \left[ D(f)^{-1} D(\sigma^2 f) D(f)^{-1} D(w) \right] = \prod_{\ell=1}^d \kappa_\ell \times J \int_{\mathcal{X}} \frac{\sigma^2(x) w(x)}{f(x)} \prod_{\ell=1}^d g_\ell(x) \, dx + o\left( \prod_{\ell=1}^d \kappa_\ell \right).$$

For the integrated squared bias, since the approximate orthogonality condition holds, both $B_2$ and $B_3$ are of smaller order and the leading term in the integrated squared bias reduces to $B_1$ only.
Table SA-1: Pointwise Results, Model 1, Evenly-spaced, $n = 1000, 5000$ Replications

| $j$ | $\kappa$ | $x = 0.2$ | $x = 0.5$ | $x = 0.8$ |
|-----|----------|-----------|-----------|-----------|
|     |          | RMSE      | CR        | IL        | RMSE      | CR        | IL        | RMSE      | CR        | IL        |
| 0   | 1.0       | 0.084     | 56.4      | 0.185     | 0.058     | 84.7      | 0.249     | 0.069     | 69.2      | 0.185     |
|     | 2.0       | 0.055     | 78.9      | 0.190     | 0.045     | 82.2      | 0.175     | 0.003     | 94.5      | 0.190     |
| $\kappa_{\text{IMSE}}$ | 3.0       | 0.007     | 94.8      | 0.252     | 0.046     | 91.5      | 0.328     | 0.037     | 90.9      | 0.252     |
|     | 4.0       | 0.029     | 93.9      | 0.378     | 0.004     | 94.7      | 0.221     | 0.022     | 93.9      | 0.378     |
|     | 5.0       | 0.002     | 94.9      | 0.319     | 0.018     | 94.4      | 0.400     | 0.002     | 94.9      | 0.319     |
| $\hat{\kappa}_{\text{ROT}}$ | 4.9       | 0.008     | 95.1      | 0.328     | 0.009     | 94.6      | 0.317     | 0.009     | 94.2      | 0.328     |
| $\hat{\kappa}_{\text{DPI}}$ | 5.1       | 0.006     | 95.2      | 0.323     | 0.007     | 94.4      | 0.318     | 0.008     | 94.5      | 0.323     |
| 1   | 1.0       | 0.032     | 90.8      | 0.205     | 0.030     | 90.3      | 0.186     | 0.056     | 80.7      | 0.205     |
|     | 2.0       | 0.026     | 93.2      | 0.269     | 0.013     | 94.1      | 0.251     | 0.033     | 92.4      | 0.269     |
| $\kappa_{\text{IMSE}}$ | 3.0       | 0.002     | 95.3      | 0.304     | 0.003     | 94.8      | 0.226     | 0.001     | 94.8      | 0.304     |
|     | 4.0       | 0.003     | 95.2      | 0.278     | 0.008     | 94.7      | 0.308     | 0.011     | 94.4      | 0.278     |
|     | 5.0       | 0.001     | 94.7      | 0.297     | 0.008     | 94.9      | 0.267     | 0.009     | 94.6      | 0.297     |
| $\hat{\kappa}_{\text{ROT}}$ | 4.9       | 0.001     | 95.3      | 0.304     | 0.006     | 95.0      | 0.298     | 0.008     | 95.0      | 0.304     |
| $\hat{\kappa}_{\text{DPI}}$ | 5.1       | 0.001     | 95.0      | 0.315     | 0.006     | 95.1      | 0.306     | 0.007     | 94.7      | 0.315     |
| 2   | 1.0       | 0.029     | 91.4      | 0.205     | 0.018     | 93.4      | 0.194     | 0.054     | 82.2      | 0.205     |
|     | 2.0       | 0.032     | 92.5      | 0.270     | 0.013     | 94.2      | 0.251     | 0.038     | 91.3      | 0.270     |
| $\kappa_{\text{IMSE}}$ | 3.0       | 0.003     | 95.1      | 0.315     | 0.004     | 94.7      | 0.268     | 0.002     | 94.8      | 0.315     |
|     | 4.0       | 0.004     | 94.5      | 0.314     | 0.009     | 94.8      | 0.311     | 0.001     | 94.9      | 0.314     |
|     | 5.0       | 0.000     | 94.4      | 0.311     | 0.001     | 94.6      | 0.340     | 0.006     | 94.8      | 0.311     |
| $\hat{\kappa}_{\text{ROT}}$ | 4.9       | 0.001     | 94.7      | 0.323     | 0.003     | 95.0      | 0.336     | 0.002     | 95.2      | 0.323     |
| $\hat{\kappa}_{\text{DPI}}$ | 5.1       | 0.001     | 94.5      | 0.332     | 0.003     | 94.9      | 0.342     | 0.001     | 95.1      | 0.332     |
| 3   | 1.0       | 0.046     | 85.1      | 0.198     | 0.184     | 22.8      | 0.269     | 0.038     | 88.2      | 0.198     |
|     | 2.0       | 0.005     | 95.0      | 0.223     | 0.012     | 94.1      | 0.198     | 0.043     | 87.7      | 0.223     |
| $\kappa_{\text{IMSE}}$ | 3.0       | 0.005     | 95.0      | 0.248     | 0.034     | 92.7      | 0.321     | 0.036     | 90.6      | 0.248     |
|     | 4.0       | 0.001     | 94.6      | 0.403     | 0.008     | 94.5      | 0.250     | 0.034     | 92.9      | 0.341     |
|     | 5.0       | 0.002     | 95.0      | 0.318     | 0.021     | 93.8      | 0.391     | 0.003     | 94.8      | 0.318     |
| $\hat{\kappa}_{\text{ROT}}$ | 4.9       | 0.002     | 95.4      | 0.342     | 0.006     | 94.8      | 0.328     | 0.004     | 93.9      | 0.322     |
| $\hat{\kappa}_{\text{DPI}}$ | 5.1       | 0.000     | 95.3      | 0.339     | 0.005     | 94.3      | 0.331     | 0.005     | 94.8      | 0.321     |
|                | $x = 0.2$ |          |          | $x = 0.5$ |          |          | $x = 0.8$ |          |          |
|----------------|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|
| $\kappa$               | RMSE     | CR      | IL       | RMSE     | CR      | IL       | RMSE     | CR      | IL       |
| $j = 0$ | 5.0       | 0.066   | 86.9 0.320 | 0.083 | 87.0 0.403 | 0.001 | 94.9 0.319 |
|                 | 6.0       | 0.042   | 90.7 0.277 | 0.096 | 69.6 0.262 | 0.004 | 94.8 0.276 |
| $\kappa_{\text{IMSE}}$ | 7.0       | 0.007   | 94.8 0.291 | 0.079 | 89.4 0.464 | 0.007 | 94.6 0.291 |
|                 | 8.0       | 0.019   | 93.9 0.377 | 0.046 | 90.2 0.297 | 0.002 | 94.8 0.378 |
| $\hat{\kappa}_{\text{ROT}}$ | 9.0       | 0.026   | 93.9 0.517 | 0.061 | 91.8 0.518 | 0.004 | 94.7 0.519 |
| $\hat{\kappa}_{\text{DPI}}$ | 8.5       | 0.022   | 93.6 0.424 | 0.007 | 91.2 0.396 | 0.002 | 94.0 0.424 |
| $j = 1$ | 5.0       | 0.066   | 85.5 0.297 | 0.098 | 69.6 0.268 | 0.049 | 89.8 0.297 |
|                 | 6.0       | 0.000   | 94.0 0.360 | 0.020 | 94.1 0.363 | 0.027 | 93.7 0.360 |
| $\kappa_{\text{IMSE}}$ | 7.0       | 0.010   | 94.2 0.395 | 0.032 | 92.8 0.307 | 0.005 | 94.6 0.396 |
|                 | 8.0       | 0.012   | 94.2 0.381 | 0.019 | 94.2 0.411 | 0.001 | 94.9 0.381 |
| $\hat{\kappa}_{\text{ROT}}$ | 9.0       | 0.001   | 94.5 0.351 | 0.007 | 94.6 0.343 | 0.000 | 94.9 0.351 |
| $\hat{\kappa}_{\text{DPI}}$ | 8.3       | 0.008   | 94.3 0.371 | 0.005 | 94.4 0.378 | 0.000 | 94.7 0.371 |
| $j = 2$ | 5.0       | 0.046   | 91.1 0.311 | 0.013 | 94.4 0.340 | 0.028 | 93.4 0.311 |
|                 | 6.0       | 0.005   | 94.3 0.362 | 0.024 | 93.9 0.366 | 0.031 | 93.4 0.363 |
| $\kappa_{\text{IMSE}}$ | 7.0       | 0.014   | 94.3 0.403 | 0.001 | 94.5 0.399 | 0.013 | 94.4 0.403 |
|                 | 8.0       | 0.001   | 94.4 0.413 | 0.022 | 94.4 0.415 | 0.015 | 94.3 0.413 |
| $\hat{\kappa}_{\text{ROT}}$ | 9.0       | 0.001   | 94.2 0.435 | 0.002 | 94.3 0.447 | 0.002 | 94.7 0.436 |
| $\hat{\kappa}_{\text{DPI}}$ | 8.3       | 0.001   | 94.3 0.420 | 0.013 | 94.4 0.426 | 0.009 | 94.3 0.421 |
| $j = 3$ | 5.0       | 0.073   | 85.1 0.320 | 0.103 | 82.4 0.394 | 0.002 | 94.6 0.318 |
|                 | 6.0       | 0.018   | 94.0 0.302 | 0.041 | 91.4 0.297 | 0.013 | 94.7 0.302 |
| $\kappa_{\text{IMSE}}$ | 7.0       | 0.001   | 94.5 0.325 | 0.056 | 91.5 0.454 | 0.008 | 94.8 0.325 |
|                 | 8.0       | 0.019   | 93.9 0.375 | 0.010 | 94.5 0.337 | 0.002 | 94.7 0.375 |
| $\hat{\kappa}_{\text{ROT}}$ | 9.0       | 0.029   | 93.9 0.520 | 0.042 | 92.7 0.508 | 0.003 | 94.8 0.492 |
| $\hat{\kappa}_{\text{DPI}}$ | 8.3       | 0.001   | 93.3 0.425 | 0.025 | 93.4 0.409 | 0.002 | 94.9 0.415 |
|                 | 8.5       | 0.003   | 93.6 0.427 | 0.023 | 93.2 0.414 | 0.000 | 94.0 0.417 |

119
Table SA-3: Pointwise Results, Model 3, Evenly-spaced, \( n = 1000, 5000 \) Replications

| \( j \) | \( \kappa \) | \( x = 0.2 \) | \( x = 0.5 \) | \( x = 0.8 \) |
|-------|-------|-------|-------|-------|
|       | RMSE  | CR    | IL    | RMSE  | CR    | IL    | RMSE  | CR    | IL    |
| \( j = 0 \) |       |       |       |       |       |       |       |       |       |
| 10.0  | 0.001 | 94.5  | 0.419 | 0.489 | 0.04  | 0.343 | 0.002 | 94.4  | 0.419 |
| 11.0  | 0.003 | 94.5  | 0.355 | 0.244 | 61.00 | 0.570 | 0.003 | 94.4  | 0.355 |
| \( \kappa_{\text{IMSE}} \) 12.0 | 0.004 | 94.1  | 0.369 | 0.350 | 3.64  | 0.365 | 0.004 | 94.9  | 0.369 |
| 13.0  | 0.005 | 94.6  | 0.470 | 0.240 | 66.50 | 0.616 | 0.004 | 94.7  | 0.472 |
| 14.0  | 0.002 | 94.1  | 0.632 | 0.253 | 27.80 | 0.389 | 0.000 | 94.7  | 0.634 |
| \( \hat{\kappa}_{\text{ROT}} \) 7.3  | 0.018 | 95.6  | 0.319 | 0.193 | 59.60 | 0.414 | 0.019 | 95.2  | 0.319 |
| \( \hat{\kappa}_{\text{DPI}} \) 10.4 | 0.002 | 94.4  | 0.423 | 0.178 | 31.40 | 0.430 | 0.003 | 95.1  | 0.424 |
| \( j = 1 \) |       |       |       |       |       |       |       |       |       |
| 10.0  | 0.000 | 94.3  | 0.400 | 0.036 | 93.40 | 0.453 | 0.000 | 94.8  | 0.400 |
| 11.0  | 0.028 | 93.5  | 0.467 | 0.283 | 17.30 | 0.382 | 0.028 | 94.1  | 0.468 |
| \( \kappa_{\text{IMSE}} \) 12.0 | 0.005 | 94.1  | 0.489 | 0.010 | 94.00 | 0.493 | 0.003 | 94.7  | 0.490 |
| 13.0  | 0.002 | 93.9  | 0.456 | 0.178 | 60.00 | 0.409 | 0.002 | 94.6  | 0.457 |
| 14.0  | 0.004 | 94.2  | 0.422 | 0.025 | 93.70 | 0.530 | 0.004 | 94.6  | 0.422 |
| \( \hat{\kappa}_{\text{ROT}} \) 7.3  | 0.048 | 88.8  | 0.391 | 0.482 | 26.30 | 0.358 | 0.047 | 88.7  | 0.391 |
| \( \hat{\kappa}_{\text{DPI}} \) 10.4 | 0.008 | 93.7  | 0.418 | 0.163 | 60.20 | 0.426 | 0.007 | 94.2  | 0.418 |
| \( j = 2 \) |       |       |       |       |       |       |       |       |       |
| 10.0  | 0.001 | 94.1  | 0.433 | 0.038 | 93.40 | 0.459 | 0.000 | 94.7  | 0.434 |
| 11.0  | 0.037 | 93.4  | 0.474 | 0.044 | 92.70 | 0.491 | 0.036 | 93.6  | 0.475 |
| \( \kappa_{\text{IMSE}} \) 12.0 | 0.009 | 94.1  | 0.499 | 0.019 | 93.90 | 0.500 | 0.008 | 94.7  | 0.500 |
| 13.0  | 0.012 | 94.0  | 0.502 | 0.028 | 93.70 | 0.531 | 0.014 | 94.4  | 0.504 |
| 14.0  | 0.002 | 94.1  | 0.542 | 0.037 | 93.40 | 0.537 | 0.004 | 94.2  | 0.544 |
| \( \hat{\kappa}_{\text{ROT}} \) 7.3  | 0.101 | 70.0  | 0.407 | 0.194 | 54.60 | 0.408 | 0.100 | 70.5  | 0.408 |
| \( \hat{\kappa}_{\text{DPI}} \) 10.4 | 0.021 | 92.5  | 0.454 | 0.040 | 89.50 | 0.471 | 0.020 | 93.6  | 0.455 |
| \( j = 3 \) |       |       |       |       |       |       |       |       |       |
| 10.0  | 0.001 | 94.4  | 0.417 | 0.282 | 18.70 | 0.385 | 0.002 | 94.4  | 0.418 |
| 11.0  | 0.007 | 94.2  | 0.392 | 0.077 | 90.30 | 0.556 | 0.007 | 94.3  | 0.392 |
| \( \kappa_{\text{IMSE}} \) 12.0 | 0.006 | 94.2  | 0.409 | 0.172 | 63.10 | 0.412 | 0.005 | 94.8  | 0.410 |
| 13.0  | 0.006 | 94.6  | 0.468 | 0.079 | 90.30 | 0.601 | 0.004 | 94.8  | 0.469 |
| 14.0  | 0.003 | 94.2  | 0.624 | 0.104 | 83.10 | 0.440 | 0.002 | 94.6  | 0.617 |
| \( \hat{\kappa}_{\text{ROT}} \) 7.3  | 0.046 | 91.9  | 0.341 | 0.212 | 57.20 | 0.422 | 0.047 | 91.2  | 0.341 |
| \( \hat{\kappa}_{\text{DPI}} \) 10.4 | 0.020 | 93.2  | 0.436 | 0.126 | 60.60 | 0.450 | 0.006 | 94.9  | 0.431 |
Table SA-4: Pointwise Results, Model 4, Evenly-spaced, $n = 1000$, 5000 Replications

|   | $\kappa$ | $\kappa_{\text{IMSE}}$ | $\kappa_{\text{ROT}}$ | $\kappa_{\text{DPI}}$ | $x = (0.5, 0.5)$ | $x = (0.1, 0.5)$ | $x = (0.1, 0.1)$ |
|---|----------|--------------------------|------------------------|------------------------|------------------|------------------|------------------|
| $j = 0$ |          |                          |                        |                        | RMSE  | CR  | IL  | RMSE  | CR  | IL  | RMSE  | CR  | IL  |
| 2.0 | 0.515   | 0.0 0.266                | 0.405 1.68 0.384       | 0.087 89.4 0.574       |
| 3.0 | 0.324   | 68.9 0.880               | 0.135 87.70 0.696      | 0.116 85.3 0.556       |
| 4.0 | 0.183   | 57.0 0.401               | 0.142 78.10 0.470      | 0.074 90.3 0.550       |
| 5.0 | 0.191   | 89.9 1.310               | 0.109 91.30 0.886      | 0.033 93.5 0.597       |
| 6.0 | 0.070   | 91.5 0.563               | 0.045 93.60 0.649      | 0.002 93.6 0.745       |
| 5.5 | 0.072   | 90.8 0.968               | 0.034 93.90 0.776      | 0.014 93.5 0.664       |
| 4.7 | 0.075   | 77.9 1.020               | 0.033 84.80 0.755      | 0.043 92.5 0.582       |
| $j = 1$ |          |                          |                        |                        | RMSE  | CR  | IL  | RMSE  | CR  | IL  | RMSE  | CR  | IL  |
| 2.0 | 0.027   | 93.7 0.510               | 0.000 95.20 0.516      | 0.006 94.5 0.526       |
| 3.0 | 0.164   | 66.1 0.417               | 0.119 85.60 0.516      | 0.078 92.2 0.636       |
| 4.0 | 0.042   | 93.3 0.772               | 0.035 94.20 0.821      | 0.039 93.5 0.871       |
| 5.0 | 0.024   | 93.9 0.583               | 0.012 94.30 0.829      | 0.013 93.6 1.180       |
| 6.0 | 0.013   | 93.8 1.070               | 0.023 93.40 1.240      | 0.006 93.0 1.450       |
| 5.5 | 0.007   | 93.8 0.807               | 0.003 93.70 1.020      | 0.010 93.2 1.300       |
| 4.7 | 0.004   | 94.0 0.643               | 0.003 94.40 0.827      | 0.024 93.6 1.080       |
| $j = 2$ |          |                          |                        |                        | RMSE  | CR  | IL  | RMSE  | CR  | IL  | RMSE  | CR  | IL  |
| 2.0 | 0.028   | 93.7 0.510               | 0.001 95.10 0.519      | 0.000 94.4 0.538       |
| 3.0 | 0.029   | 93.8 0.659               | 0.035 94.60 0.597      | 0.073 92.3 0.642       |
| 4.0 | 0.052   | 93.2 0.779               | 0.041 94.30 0.827      | 0.040 93.6 0.875       |
| 5.0 | 0.017   | 94.1 1.070               | 0.004 94.00 0.957      | 0.013 93.6 1.190       |
| 6.0 | 0.014   | 94.4 1.090               | 0.025 93.80 1.270      | 0.007 93.3 1.480       |
| 5.5 | 0.003   | 94.3 1.080               | 0.008 93.60 1.100      | 0.011 93.4 1.320       |
| 4.7 | 0.006   | 94.2 0.977               | 0.011 94.40 0.916      | 0.024 93.6 1.090       |
| $j = 3$ |          |                          |                        |                        | RMSE  | CR  | IL  | RMSE  | CR  | IL  | RMSE  | CR  | IL  |
| 2.0 | 0.255   | 27.7 0.390               | 0.189 59.90 0.434      | 0.028 94.3 0.549       |
| 3.0 | 0.275   | 74.5 0.838               | 0.124 89.20 0.696      | 0.064 93.1 0.626       |
| 4.0 | 0.045   | 93.0 0.563               | 0.028 94.90 0.686      | 0.039 93.9 0.818       |
| 5.0 | 0.111   | 92.5 1.260               | 0.070 93.40 0.968      | 0.023 93.5 0.998       |
| 6.0 | 0.006   | 94.1 0.788               | 0.007 94.10 0.896      | 0.020 93.4 1.020       |
| 5.5 | 0.057   | 93.3 1.050               | 0.039 94.10 0.934      | 0.022 93.5 1.010       |
| 4.7 | 0.063   | 92.3 1.040               | 0.039 93.20 0.879      | 0.030 93.5 0.941       |
Table SA-5: Pointwise Results, Model 1, Quantile-spaced, $n = 1000$, 5000 Replications

| $\kappa$ | $x = 0.2$ | | $x = 0.5$ | | $x = 0.8$ | |
|----------|-----------|---|---|---|---|---|
|          | RMSE      | CR | IL | RMSE | CR | IL | RMSE | CR | IL | RMSE | CR | IL |
| $j = 0$  |           |    |    |      |    |    |      |    |    |      |    |    |
| 1.0      | 0.084     | 56.5 | 0.185 | 0.046 | 87.6 | 0.240 | 0.068 | 69.6 | 0.185 |       |    |    |
| 2.0      | 0.054     | 79.0 | 0.190 | 0.045 | 81.9 | 0.176 | 0.036 | 90.8 | 0.253 |       |    |    |
| $\kappa_{IMSE}$ 3.0 | 0.006 | 94.8 | 0.252 | 0.036 | 92.1 | 0.308 | 0.036 | 90.8 | 0.253 |       |    |    |
| 4.0      | 0.021     | 94.3 | 0.352 | 0.003 | 94.5 | 0.225 | 0.018 | 94.0 | 0.352 |       |    |    |
| 5.0      | 0.003     | 94.6 | 0.322 | 0.013 | 94.5 | 0.364 | 0.001 | 94.9 | 0.321 |       |    |    |
| $\hat{\kappa}_{ROT}$ 4.9 | 0.006 | 94.8 | 0.323 | 0.006 | 94.7 | 0.302 | 0.007 | 94.3 | 0.322 |       |    |    |
| $\hat{\kappa}_{DPI}$ 5.1 | 0.005 | 95.0 | 0.318 | 0.005 | 94.1 | 0.305 | 0.006 | 94.6 | 0.317 |       |    |    |
| $j = 1$  |           |    |    |      |    |    |      |    |    |      |    |    |
| 1.0      | 0.032     | 90.7 | 0.205 | 0.030 | 89.5 | 0.187 | 0.056 | 80.7 | 0.205 |       |    |    |
| 2.0      | 0.025     | 93.3 | 0.269 | 0.012 | 94.2 | 0.251 | 0.033 | 92.3 | 0.269 |       |    |    |
| $\kappa_{IMSE}$ 3.0 | 0.002 | 95.4 | 0.303 | 0.004 | 94.6 | 0.228 | 0.001 | 94.8 | 0.303 |       |    |    |
| 4.0      | 0.003     | 95.1 | 0.281 | 0.008 | 94.6 | 0.307 | 0.010 | 94.4 | 0.280 |       |    |    |
| 5.0      | 0.001     | 94.7 | 0.298 | 0.008 | 94.7 | 0.274 | 0.008 | 94.7 | 0.298 |       |    |    |
| $\hat{\kappa}_{ROT}$ 4.9 | 0.002 | 95.5 | 0.305 | 0.006 | 94.9 | 0.300 | 0.008 | 95.0 | 0.306 |       |    |    |
| $\hat{\kappa}_{DPI}$ 5.1 | 0.001 | 94.9 | 0.316 | 0.006 | 95.0 | 0.307 | 0.006 | 94.7 | 0.316 |       |    |    |
| $j = 2$  |           |    |    |      |    |    |      |    |    |      |    |    |
| 1.0      | 0.029     | 91.2 | 0.205 | 0.019 | 93.1 | 0.194 | 0.054 | 82.0 | 0.206 |       |    |    |
| 2.0      | 0.031     | 92.4 | 0.270 | 0.013 | 94.2 | 0.251 | 0.038 | 91.2 | 0.270 |       |    |    |
| $\kappa_{IMSE}$ 3.0 | 0.002 | 95.0 | 0.314 | 0.004 | 94.5 | 0.264 | 0.001 | 94.8 | 0.313 |       |    |    |
| 4.0      | 0.003     | 94.6 | 0.310 | 0.009 | 94.6 | 0.310 | 0.000 | 95.0 | 0.310 |       |    |    |
| 5.0      | 0.000     | 94.3 | 0.314 | 0.002 | 94.5 | 0.329 | 0.005 | 94.8 | 0.314 |       |    |    |
| $\hat{\kappa}_{ROT}$ 4.9 | 0.001 | 94.7 | 0.323 | 0.004 | 94.9 | 0.330 | 0.002 | 95.2 | 0.323 |       |    |    |
| $\hat{\kappa}_{DPI}$ 5.1 | 0.001 | 94.5 | 0.332 | 0.004 | 94.8 | 0.337 | 0.001 | 94.9 | 0.332 |       |    |    |
| $j = 3$  |           |    |    |      |    |    |      |    |    |      |    |    |
| 1.0      | 0.046     | 85.0 | 0.198 | 0.022 | 35.3 | 0.254 | 0.038 | 88.3 | 0.198 |       |    |    |
| 2.0      | 0.005     | 95.0 | 0.223 | 0.012 | 93.9 | 0.198 | 0.043 | 87.6 | 0.223 |       |    |    |
| $\kappa_{IMSE}$ 3.0 | 0.005 | 94.9 | 0.249 | 0.012 | 90.0 | 0.299 | 0.037 | 90.7 | 0.249 |       |    |    |
| 4.0      | 0.000     | 94.8 | 0.341 | 0.008 | 94.7 | 0.251 | 0.005 | 93.6 | 0.341 |       |    |    |
| 5.0      | 0.002     | 94.9 | 0.325 | 0.003 | 93.8 | 0.355 | 0.004 | 94.8 | 0.324 |       |    |    |
| $\hat{\kappa}_{ROT}$ 4.9 | 0.002 | 94.9 | 0.326 | 0.006 | 94.8 | 0.311 | 0.001 | 94.7 | 0.324 |       |    |    |
| $\hat{\kappa}_{DPI}$ 5.1 | 0.001 | 94.7 | 0.324 | 0.005 | 94.4 | 0.316 | 0.000 | 94.8 | 0.324 |       |    |    |
Table SA-6: Pointwise Results, Model 2, Quantile-spaced, $n = 1000$, 5000 Replications

| $j$ | $\kappa$ | $x = 0.2$ | $x = 0.5$ | $x = 0.8$ |
|-----|----------|------------|------------|------------|
|     | RMSE     | CR         | IL         | RMSE       | CR         | IL         | RMSE       | CR         | IL         |
| $j = 0$ |          |            |            |            |            |            |            |            |            |
| 5.0 | 0.058    | 87.7 0.323 | 0.039 88.5 0.367 | 0.003 94.3 0.321 |
| 6.0 | 0.041    | 90.7 0.282 | 0.087 72.0 0.271 | 0.003 94.8 0.281 |
| $\kappa_{IMSE}$ | 7.0 | 0.009 94.2 0.296 | 0.036 91.1 0.409 | 0.006 94.6 0.297 |
|     | 8.0 | 0.015 94.2 0.377 | 0.037 90.7 0.313 | 0.001 94.6 0.380 |
|     | 9.0 | 0.014 94.0 0.456 | 0.026 92.5 0.443 | 0.001 94.7 0.456 |
| $\hat{\kappa}_{ROT}$ | 8.3 | 0.012 94.1 0.399 | 0.008 91.5 0.370 | 0.002 94.8 0.402 |
| $\hat{\kappa}_{DPI}$ | 8.6 | 0.015 93.6 0.406 | 0.002 91.5 0.378 | 0.001 94.2 0.407 |
| $j = 1$ |          |            |            |            |            |            |            |            |            |
| 5.0 | 0.060    | 86.6 0.298 | 0.091 71.2 0.274 | 0.046 90.3 0.299 |
| 6.0 | 0.001    | 94.1 0.357 | 0.014 93.4 0.358 | 0.026 93.8 0.358 |
| $\kappa_{IMSE}$ | 7.0 | 0.008 94.1 0.391 | 0.026 92.6 0.319 | 0.004 94.7 0.391 |
|     | 8.0 | 0.009 94.3 0.380 | 0.014 93.6 0.401 | 0.001 94.8 0.379 |
|     | 9.0 | 0.002 94.3 0.364 | 0.002 94.2 0.362 | 0.000 94.7 0.364 |
| $\hat{\kappa}_{ROT}$ | 8.3 | 0.006 94.3 0.376 | 0.005 93.8 0.381 | 0.000 94.6 0.375 |
| $\hat{\kappa}_{DPI}$ | 8.6 | 0.006 94.8 0.378 | 0.004 93.4 0.385 | 0.001 95.1 0.379 |
| $j = 2$ |          |            |            |            |            |            |            |            |            |
| 5.0 | 0.040    | 91.2 0.314 | 0.020 93.0 0.329 | 0.027 93.2 0.314 |
| 6.0 | 0.004    | 94.0 0.361 | 0.022 93.9 0.365 | 0.028 93.5 0.362 |
| $\kappa_{IMSE}$ | 7.0 | 0.012 94.2 0.400 | 0.002 94.4 0.382 | 0.012 94.3 0.400 |
|     | 8.0 | 0.000 94.4 0.413 | 0.020 94.3 0.412 | 0.013 94.2 0.414 |
|     | 9.0 | 0.001 94.2 0.422 | 0.004 93.9 0.427 | 0.001 94.7 0.423 |
| $\hat{\kappa}_{ROT}$ | 8.3 | 0.000 94.5 0.415 | 0.012 94.0 0.416 | 0.008 94.3 0.416 |
| $\hat{\kappa}_{DPI}$ | 8.6 | 0.000 94.3 0.419 | 0.011 94.2 0.422 | 0.006 94.5 0.420 |
| $j = 3$ |          |            |            |            |            |            |            |            |            |
| 5.0 | 0.078    | 83.0 0.326 | 0.010 71.0 0.358 | 0.001 94.7 0.324 |
| 6.0 | 0.020    | 93.6 0.305 | 0.040 90.1 0.300 | 0.012 94.5 0.305 |
| $\kappa_{IMSE}$ | 7.0 | 0.001 94.4 0.327 | 0.012 87.7 0.400 | 0.008 94.7 0.327 |
|     | 8.0 | 0.016 93.9 0.379 | 0.008 93.7 0.344 | 0.001 94.4 0.381 |
|     | 9.0 | 0.004 93.6 0.445 | 0.010 91.4 0.437 | 0.002 95.0 0.446 |
| $\hat{\kappa}_{ROT}$ | 8.3 | 0.010 93.6 0.399 | 0.000 92.7 0.384 | 0.000 94.9 0.402 |
| $\hat{\kappa}_{DPI}$ | 8.6 | 0.012 93.3 0.406 | 0.004 92.5 0.391 | 0.000 94.4 0.407 |
Table SA-7: Pointwise Results, Model 3, Quantile-spaced, \( n = 1000, 5000 \) Replications

|       | \( \kappa \) | \( x = 0.2 \) | \( x = 0.5 \) | \( x = 0.8 \) |
|-------|--------------|---------------|---------------|---------------|
|       |              | RMSE | CR | IL    | RMSE | CR | IL    | RMSE | CR | IL    |
| \( j = 0 \) |              |      |    |       |      |    |       |      |    |       |
| \( \kappa_{\text{IMSE}} \) | 12.0         | 0.003 | 94.5 | 0.389 | 0.250 | 32.2 | 0.403 | 0.003 | 94.5 | 0.391 |
| \( \kappa_{\text{ROT}} \) | 7.3          | 0.015 | 95.1 | 0.322 | 0.316 | 39.2 | 0.386 | 0.016 | 95.1 | 0.323 |
| \( \kappa_{\text{DPI}} \) | 10.5         | 0.001 | 94.6 | 0.413 | 0.236 | 38.0 | 0.411 | 0.001 | 94.8 | 0.414 |
| \( j = 1 \) |              |      |    |       |      |    |       |      |    |       |
| \( \kappa_{\text{IMSE}} \) | 12.0         | 0.007 | 94.1 | 0.476 | 0.049 | 87.3 | 0.472 | 0.006 | 94.6 | 0.476 |
| \( \kappa_{\text{ROT}} \) | 7.3          | 0.048 | 88.6 | 0.388 | 0.464 | 21.3 | 0.364 | 0.048 | 88.7 | 0.388 |
| \( \kappa_{\text{DPI}} \) | 10.5         | 0.002 | 94.0 | 0.423 | 0.154 | 64.7 | 0.431 | 0.002 | 94.7 | 0.424 |
| \( j = 2 \) |              |      |    |       |      |    |       |      |    |       |
| \( \kappa_{\text{IMSE}} \) | 12.0         | 0.009 | 94.1 | 0.495 | 0.004 | 93.5 | 0.494 | 0.007 | 94.4 | 0.496 |
| \( \kappa_{\text{ROT}} \) | 7.3          | 0.095 | 72.9 | 0.405 | 0.232 | 40.6 | 0.397 | 0.095 | 72.4 | 0.406 |
| \( \kappa_{\text{DPI}} \) | 10.5         | 0.015 | 93.7 | 0.455 | 0.057 | 86.0 | 0.463 | 0.015 | 94.2 | 0.457 |
| \( j = 3 \) |              |      |    |       |      |    |       |      |    |       |
| \( \kappa_{\text{IMSE}} \) | 12.0         | 0.005 | 94.4 | 0.429 | 0.259 | 29.3 | 0.395 | 0.004 | 94.4 | 0.428 |
| \( \kappa_{\text{ROT}} \) | 7.3          | 0.042 | 92.9 | 0.344 | 0.319 | 24.8 | 0.391 | 0.042 | 92.0 | 0.345 |
| \( \kappa_{\text{DPI}} \) | 10.5         | 0.004 | 94.6 | 0.425 | 0.179 | 54.7 | 0.426 | 0.005 | 94.4 | 0.426 |
Table SA-8: Pointwise Results, Model 4, Quantile-spaced, \( n = 1000, 5000 \) Replications

| \( j \) | \( \kappa \) | \( x = (0.5, 0.5) \) | \( x = (0.1, 0.5) \) | \( x = (0.1, 0.1) \) |
|---|---|---|---|---|
| 0 | 2.0 | 0.511 0.0 0.269 | 0.402 2.16 0.385 | 0.087 89.2 0.573 |
| | 3.0 | 0.214 79.5 0.774 | 0.081 91.00 0.652 | 0.115 85.4 0.556 |
| \( \kappa_{\text{IMSE}} \) | 4.0 | 0.174 61.3 0.414 | 0.136 79.40 0.477 | 0.073 90.2 0.552 |
| | 5.0 | 0.113 91.8 1.080 | 0.062 93.30 0.807 | 0.034 93.5 0.602 |
| | 6.0 | 0.062 92.0 0.600 | 0.039 93.70 0.673 | 0.003 93.9 0.755 |
| \( \hat{\kappa}_{\text{ROT}} \) | 5.5 | 0.034 91.8 0.863 | 0.013 94.20 0.744 | 0.015 93.9 0.671 |
| \( \hat{\kappa}_{\text{DPI}} \) | 4.7 | 0.032 82.4 0.890 | 0.008 88.00 0.712 | 0.044 92.8 0.587 |
| 1 | 2.0 | 0.030 93.5 0.509 | 0.002 95.00 0.516 | 0.005 94.2 0.527 |
| | 3.0 | 0.158 68.5 0.426 | 0.115 86.20 0.522 | 0.077 92.0 0.637 |
| \( \kappa_{\text{IMSE}} \) | 4.0 | 0.038 93.6 0.764 | 0.031 94.30 0.814 | 0.039 93.7 0.870 |
| | 5.0 | 0.020 93.7 0.611 | 0.008 94.40 0.844 | 0.013 93.5 1.170 |
| | 6.0 | 0.011 93.6 1.040 | 0.020 93.90 1.220 | 0.006 93.1 1.430 |
| \( \hat{\kappa}_{\text{ROT}} \) | 5.5 | 0.006 93.6 0.808 | 0.003 94.00 1.010 | 0.010 93.2 1.280 |
| \( \hat{\kappa}_{\text{DPI}} \) | 4.7 | 0.004 93.9 0.654 | 0.003 94.40 0.836 | 0.022 93.5 1.080 |
| 2 | 2.0 | 0.030 93.6 0.509 | 0.003 95.20 0.519 | 0.000 94.5 0.538 |
| | 3.0 | 0.041 93.2 0.611 | 0.040 94.30 0.589 | 0.073 92.2 0.642 |
| \( \kappa_{\text{IMSE}} \) | 4.0 | 0.048 93.4 0.773 | 0.040 94.30 0.822 | 0.040 93.8 0.874 |
| | 5.0 | 0.013 94.0 0.938 | 0.001 94.30 0.943 | 0.013 93.7 1.180 |
| | 6.0 | 0.013 94.3 1.070 | 0.023 94.10 1.240 | 0.006 93.1 1.460 |
| \( \hat{\kappa}_{\text{ROT}} \) | 5.5 | 0.002 94.3 0.998 | 0.009 93.80 1.080 | 0.011 93.4 1.310 |
| \( \hat{\kappa}_{\text{DPI}} \) | 4.7 | 0.005 94.2 0.890 | 0.011 94.60 0.908 | 0.022 93.7 1.090 |
| 3 | 2.0 | 0.258 26.7 0.388 | 0.190 59.30 0.435 | 0.029 94.4 0.550 |
| | 3.0 | 0.070 89.9 0.742 | 0.022 93.00 0.646 | 0.064 93.0 0.626 |
| \( \kappa_{\text{IMSE}} \) | 4.0 | 0.045 92.9 0.560 | 0.027 95.20 0.681 | 0.040 94.0 0.814 |
| | 5.0 | 0.026 93.9 1.050 | 0.023 94.20 0.868 | 0.023 93.5 0.986 |
| | 6.0 | 0.005 93.9 0.783 | 0.008 94.50 0.889 | 0.020 93.4 1.020 |
| \( \hat{\kappa}_{\text{ROT}} \) | 5.5 | 0.014 93.7 0.928 | 0.016 94.20 0.879 | 0.021 93.5 1.000 |
| \( \hat{\kappa}_{\text{DPI}} \) | 4.7 | 0.005 93.5 0.908 | 0.008 94.20 0.814 | 0.028 93.5 0.937 |
Table SA-9: Tuning parameters, $n = 1000$, 5000 Replications

| Model 1, $\kappa_{\text{IMSE}} = 3.4$ | Min | 1Q | Med | Mean | 3Q | Max | Sd |
|--------------------------------------|-----|----|-----|------|----|-----|----|
| $\kappa_{\text{ROT}}$               | 3   | 4  | 5   | 4.88 | 5  | 8   | 0.78 |
| $\kappa_{\text{DPI}}$               | 3   | 4  | 5   | 5.11 | 6  | 10  | 0.99 |

| Model 2, $\kappa_{\text{IMSE}} = 7.2$ | Min | 1Q | Med | Mean | 3Q | Max | Sd |
|--------------------------------------|-----|----|-----|------|----|-----|----|
| $\kappa_{\text{ROT}}$               | 6   | 8  | 8   | 8.31 | 9  | 10  | 0.63 |
| $\kappa_{\text{DPI}}$               | 6   | 8  | 8   | 8.53 | 9  | 12  | 0.85 |

| Model 3, $\kappa_{\text{IMSE}} = 11.9$ | Min | 1Q | Med | Mean | 3Q | Max | Sd |
|---------------------------------------|-----|----|-----|------|----|-----|----|
| $\kappa_{\text{ROT}}$                | 6   | 7  | 7   | 7.30 | 8  | 9   | 0.51 |
| $\kappa_{\text{DPI}}$                | 8   | 10 | 10  | 10.37| 11 | 13  | 1.14 |

| Model 4, $\kappa_{\text{IMSE}} = 4.4$ | Min | 1Q | Med | Mean | 3Q | Max | Sd |
|--------------------------------------|-----|----|-----|------|----|-----|----|
| $\kappa_{\text{ROT}}$               | 5   | 5  | 5   | 5.46 | 6  | 6   | 0.50 |
| $\kappa_{\text{DPI}}$               | 4   | 4  | 5   | 4.69 | 5  | 6   | 0.46 |
Table SA-10: Uniform Results, Model 1, $n = 1000$, 5000 Replications.

|                | Evenly-spaced |             | Quantile-spaced |             |
|----------------|---------------|--------------|-----------------|--------------|
|                | CP  | ACE | AW  | UCR |          | CP  | ACE | AW  | UCR |
| **Plug-in, $j = 0$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 0.92 | 0.017 | 0.384 | 79.68 | 0.92 | 0.017 | 0.383 | 79.96 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.005 | 0.469 | 92.22 | 1.00 | 0.005 | 0.468 | 92.22 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.006 | 0.478 | 91.40 | 1.00 | 0.006 | 0.478 | 91.56 |
| **Plug-in, $j = 1$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 1.00 | 0.005 | 0.426 | 93.86 | 1.00 | 0.005 | 0.426 | 93.52 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.506 | 93.72 | 1.00 | 0.004 | 0.505 | 93.44 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.003 | 0.514 | 93.44 | 1.00 | 0.003 | 0.515 | 93.44 |
| **Plug-in, $j = 2$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 1.00 | 0.005 | 0.443 | 94.06 | 1.00 | 0.005 | 0.443 | 94.02 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.003 | 0.536 | 93.78 | 1.00 | 0.003 | 0.536 | 93.72 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.003 | 0.546 | 93.34 | 1.00 | 0.003 | 0.547 | 93.50 |
| **Plug-in, $j = 3$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 1.00 | 0.008 | 0.413 | 88.98 | 1.00 | 0.008 | 0.413 | 88.76 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.499 | 93.56 | 1.00 | 0.004 | 0.498 | 93.34 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.509 | 93.04 | 1.00 | 0.004 | 0.509 | 93.08 |
| **Bootstrap, $j = 0$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 0.92 | 0.017 | 0.382 | 79.50 | 0.92 | 0.017 | 0.382 | 79.62 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.006 | 0.466 | 91.82 | 1.00 | 0.006 | 0.464 | 91.74 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.006 | 0.475 | 91.16 | 1.00 | 0.006 | 0.474 | 90.92 |
| **Bootstrap, $j = 1$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 1.00 | 0.005 | 0.424 | 93.64 | 1.00 | 0.005 | 0.423 | 93.54 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.501 | 93.06 | 1.00 | 0.004 | 0.501 | 92.76 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.510 | 92.76 | 1.00 | 0.004 | 0.510 | 92.60 |
| **Bootstrap, $j = 2$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 1.00 | 0.005 | 0.440 | 94.04 | 1.00 | 0.005 | 0.440 | 93.82 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.531 | 93.20 | 1.00 | 0.004 | 0.531 | 93.12 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.541 | 92.74 | 1.00 | 0.003 | 0.542 | 93.04 |
| **Bootstrap, $j = 3$** |     |     |     |     |          |     |     |     |     |
| $\hat{\kappa}_{\text{MSE}}$ | 1.00 | 0.008 | 0.411 | 88.36 | 1.00 | 0.008 | 0.411 | 88.48 |
| $\hat{\kappa}_{\text{ROT}}$   | 1.00 | 0.004 | 0.495 | 93.22 | 1.00 | 0.004 | 0.494 | 92.76 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00 | 0.004 | 0.504 | 92.58 | 1.00 | 0.004 | 0.504 | 92.56 |
Table SA-11: Uniform Results, Model 2, $n = 1000, 5000$ Replications.

|                      | Evenly-spaced | Quantile-spaced |
|----------------------|---------------|-----------------|
|                      | CP  | ACE  | AW  | UCR | CP  | ACE  | AW  | UCR |
| **Plug-in, $j = 0$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 0.95 | 0.013 | 0.556 | 68.54 | 0.95 | 0.014 | 0.554 | 68.68 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.006 | 0.604 | 85.58 | 1.00 | 0.006 | 0.601 | 84.36 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.006 | 0.611 | 84.46 | 1.00 | 0.006 | 0.611 | 84.52 |
| **Plug-in, $j = 1$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 1.00 | 0.004 | 0.588 | 91.84 | 1.00 | 0.004 | 0.588 | 90.84 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.004 | 0.634 | 90.74 | 1.00 | 0.004 | 0.634 | 90.76 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.003 | 0.642 | 91.28 | 1.00 | 0.003 | 0.644 | 91.46 |
| **Plug-in, $j = 2$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 1.00 | 0.003 | 0.632 | 92.40 | 1.00 | 0.003 | 0.631 | 92.22 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.003 | 0.685 | 92.26 | 1.00 | 0.003 | 0.684 | 92.20 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.003 | 0.693 | 92.32 | 1.00 | 0.003 | 0.695 | 92.22 |
| **Plug-in, $j = 3$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 1.00 | 0.006 | 0.587 | 84.36 | 1.00 | 0.006 | 0.586 | 84.56 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.003 | 0.636 | 90.32 | 1.00 | 0.003 | 0.635 | 89.94 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.004 | 0.643 | 90.00 | 1.00 | 0.004 | 0.645 | 89.92 |
| **Bootstrap, $j = 0$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 0.95 | 0.014 | 0.551 | 67.30 | 0.95 | 0.015 | 0.549 | 67.24 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.006 | 0.597 | 84.68 | 1.00 | 0.007 | 0.594 | 83.30 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.007 | 0.604 | 83.18 | 1.00 | 0.007 | 0.604 | 83.16 |
| **Bootstrap, $j = 1$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 1.00 | 0.004 | 0.581 | 91.14 | 1.00 | 0.005 | 0.581 | 89.86 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.004 | 0.626 | 89.74 | 1.00 | 0.004 | 0.626 | 89.74 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.004 | 0.633 | 90.12 | 1.00 | 0.004 | 0.635 | 90.36 |
| **Bootstrap, $j = 2$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 1.00 | 0.004 | 0.624 | 91.58 | 1.00 | 0.004 | 0.623 | 91.44 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.004 | 0.675 | 91.22 | 1.00 | 0.003 | 0.674 | 91.24 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.003 | 0.683 | 91.32 | 1.00 | 0.003 | 0.685 | 91.46 |
| **Bootstrap, $j = 3$** |     |      |     |     |     |      |     |     |
| $\kappa_{\text{IMSE}}$     | 1.00 | 0.007 | 0.581 | 83.18 | 1.00 | 0.007 | 0.580 | 83.64 |
| $\hat{k}_{\text{ROT}}$    | 1.00 | 0.004 | 0.627 | 89.32 | 1.00 | 0.004 | 0.626 | 88.96 |
| $\hat{k}_{\text{DPI}}$    | 1.00 | 0.004 | 0.634 | 88.74 | 1.00 | 0.004 | 0.636 | 88.94 |
Table SA-12: Uniform Results, Model 3, $n = 1000, 5000$ Replications.

|                      | Evenly-spaced | Quantile-spaced |
|----------------------|---------------|-----------------|
|                      | CP  | ACE | AW  | UCR | CP  | ACE | AW  | UCR |
| Plug-in, $j = 0$     |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 0.88| 0.033| 0.731| 8.98| 0.87| 0.027| 0.726| 26.22 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.73| 0.082| 0.571| 38.70| 0.62| 0.108| 0.569| 10.42 |
| $\hat{\kappa}_{\text{DPI}}$   | 0.82| 0.046| 0.677| 35.62| 0.78| 0.048| 0.679| 22.64 |
| Plug-in, $j = 1$     |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00| 0.003| 0.754| 91.14| 1.00| 0.006| 0.754| 81.42 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.62| 0.152| 0.607| 24.18| 0.55| 0.151| 0.607| 12.32 |
| $\hat{\kappa}_{\text{DPI}}$   | 0.84| 0.032| 0.705| 58.56| 0.84| 0.026| 0.709| 55.24 |
| Plug-in, $j = 2$     |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00| 0.002| 0.822| 91.54| 1.00| 0.003| 0.820| 90.30 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.63| 0.067| 0.649| 27.62| 0.64| 0.066| 0.648| 25.28 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00| 0.005| 0.764| 84.12| 1.00| 0.005| 0.768| 84.66 |
| Plug-in, $j = 3$     |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 0.91| 0.012| 0.765| 60.42| 0.94| 0.010| 0.761| 65.40 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.70| 0.081| 0.603| 11.34| 0.66| 0.097| 0.603| 4.96  |
| $\hat{\kappa}_{\text{DPI}}$   | 0.83| 0.023| 0.710| 48.82| 0.82| 0.024| 0.713| 45.12 |
| Bootstrap, $j = 0$   |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 0.88| 0.035| 0.718| 7.46 | 0.85| 0.029| 0.713| 24.20 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.71| 0.084| 0.565| 37.90| 0.62| 0.110| 0.564| 10.06 |
| $\hat{\kappa}_{\text{DPI}}$   | 0.82| 0.048| 0.668| 34.78| 0.76| 0.050| 0.669| 21.16 |
| Bootstrap, $j = 1$   |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00| 0.003| 0.741| 89.80| 1.00| 0.006| 0.741| 79.26 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.60| 0.154| 0.600| 23.40| 0.54| 0.155| 0.599| 11.82 |
| $\hat{\kappa}_{\text{DPI}}$   | 0.84| 0.033| 0.693| 57.44| 0.84| 0.027| 0.697| 53.56 |
| Bootstrap, $j = 2$   |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 1.00| 0.003| 0.806| 90.16| 1.00| 0.003| 0.804| 88.90 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.60| 0.071| 0.641| 26.88| 0.61| 0.069| 0.640| 24.34 |
| $\hat{\kappa}_{\text{DPI}}$   | 1.00| 0.006| 0.751| 82.16| 1.00| 0.006| 0.754| 82.66 |
| Bootstrap, $j = 3$   |     |     |     |     |     |     |     |     |
| $\hat{\kappa}_{\text{IMSE}}$ | 0.91| 0.013| 0.750| 56.64| 0.90| 0.011| 0.747| 62.38 |
| $\hat{\kappa}_{\text{ROT}}$   | 0.69| 0.084| 0.596| 10.00| 0.64| 0.100| 0.596| 4.40  |
| $\hat{\kappa}_{\text{DPI}}$   | 0.82| 0.025| 0.698| 46.46| 0.81| 0.026| 0.702| 42.60 |
Table SA-13: Uniform Results, Model 4, $n = 1000$, 5000 Replications.

| Plug-in, $j = 0$ | Evenly-spaced | Quantile-spaced |
|------------------|---------------|-----------------|
|                  | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 1.00 | 0.007 | 1.086 | 57.32 | 0.98 | 0.008 | 1.096 | 52.30 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.002 | 1.368 | 82.02 | 1.00 | 0.003 | 1.382 | 80.04 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.004 | 1.221 | 71.34 | 1.00 | 0.005 | 1.241 | 69.68 |

| Plug-in, $j = 1$ | Evenly-spaced | Quantile-spaced |
|------------------|---------------|-----------------|
|                  | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 1.00 | 0.002 | 1.256 | 86.48 | 1.00 | 0.002 | 1.253 | 86.94 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.002 | 1.590 | 86.54 | 1.00 | 0.002 | 1.588 | 86.52 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.002 | 1.399 | 87.22 | 1.00 | 0.002 | 1.405 | 87.10 |

| Plug-in, $j = 2$ | Evenly-spaced | Quantile-spaced |
|------------------|---------------|-----------------|
|                  | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 1.00 | 0.002 | 1.356 | 87.86 | 1.00 | 0.002 | 1.357 | 88.04 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.002 | 1.741 | 85.42 | 1.00 | 0.002 | 1.748 | 85.40 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.002 | 1.392 | 87.80 | 1.00 | 0.002 | 1.537 | 87.64 |

| Plug-in, $j = 3$ | Evenly-spaced | Quantile-spaced |
|------------------|---------------|-----------------|
|                  | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 1.00 | 0.003 | 1.133 | 80.32 | 1.00 | 0.003 | 1.158 | 78.64 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.002 | 1.528 | 86.08 | 1.00 | 0.002 | 1.550 | 86.02 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.002 | 1.352 | 83.90 | 1.00 | 0.002 | 1.379 | 84.22 |

| Bootstrap, $j = 0$ | Evenly-spaced | Quantile-spaced |
|-------------------|---------------|-----------------|
|                   | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 0.98 | 0.009 | 1.042 | 47.92 | 0.95 | 0.011 | 1.052 | 43.14 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.004 | 1.296 | 73.72 | 1.00 | 0.004 | 1.310 | 70.58 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.006 | 1.165 | 62.46 | 1.00 | 0.006 | 1.184 | 60.30 |

| Bootstrap, $j = 1$ | Evenly-spaced | Quantile-spaced |
|-------------------|---------------|-----------------|
|                   | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 1.00 | 0.003 | 1.212 | 82.16 | 1.00 | 0.003 | 1.209 | 82.60 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.003 | 1.507 | 79.60 | 1.00 | 0.003 | 1.506 | 80.32 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.003 | 1.340 | 81.54 | 1.00 | 0.003 | 1.346 | 81.42 |

| Bootstrap, $j = 2$ | Evenly-spaced | Quantile-spaced |
|-------------------|---------------|-----------------|
|                   | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 1.00 | 0.002 | 1.305 | 83.14 | 1.00 | 0.002 | 1.305 | 83.88 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.003 | 1.638 | 77.16 | 1.00 | 0.003 | 1.644 | 77.10 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.002 | 1.449 | 81.04 | 1.00 | 0.002 | 1.462 | 80.58 |

| Bootstrap, $j = 3$ | Evenly-spaced | Quantile-spaced |
|-------------------|---------------|-----------------|
|                   | CP  | ACE | AW   | UCE  | CP  | ACE | AW   | UCE  |
| $\kappa^*_{\text{MSE}}$ | 1.00 | 0.004 | 1.094 | 74.42 | 1.00 | 0.004 | 1.118 | 72.30 |
| $\hat{\kappa}^*_{\text{ROT}}$ | 1.00 | 0.003 | 1.442 | 77.72 | 1.00 | 0.003 | 1.465 | 78.12 |
| $\hat{\kappa}^*_{\text{DPI}}$ | 1.00 | 0.003 | 1.290 | 77.44 | 1.00 | 0.003 | 1.317 | 77.64 |
References

[1] Agarwal, G. G., and Studden, W. (1980), “Asymptotic Integrated Mean Square Error Using Least Squares and Bias Minimizing Splines,” Annals of Statistics, 8, 1307–1325.

[2] Barrow, D. L., and Smith, P. W. (1978), “Asymptotic Properties of Best $L_2[0,1]$ Approximation by Splines with Variable Knots,” Quarterly of Applied Mathematics, 36, 293–304.

[3] Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), “Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results,” Journal of Econometrics, 186, 345–366.

[4] Bhatia, R. (2013), Matrix Analysis, Springer.

[5] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), Classification and regression trees, CRC press.

[6] Buse, A., and Lim, L. (1977), “Cubic splines as a special case of restricted least squares,” Journal of the American Statistical Association, 72, 64–68.

[7] Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018), “Coverage Error Optimal Confidence Intervals,” working paper, University of Michigan.

[8] ——— (2018), “On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference,” Journal of the American Statistical Association, forthcoming.

[9] Calonico, S., Cattaneo, M. D., and Titiumik, R. (2015), “Optimal Data-Driven Regression Discontinuity Plots,” Journal of the American Statistical Association, 110, 1753–1769.

[10] Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Schaumburg, E. (2018), “Characteristic-Sorted Portfolios: Estimation and Inference,” working paper.

[11] 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 (Advances in Econometrics, vol. 27), ed. D. Drukker, Emerald Group Publishing, pp. 93–127.
12 —— (2013), “Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators,” *Journal of Econometrics*, 174, 127–143.

13 Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2018), “lspartition: Partitioning-Based Least Squares Regression,” working paper, University of Michigan.

14 Chen, X. (2007), “Large Sample Sieve Estimation of Semi-Nonparametric Models,” in *Handbook of Econometrics, Volume VI*, eds. J. J. Heckman and E. Leamer, New York: Elsevier Science B.V., pp. 5549–5632.

15 Chen, X., and Christensen, T. M. (2015), “Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions,” *Journal of Econometrics*, 188, 447–465.

16 Chernozhukov, V., Chetverikov, D., and Kato, K. (2014), “Gaussian Approximation of Suprema of Empirical Processes,” *Annals of Statistics*, 42, 1564–1597.

17 —— (2014), “Anti-Concentration and Honest Adaptive Confidence Bands,” *Annals of Statistics*, 42, 1787–1818.

18 —— (2015), “Comparison and Anti-concentration Bounds for Maxima of Gaussian Random Vectors,” *Probability Theory and Related Fields*, 162, 47–70.

19 —— (2016), “Empirical and Multiplier Bootstraps for Suprema of Empirical Processes of Increasing Complexity, and Related Gaussian Couplings,” *Stochastic Processes and their Applications*, 126, 3632–3651.

20 Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018), “Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments,” *arXiv preprint arXiv:1712.04802*.

21 Chernozhukov, V., Lee, S., and Rosen, A. M. (2013), “Intersection bounds: estimation and inference,” *Econometrica*, 81, 667–737.

22 Chernozhukov, V., Newey, W. K., and Santos, A. (2015), “Constrained Conditional Moment Restriction Models,” *arXiv:1509.06311*. 

132
[23] Chui, C. K. (2016), *An introduction to wavelets*, Elsevier.

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

[25] Davydov, O. (2001), “Stable Local Bases for Multivariate Spline Spaces,” *Journal of Approximation Theory*, 111, 267–297.

[26] Demko, S. (1977), “Inverses of Band Matrices and Local Convergence of Spline Projections,” *SIAM Journal on Numerical Analysis*, 14, 616–619.

[27] Donoho, D. L. (1994), “Statistical Estimation and Optimal Recovery,” *Annals of Statistics*, 22, 238–270.

[28] Eggermont, P. P. B., and LaRiccia, V. N. (2009), *Maximum Penalized Likelihood Estimation: Regression*, New York, NY: Springer.

[29] Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, New York: Chapman & Hall/CRC.

[30] Fithian, W., Sun, D., and Taylor, J. (2017), “Optimal inference after model selection,” *arXiv preprint arXiv:1410.2597*.

[31] Gazzah, H., Regalia, P. A., and Delmas, J.-P. (2001), “Asymptotic eigenvalue distribution of block Toeplitz matrices and application to blind SIMO channel identification,” *IEEE Transactions on Information Theory*, 47, 1243–1251.

[32] Genovese, C., and Wasserman, L. (2008), “Adaptive confidence bands,” *Annals of Statistics*, 36, 875–905.

[33] Genovese, C. R., and Wasserman, L. (2005), “Confidence Sets for Nonparametric Wavelet Regression,” *Annals of statistics*, 33, 698–729.

[34] Grenander, U., and Szegö, G. (2001), *Toeplitz forms and their applications*, Vol. 321, University of California Press.

[35] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002), *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag.
[36] Hall, P., and Horowitz, J. (2013), “A simple bootstrap method for constructing nonparametric confidence bands for functions,” *The Annals of Statistics*, 41, 1892–1921.

[37] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2012), *Wavelets, Approximation, and Statistical Applications*, Springer.

[38] Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning*, Springer Series in Statistics, New York: Springer-Verlag.

[39] Hiriart-Urruty, J.-B., and Ye, D. (1995), “Sensitivity analysis of all eigenvalues of a symmetric matrix,” *Numerische Mathematik*, 70, 45–72.

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

[41] Huang, J. Z. (1998), “Projection Estimation in Multiple Regression with Application to Functional ANOVA Models,” *Annals of Statistics*, 26, 242–272.

[42] ——— (2003), “Local Asymptotics for Polynomial Spline Regression,” *Annals of Statistics*, 31, 1600–1635.

[43] Koltchinskii, V. I. (1994), “Komlós-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions,” *Journal of Theoretical Probability*, 7, 73–118.

[44] Komlós, J., Major, P., and Tusnády, G. (1975), “An approximation of partial sums of independent RV’s, and the sample DF. I,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32, 111–131.

[45] ——— (1976), “An approximation of partial sums of independent RV’s, and the sample DF. II,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34, 33–58.

[46] Massart, P. (1989), “Strong approximation for multivariate empirical and related processes, via KMT constructions,” *Annals of probability*, 17, 266–291.

[47] Meyer, Y. (1995), *Wavelets and Operators*, Cambridge university press.

[48] Micula, G., and Micula, S. (1999), *Handbook of Splines*, Dordrecht: Springer Netherlands.
[49] Newey, W. K. (1997), “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, 79, 147–168.

[50] Nobel, A. (1996), “Histogram Regression Estimation Using Data-Dependent Partitions,” *Annals of Statistics*, 24, 1084–1105.

[51] Rio, E. (1994), “Local invariance principles and their application to density estimation,” *Probability Theory and Related Fields*, 98, 21–45.

[52] Ruppert, D., Wand, M. P., and Carroll, R. (2009), *Semiparametric Regression*, New York: Cambridge University Press.

[53] Sakhanenko, A. (1985), “Convergence Rate in the Invariance Principle for Non-identically Distributed Variables with Exponential Moments,” *Advances in Probability Theory: Limit Theorems for Sums of Random Variables*, 2–73.

[54] Schumaker, L. (2007), *Spline functions: basic theory*, Cambridge University Press.

[55] Stone, C. J. (1982), “Optimal Global Rates of Convergence for Nonparametric Regression,” *Annals of Statistics*, 10, 1040–1053.

[56] Sweldens, W., and Piessens, R. (1994), “Asymptotic error expansion of wavelet approximations of smooth functions II,” *Numerische Mathematik*, 68, 377–401.

[57] Tibshirani, R. J. (2014), “Adaptive Piecewise Polynomial Estimation via Trend Filtering,” *The Annals of Statistics*, 42, 285–323.

[58] Tropp, J. A. (2012), “User-friendly tail bounds for sums of random matrices,” *Foundations of computational mathematics*, 12, 389–434.

[59] Tukey, J. W. (1961), “Curves As Parameters, and Touch Estimation,” in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Vol. 1, pp. 681–694.

[60] Tyrtyshnikov, E. E. (1997), *A brief introduction to numerical analysis*, Boston, MA: Birkhuser Boston.

[61] Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer Science & Business Media.
[62] Yurinskii, V. V. (1978), “On the error of the Gaussian approximation for convolutions,” *Theory of Probability & Its Applications*, 22, 236–247.

[63] Zhai, A. (2018), “A High-Dimensional CLT in W2 Distance with Near Optimal Convergence Rate,” *Theoretical Probability and Related Fields*, forthcoming.

[64] Zhang, H., and Singer, B. H. (2010), *Recursive Partitioning and Applications*, Springer.

[65] Zhou, S., Shen, X., and Wolfe, D. (1998), “Local Asymptotics for Regression Splines and Confidence Regions,” *Annals of Statistics*, 26, 1760–1782.

[66] Zhou, S., and Wolfe, D. A. (2000), “On Derivative Estimation in Spline Regression,” *Statistica Sinica*, 10, 93–108.