Journal of Econometrics 174 (2013) 127–143

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators☆

Matias D. Cattaneo *, Max H. Farrell

Department of Economics, University of Michigan, United States

ARTICLE INFO

Article history:
Received 31 January 2011
Received in revised form
10 December 2012
Accepted 12 February 2013
Available online 24 February 2013

JEL classification:
C14
C21

Keywords:
Nonparametric estimation
Partitioning
Subclassification
Convergence rates
Bahadur representation
Asymptotic normality

ABSTRACT

This paper studies the asymptotic properties of partitioning estimators of the conditional expectation function and its derivatives. Mean-square and uniform convergence rates are established and shown to be optimal under simple and intuitive conditions. The uniform rate explicitly accounts for the effect of moment assumptions, which is useful in semiparametric inference. A general asymptotic integrated mean-square error approximation is obtained and used to derive an optimal plug-in tuning parameter selector. A uniform Bahadur representation is developed for linear functionals of the estimator. Using this representation, asymptotic normality is established, along with consistency of a standard-error estimator. The finite-sample performance of the partitioning estimator is examined and compared to other nonparametric techniques in an extensive simulation study.

1. Introduction

Nonparametric estimation of an unknown conditional expectation function and its derivatives is an important problem in econometrics (see, e.g., Ichimura and Todd, 2007 and references therein). In many applications the object of interest is a conditional expectation, its derivative, or functional thereof, while in other cases their nonparametric estimators are employed as a first step in a semiparametric procedure. The implementation of nonparametric estimators requires suitable large sample properties, including sufficiently rapid rates of convergence and known asymptotic distributions. Series- and kernel-based methods are examples whose properties are now well understood.

This paper studies the large sample properties of an estimator of the regression function and its derivatives known as *partitioning*. This estimation strategy is alternatively referred to as *blocking*, *subclassification*, or *stratification*. The estimator is constructed by partitioning the support of the conditioning variables into disjoint cells, which become smaller with the sample size, and within each the unknown regression function (and its derivatives) is approximated by linear least-squares using a fixed-order polynomial basis (other bases are possible). Consistent estimation is achieved as the cells become small enough to remove the error of the parametric approximation. For a recent textbook discussion of this estimation strategy, see Györfi et al. (2002, Chapter 4). After the necessary notation and assumptions are introduced, we provide a detailed comparison between partitioning estimators and other nonparametric estimators in Section 2.2.

The partitioning estimator, although simple and intuitive, has not received a thorough treatment in the econometrics or statistics literature. The available results typically concern mean-square rates for special cases (see, e.g., Kohler et al., 2006 and references therein). The main goal of this paper is to provide a general asymptotic treatment of partitioning estimators. Our analysis yields the following new insights. First, employing simple and intuitive sufficient conditions, in most cases weaker than those in the existing literature, mean-square and uniform convergence rates of the partitioning estimator are established and shown to be optimal. More generally, the uniform convergence rate explicitly

☆ We thank Yves Atchade, Xiaohong Chen, Richard Crump, Guido Imbens, Lutz Kilian, Jim Powell, seminar participants at Berkeley (2008) and Michigan (2009), and conference participants at the 2010 Advances in Econometrics Conference for comments. We are especially grateful to Victor Chernozhukov and Michael Jansson for detailed discussions and suggestions. We also thank the co-Editor, Peter Robinson, an Associate Editor, and four reviewers for detailed comments that greatly improved the paper. An earlier version of this paper was circulated under the title "Optimal Rates of Convergence and Asymptotic Normality of Block Regression Estimators". The authors gratefully acknowledge financial support from the National Science Foundation (SES 0921505 and SES 1122994).

* Correspondence to: Department of Economics, University of Michigan, 238 Lorch Hall, 611 Tappan Street, Ann Arbor, MI 48109-1220, United States. Tel.: +1 734 763 1306; fax: +1 734 764 2769.
E-mail address: cattaneo@umich.edu (M.D. Cattaneo).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.02.002

highlights a natural trade-off between moment assumptions and rate restrictions. Second, we characterize the leading terms of a conditional integrated mean-square error expansion and provide an optimal plug-in selector for the tuning parameter. Third, we derive a uniform Bahadur-type representation of linear functionals of the partitioning estimator, which is used to establish asymptotic normality under simple and intuitive conditions, with a suitable standard-error estimator. We cover both regular and irregular estimands. The applicability of the new results is illustrated with three examples: (i) derivative of the regression function at a point, (ii) partial and full means, and (iii) weighted average derivatives. Our results are also useful in other contexts in econometrics, as discussed in Section 1.1.

The paper proceeds as follows. In the remainder of this section we give the main motivations for our work, discussing in particular the importance of our results for both empirical and theoretical econometrics. Section 2 describes the partitioning estimator formally and also provides a comparison to other nonparametric estimators. Rates of convergence and a general integrated mean-square error expansion for the partitioning estimator are given in Section 3, while a Bahadur-type representation for linear functionals of the estimator and asymptotic normality with valid standard-error estimators are developed in Section 4. The results of a Monte Carlo study are summarized in Section 5. Finally, Section 6 concludes. Proofs are gathered in the Appendix. A supplement is available upon request containing detailed technical proofs and greatly expanded simulation results.

1.1. Motivation and preliminary discussion

Studying the large-sample properties of partitioning estimators may be interesting and important for a variety of reasons, some theoretical and others methodological. The partitioning estimator has specific features and asymptotic optimality properties that make it a useful addition to the econometrics toolkit: a complement, not a substitute, to the arsenal of nonparametric procedures commonly employed in econometrics. This estimator is attractive because it is very tractable and enjoys useful asymptotic representations leading to intuitive results, as well as other features that may be useful in econometric applications.

In particular, the partitioning estimator is potentially discontinuous in finite samples (just like nearest-neighbor estimators). This specific characteristic may be an advantage from a practical point of view, and could also lead to an estimator with desirable theoretical properties. The "binning" underlying the partitioning estimator arises naturally in many economic problems, where units (people, firms, etc.) in the same bin share similar economic behavior, and therefore partitioning-based inference procedures have been proposed to retain this natural interpretability (see applications below). From a theoretical perspective, we are interested in understanding the asymptotic properties of partitioning given its potential discontinuity in finite samples, and how they compare with results for other nonparametric procedures. We briefly discuss three implications of this discontinuity, which make the partitioning estimator theoretically and practically interesting in our view.

1. *Shape restrictions: convergence rates*. Nonparametric estimation typically assumes the estimand is smooth and most estimators are constructed imposing some of the underlying smoothness assumed. The partitioning estimator does not impose smoothness and therefore allows us to understand what effects imposing this shape restriction may have on asymptotic properties, which arguably is of theoretical interest. For instance, we establish optimal uniform convergence rates for partitioning, showing (by example) that imposing smoothness is not necessary for this result. This finding is not ex-ante obvious in our view, especially given other known results (see Section 2.2).
2. *Shape restrictions: bias–variance trade-off*. From a more practical perspective, removing the smoothness restriction may be interpreted as "freeing up restrictions". This means that the estimator will have a different bias–variance behavior in finite samples. To fix ideas, consider the linear partitioning and linear regression spline estimators of a univariate regression function. For each sample size, both are (piecewise linear) least squares fits, and differ only in that the spline is required to be continuous (see Section 2.2). That is, the linear spline is a restricted least squares problem compared to the partitioning estimate. From linear model results, it follows that the spline has larger bias than the partitioning estimate, but smaller variance.[1] Neither can be strictly superior or inferior, based on the usual bias–variance trade-off, and in fact the partitioning estimator may have better properties from a theoretical point of view.
3. *Diagnostics*. The potential discontinuity of the partitioning estimator in finite samples makes it a useful complement to existing smooth estimates already available in the literature. Specifically, the partitioning estimate may be used as a diagnostic check on the underlying smoothness assumptions imposed by other procedures, particularly if such assumptions are in question for a certain region of the support. Furthermore, the discontinuous partitioning estimate can be used to characterize the overall variability of the data relative to a smoothed-out estimate (see the regression discontinuity application below for an example).

Further motivation for our work stems from the role of partitioning estimators in empirical economics. Perhaps originating with the regressogram of Tukey (1947), partitioning-based procedures have been suggested in many contexts where "binning" has a natural interpretation, despite their formal properties being unknown in most cases. We close this section by briefly discussing four examples where partitioning estimation arises in econometrics: as an exploratory device, a nonparametric estimator, and two semiparametric cases.

Application (*Regression Discontinuity*)**.** Partitioning estimators are used heuristically in the regression discontinuity (RD) design for two purposes: (i) to plot a smoothed-out cloud of points along with global polynomial fits of the underlying regression function for control and treatment units, and (ii) to investigate whether the data suggests the presence of other possible discontinuities in the underlying conditional expectation of potential outcomes, as a form of falsification test. Imbens and Lemieux (2008) review the RD literature, and explicitly advocate partitioning (calling it a "histogram-type estimate") to assess the plausibility of the RD design. Our general result in Theorem 3 is employed in Calonico et al. (2012) to derive an optimal choice of partition length in this context, thereby providing a systematic way of plotting RD data. □

Application (*Portfolio Sorting*)**.** In understanding anomalous asset returns, a common approach is "portfolio sorts", in which assets are partitioned into homogeneous groups according to characteristics that may drive anomalies. A number of informal and formal analyses are then performed on the sorted assets, including tests of monotonicity and comparison of extremes. See, e.g., Fama and French (2008). Our results may be used to develop formal nonparametric inference for this type of application. □

[1] This claim assumes that the estimators are misspecified in finite samples, as is the case with nonparametric estimators in general. This remains true when comparing to kernel-based estimators.

Application (*Subclassification on Observables*)**.** In many econometric contexts units are divided in groups according to their observed characteristics, and then inference is conducted first within each subclass and then overall. Under an ignorability assumption, for example, subclassification (or partitioning) has been proposed in multiple forms to estimate treatment effects. Imbens and Wooldridge (2009) give a recent survey of the program evaluation literature, which includes several examples of partitioning-based procedures. Despite many such procedures have been proposed and used in empirical work, there is a paucity of rigorous asymptotic theory. The theoretical results presented herein may be used to characterize the large-sample properties of those partitioning-based procedures. For instance, in Cattaneo and Farrell (2011) we employ these results to formalize the properties of a partitioning-based estimator of the average treatment effect and dose-response function. □

Application (*Average Derivatives*)**.** Partitioning also yields simple and intuitive estimators for derivatives of the regression function. Based on this observation, Banerjee (2007) recently proposed a partitioning-based semiparametric average derivative estimator. In Section 4 we discuss an alternative semiparametric estimator for (weighted) average derivatives, and establish its asymptotic properties under general, easy-to-interpret sufficient conditions. □

2. The partitioning estimator

2.1. Setup and estimator

Before describing the estimator we introduce some notation. For a scalar, vector, or matrix A we denote $|A| = \sqrt{\text{tr}(A'A)}$. For a multi-index $k = (k_1, k_2, \ldots, k_d) \in \mathbb{Z}_+^d$, we let $[k] = k_1 + \cdots + k_d$, $x^k = x_1^{k_1} \cdots x_d^{k_d}$ for $x = (x_1, \ldots, x_d)' \in \mathbb{R}^d$, and $\partial^k h(x) = \partial^{[k]} h(x)/(\partial^{k_1} x_1 \cdots \partial^{k_d} x_d)$ for smooth enough function $h(x)$.

We impose the following assumption on the data generating process throughout.

Assumption 1. (a) $(Y_1, X_1'), \ldots, (Y_n, X_n')$ is an i.i.d. sample from (Y, X'), and $X \in \mathcal{X}$ is continuously distributed with Lebesgue density $f(x)$.
(b) $\mathcal{X} \subset \mathbb{R}^d$ is given by $\mathcal{X} = \times_{\ell=1}^d \mathcal{X}_\ell$, a Cartesian product of compact, convex intervals.
(c) $\mathbb{E}[|Y|^{2+\eta} \mid X]$ is bounded for some $\eta \geq 0$.
(d) $f(x)$ is bounded and bounded away from zero on $\mathcal{X}$.
(e) $\mu(x) = \mathbb{E}[Y|X = x]$ is S-times continuously differentiable on (an extension of) $\mathcal{X}$, and satisfies $\left|\partial^m \mu(x) - \partial^m \mu(x')\right| \leq C \left|x - x'\right|^\alpha$, for some constants $C > 0$ and $\alpha \in (0, 1]$, and all $x, x' \in \mathcal{X}$ and $[m] = S$.

We discuss the salient features of this assumption in the following remarks.

- Part (a) restricts attention to cross-sectional contexts with continuous regressors. Our results can be extended to cover some form of time-dependent data, or to include discrete regressors by working conditionally, although we do not consider these extensions here to simplify the discussion and notation.
- Part (b) requires regressors with compact support. The assumed rectangular structure is without loss of generality for most of the results presented here. The compact support assumption has the main advantage of allowing for the density $f(x)$ to be bounded away from zero on the full support of X, but has the potential drawback of introducing bias at the boundary of the support. This assumption is also imposed for nonparametric series estimators (Newey, 1997) and nonparametric local polynomials (Fan and Gijbels, 1996), but it can be relaxed in semiparametric inference by considering weaker (weighted) norms (Chen, 2007). In this paper we only focus on the conventional mean-square and uniform norms.
  This assumption is important because it can affect the attainable convergence rates for nonparametric regression estimators in general. Specifically, in the case of mean-square convergence, Kohler et al. (2009) show that it is possible to attain Stone's (1982) optimal L_2 convergence rate even without compactness as long as certain moment conditions hold, and Kohler et al. (2006) show that a cleverly constructed special partitioning estimator attains this rate. In the case of the uniform convergence rate, it appears to be an open question whether Stone's (1982) bound is achievable without compactness.
- Part (c) allows for the case of $\eta = 0$ (i.e., bounded second conditional moment only), and the generality will be useful in the derivation of the uniform convergence rate.
- Part (d) ensures that all cells in the partition will contain enough observations asymptotically, and appears difficult to relax without affecting the rates of convergence.
- Part (e) is a classical smoothness condition controlling the amount of bias reduction possible, when coupled with an appropriate basis choice employed within each cell.

To describe the nonparametric procedure, we first give a precise description of the partitioning scheme. For a sequence $J_n \to \infty$ as $n \to \infty$, partition each $\mathcal{X}_\ell$ into the J_n disjoint intervals $[p_{\ell,j-1}, p_{\ell,j})$, $j = 1, \ldots, J_n - 1$, and $[p_{\ell,J_n-1}, p_{\ell,J_n}]$, with $p_{\ell,j-1} < p_{\ell,j}$ for all j. The complete partition of $\mathcal{X}$ consists of the J_n^d cells formed as Cartesian products of all such intervals. Let $P_j \subset \mathbb{R}^d$ denote a generic cell of the partition, $j = 1, \ldots, J_n^d$, and for $x \in \mathbb{R}^d$, let $\mathbb{1}_{P_j}(x)$ be the indicator for $x \in P_j$. Throughout, we suppress the dependence on n for notational convenience: all aspects of the partition implicitly depend on n.

To guarantee that each cell is well defined we require that $\left|p_{\ell,j} - p_{\ell,j-1}\right| \asymp J_n^{-1}$ for all $\ell = 1, \ldots, d$ and $j = 1, \ldots, J_n$, where for scalars a and b, $a \asymp b$ denotes that $C_* b \leq a \leq C^* b$ for positive constants C_* and C^* that do not depend on $j = 1, \ldots, J_n$ nor n. Hence, by construction the partition satisfies $\text{vol}(P_j) \asymp J_n^{-d}$, where $\text{vol}(P_j)$ denotes the volume of cell P_j. A simple, natural partitioning scheme meeting this requirement is evenly dividing the support of each covariate, although other possibilities are allowed so long as all intervals decrease proportionally to J_n.

Within each cell the unknown conditional expectation is approximated by solving a least squares problem. For fixed $K \in \mathbb{N}$, let $r(x_\ell) = (1, x_\ell, x_\ell^2, \ldots, x_\ell^{K-1})'$ denote the vector of powers up to degree $K - 1$ on a single covariate $x_\ell \in \mathcal{X}_\ell$. Let $R(x)$ represent a column vector containing the complete polynomial basis of degree $K - 1$ formed as the Kronecker product of the $r(x_\ell)$, discarding terms with degree exceeding $K - 1$. Thus, each element of $R(x)$ is given by $x^k = x_1^{k_1} \cdots x_d^{k_d}$ for a unique $k \in \{k \in \mathbb{Z}_+^d : [k] \leq K - 1\}$. We assume $R(x)$ is ordered ascendingly in $k \in \mathbb{Z}_+^d$ and $\ell = 1, \ldots, d$. For example, if $K = 1$ then $R(x) = (1)$ and sample means are fitted in each cell, while if $K = 2$ then $R(x) = (1, x_1, \ldots, x_d)'$, corresponding to ordinary linear least squares. This construction is explicitly meant to cover the general, unrestricted case, although in applications other bases may be of interest. For example, if $\mu(x)$ additively separable, then the interactions between covariates may be excluded from the basis, leading to a simpler least squares problem. This additional flexibility is useful, for example, in estimation via control functions. The goal of this construction is to ensure that $R(x)$ is flexible enough to remove bias up to the appropriate order (see Lemma A.2).

The choice of K is intimately related to bias reduction. Setting a higher K allows for a more flexible functional form within each cell and hence lower bias, provided the underlying function is

sufficiently smooth.[2] In this sense, the partitioning scheme and the choice of K play the same role for the partitioning estimator that the choice of specific higher-order kernel plays in kernel-based estimation, while the choice of J_n is analogous to the choice of bandwidth in a kernel context. The partitioning scheme and (fixed) K represent the smoothing parameter, and $J_n \to \infty$ is the tuning parameter of the nonparametric procedure.

Let $R_j(x) = \mathbb{1}_{P_j}(x)R(x)$ denote basis restricted to the cell containing x. Using this notation the partitioning regression estimator of order K is given by:

$$\hat{\mu}(x) = \sum_{j=1}^{J_n^d} R_j(x)'\hat{\beta}_j, \qquad \hat{\beta}_j = \left(R_j'R_j\right)^- R_j'Y,$$
$$R_j = \left(R_j(X_1), \ldots, R_j(X_n)\right)', \; Y = (Y_1, \ldots, Y_n)', \tag{1}$$

where A^- denotes any generalized symmetric inverse. Under regularity conditions given below, and with proper scaling, the matrix $R_j'R_j$ will be positive definite uniformly in j with probability approaching one (see Lemma A.4), and the standard inverse will exist. The structure given in Eq. (1) implies that $\hat{\mu}(x)$ is a (random) function that has at most finitely many discontinuities, is almost everywhere differentiable, and is of bounded variation. (Qualifiers such as "almost everywhere" and "for n large enough" are usually omitted for simplicity.)

To construct an estimator of the derivatives of $\mu(x)$, let $m \in \mathbb{Z}_+^d$ be a multi-index and $\partial^m\mu(x)$ denote a partial derivative of order $[m]$. An intuitive estimator of $\partial^m\mu(x)$ is

$$\widehat{\partial^m\mu(x)} \equiv \partial^m\hat{\mu}(x) \equiv \sum_{j=1}^{J_n^d} \mathbb{1}_{P_j}(x)\left(\partial^m R(x)\right)'\hat{\beta}_j, \tag{2}$$

which we take as the definition throughout. In words, $\partial^m\hat{\mu}(x)$ is defined as the derivative of the estimated polynomial regression function, restricted to a particular cell containing x (as there are no boundary issues in differentiating $R(x)$). Because $\partial^m R(x)$ has zeros in some components, the resulting estimator employs a lower degree basis but the least squares problem within each cell is unaffected. This intuitively corresponds to estimating the rougher function $\partial^m\mu(x)$. The main results of this paper also cover the estimation of derivatives of the regression function, provided K is large enough.

2.2. Related literature

The partitioning estimator is closely related to, but different from, other nonparametric estimators available in the literature. In this section we describe how it relates to two common estimators: series and local polynomials.

From a series estimation perspective, the partitioning estimator may be recast as a linear sieve estimator. Define $\mathbf{R}_n(x) = (R_1(x)', \ldots, R_{J_n^d}(x)')'$ by collecting the bases over all J_n^d cells, and set $\mathbf{R}_n = [\mathbf{R}_n(X_1), \ldots, \mathbf{R}_n(X_n)]'$. The partition regression estimator can then be written as

$$\hat{\mu}(x) = \mathbf{R}_n(x)'\hat{\mathbf{B}}_n, \quad \hat{\mathbf{B}}_n = (\mathbf{R}_n'\mathbf{R}_n)^-\mathbf{R}_n'Y = (\hat{\beta}_1', \ldots, \hat{\beta}_{J_n^d}')'.$$

This representation implies that results available from the sieve estimation literature are in principle applicable to the partitioning estimator. But by exploiting the specific structure of the partitioning estimator we are able to obtain faster uniform convergence rates and new results such as derivative estimation, an integrated mean-square error expansion, and a Bahadur representation, while improving on rate restrictions and using simple primitive conditions, when compared to the results available in the general series estimation literature (Newey, 1997; de Jong, 2002; Belloni et al., 2012).

Regression splines are series estimators for which improved results are available (Huang, 2003). Partitioning estimators and polynomial splines are intuitively similar, but fundamentally different smoothing procedures. Both estimators rely on a refining partition of the support with fixed-order basis functions: an order K spline uses $K-1$ degree polynomials (in our notation). The key distinguishing characteristic is that at the cell boundaries (called "knots") the spline estimate is forced to be smooth whenever possible: a spline of order K is $(K-2)$-times differentiable at each knot. For precisely this reason splines are usually regarded as a "global" smoother. In contrast, partitioning estimators place no restriction on the behavior of the polynomials at the boundary of each cell, and hence the basis functions are truly local (and compactly supported). In this paper we show that the partitioning estimators have the same optimal L_2 convergence rate under the same rate restrictions as polynomial splines. We also show that the partitioning estimator achieves the optimal uniform convergence rate for levels and derivatives. General series estimators are only known to have suboptimal uniform rates (de Jong, 2002). (In personal communication, X. Chen shared preliminary work showing that spline least squares regression may achieve the optimal uniform convergence rate under certain conditions Chen and Huang, in preparation.)

Kernel-based local polynomials are another class of nonparametric estimators of the regression function and its derivatives. Partitioning estimators are conceptually (and numerically) distinct from the kernel-based local polynomial estimators discussed in Fan and Gijbels (1996) and the local polynomial estimators discussed in Eggermont and LaRiccia (2009, Chapter 16), which are also different from each other. These local polynomial approaches and the partitioning estimators differ in the way that observations are grouped: the local polynomial approaches use observations near the evaluation point, as determined by the choice of kernel and bandwidth, while partitioning estimators use observations within each cell, regardless of the particular evaluation point. This fact implies that partitioning estimators are naturally discontinuous while local polynomials are not. The partitioning estimator can be viewed as a local polynomial estimator with a particular variable bandwidth and a uniform spherical kernel.

To describe how the local polynomials and the partitioning estimators differ, consider the estimation of the regression function (a similar discussion applies to derivative estimation). Both estimation procedures solve the following weighted least-squares problem:

$$\hat{\beta}_n(x) = \underset{\beta\in\mathbb{R}^{\dim(B(\cdot))}}{\arg\min} \sum_{i=1}^n W_n(X_i, x)\left(Y_i - B(X_i, x)'\beta\right)^2,$$

where $W_n(X_i, x)$ is a non-negative weighting function and $B(X_i, x)$ is a choice of polynomial basis. Both local polynomials estimators mentioned above employ $W_n(X_i, x) = K((X_i - x)/h_n)/h_n$, for a fixed kernel function $K(\cdot)$ and a bandwidth sequence $h_n \to 0$. Moreover, the local polynomials in Fan and Gijbels (1996) are obtained by choosing $B(X, x) = R(X - x)$ and setting $\hat{\mu}(x) = e_1'\hat{\beta}_n(x)$ with $e_1 = (1, 0, 0, \ldots, 0)'$, while the local polynomial estimator in Eggermont and LaRiccia (2009, Chapter 16) employ $B(X, x) = R(X)$ and set $\hat{\mu}(x) = R(x)'\hat{\beta}_n(x)$. In contrast, the partitioning estimators use $W_n(X_i, x) = \sum_{j=1}^{J_n^d} \mathbb{1}_{P_j}(X_i)\mathbb{1}_{P_j}(x)$ and $B(X, x) = R(X)$, and set $\hat{\mu}(x) = R(x)'\hat{\beta}_n(x)$. Therefore, results from local polynomial methods cannot be applied directly to partitioning estimators.

[2] This bias reduction is asymptotic. Ruppert and Wand (1994, Remark 4) provide a very interesting discussion, for local polynomials, that highlights how in finite samples this smoothing-bias reduction could be more than offset by increased variability.

Finally, as a reviewer pointed out, Stone (1982, Section 3) also suggested another (hybrid) local polynomial procedure which bears some relation to the partitioning estimator studied here. Using our notation, Stone's estimators employ $W_n(X_i, x) = \sum_{j=1}^{J_n^d} \mathbb{1}_{P_j}(X_i)\mathbb{1}\{j : |z - x| \leq h_n, \forall z \in P_j\}/N_j$, where $N_j = \sum_{i=1}^n \mathbb{1}_{P_j}(X_i)$ is the number of observations in P_j. This estimator uses all (data in) cells falling *completely* within an h_n-ball around the evaluation point x, in contrast to partitioning which only considers observations in the cell P_j. Moreover, Stone's estimator necessitates the choice of two tuning parameters, J_n and h_n, which are required to satisfy $h_n J_n \to \infty$. The rate restriction that the cells are required to shrink faster than the bandwidth implies that the number of cells in each h_n-ball tends to infinity, and hence asymptotically the weighting is constant in the h_n-ball and symmetric about x, just like a classical local polynomial with a spherical uniform kernel with bandwidth h_n, and not like the partitioning estimators considered here.

In Section 6 we provide further discussion of the potential advantages and disadvantages of the partitioning estimators when compared to series, kernel and nearest-neighbor estimators.

3. Convergence rates and integrated mean-square expansion

Some further notation is necessary to state the results. Let $a \wedge b = \min\{a, b\}$, $a, b \in \mathbb{R}$. For a function $h(\cdot)$ let $\|h\|_p^p = \int_{\mathcal{X}} |h(x)|^p f(x)dx$ and $\|h\|_\infty = \sup_{x\in\mathcal{X}} |h(x)|$ denote the L_p and L_∞ norms; function arguments are suppressed if there is no confusion.

3.1. Rates of convergence

The following theorem gives the L_2 convergence rate for the partitioning estimate of the regression function and its derivatives.

Theorem 1. *If Assumption 1 holds and $J_n^d \log(J_n^d) = o(n)$, then for $s \leq S \wedge (K-1)$:*

$$\max_{[m]\leq s} \left\|\partial^m \hat{\mu} - \partial^m \mu\right\|_2^2 = O_p\left(\frac{J_n^{d+2s}}{n} + J_n^{-2((S+\alpha)\wedge K - s)}\right).$$

This theorem shows that, by setting J_n^d proportional to $n^{d/(2(S+\alpha)+d)}$ and $K \geq S+1$, the partitioning estimator achieves Stone's (1982) optimal rate, a property shared by other series- and kernel-based estimators. Because the partitioning estimator can be recast as a series estimator, the conclusion in Theorem 1 for the regression function (i.e., $[m] = 0$) could have been obtained directly from general results in the sieve estimation literature under high-level assumptions. A contribution of this theorem is to obtain such a result under weaker, primitive conditions. In particular, the rate restriction required, $J_n^d \log(J_n^d) = o(n)$, is weaker than the one typically imposed in the general series literature (e.g., Newey, 1997 requires the analogue of $J_n^d \max_{[m]\leq s} \|\partial^m \mathbf{R}_n(\cdot)\|_\infty^2 = o(n)$ with $\max_{[m]\leq s} \|\partial^m \mathbf{R}_n(\cdot)\|_\infty^2$ polynomial in J_n^d). This refined rate restriction was also used by Huang (2003) for multivariate regression splines and by Belloni et al. (2012) for general series estimation, but employing the operator norm instead of the (stronger) Frobenius norm used herein.

Theorem 1 also contributes to the literature in two additional ways. First, existing results for partitioning estimators of $\mu(\cdot)$ only yield the optimal rate when Y is bounded, and otherwise give suboptimal rates (see, Györfi et al., 2002, Corollaries 11.2 and 19.3). Second, this result shows that the partitioning estimator of derivatives of $\mu(\cdot)$ achieves the optimal rate under the same weak conditions. This result, which appears to be new for the partitioning estimation literature, is often useful in econometric applications (e.g., average marginal effects).

Next, we discuss the L_∞ convergence rate of the partitioning estimator.[3]

Theorem 2. *Suppose the conditions of Theorem 1 hold. If, in addition, for some $\xi \in [0, 1 \wedge \eta]$ the partition satisfies $J_n^{d\xi(1+2/\eta)} \log(J_n^d)^{2-(1+2/\eta)\xi} = O(n)$, with $0/0 \equiv 0$, then for $s \leq S \wedge (K-1)$:*

$$\max_{[m]\leq s} \left\|\partial^m \hat{\mu} - \partial^m \mu\right\|_\infty^2 = O_p\left(\frac{J_n^{(2-\xi)d+2s} \log(J_n^d)^\xi}{n} + J_n^{-2((S+\alpha)\wedge K - s)}\right).$$

The parameter ξ is a user-defined choice, which depends on the underlying moment condition of Assumption 1(c). This parameter is not a tuning parameter in the classical nonparametric sense, but rather is explicitly introduced in Theorem 2 for potential applications. As formalized in Lemma A.5, ξ allows for greater or lesser weight placed on the tails of the (conditional) distribution of the outcome variable, which in turn provides a trade-off between the rate restriction imposed and the (possibly suboptimal) rate of convergence of the estimator. Consider two examples: (i) if $\mathbb{E}[Y^4|X] < \infty$ (i.e., $\eta = 2$), then the additional requirement of Theorem 2 is $J_n^{2d} = O(n)$ for $\xi = 1$, implying essentially $(S+\alpha) \wedge K \geq d/2$, and (ii) if $\eta = 0$, which implies $\xi = 0$, only bounded conditional variance is assumed, and Theorem 2 gives the (suboptimal) rate $J_n^{2(d+s)}/n$ with convergence implying the other rate restrictions. For $0 < \xi < 1$ neither rate restriction in Theorem 2 implies the other.

With an appropriate choice of J_n, the convergence rate in Theorem 2 will be optimal if $\eta \geq 1$, allowing for $\xi = 1$, provided the rate restrictions are satisfied. Known results for general series estimation (e.g., Newey, 1997 and de Jong, 2002) do not achieve this optimal uniform rate, and impose stronger side restrictions, which implies that the rate-optimality of the partitioning estimator cannot be deduced from those results. Conventional local polynomial estimators, on the other hand, do achieve this optimal uniform rate (e.g., Masry, 1996) but these results do not apply to the partitioning estimator, as discussed above.

In semiparametric contexts it may be neither necessary nor desirable that the nonparametric component attain the optimal rate, if the goal is to minimize the restrictions imposed. Forcing the preliminary nonparametric estimator to achieve the optimal rate may require overly-restrictive conditions on the model (e.g., moments) or tuning/smoothing parameters. Theorem 2 shows that these conditions may be ameliorated by an appropriate choice of ξ. See Cattaneo and Farrell (2011) for an application of this result.

Finally, we note that the bias rate in Theorems 1 and 2, $J_n^{-2((S+\alpha)\wedge K - s)}$, highlights the fact that the partitioning estimator is not "adaptive" to the underlying smoothness of the regression function: to improve the convergence rate of the estimator, the order K must be chosen "large enough" given the unknown smoothness level, S. This feature is common to many other nonparametric estimators, including local polynomials and regression splines.[4] Although there are estimators that "adapt" to the underlying smoothness, these are usually believed to have poor finite-sample properties.

3.2. Integrated mean-square error expansion

We present a general conditional Integrated Mean-Square Error (IMSE) asymptotic expansion for the partitioning estimator.[5] We

[3] In the Appendix we also provide conditions for the result to hold almost surely.

[4] Assume $s = 0$ for simplicity. An order p local polynomial estimator (p odd) employing bandwidth $h_n \to 0$ has bias-rate $h_n^{(S+\alpha)\wedge p}$ (e.g., Fan and Gijbels, 1996 or Masry, 1996). An order p regression spline estimator employing knots $\kappa_n \to \infty$ has bias-rate $\kappa_n^{-((S+\alpha)\wedge p)}$ (e.g., Huang, 2003 or Belloni et al., 2012).

[5] The supplemental appendix also contains an unconditional IMSE expansion for the special case of $K = 1$.

focus on evenly split partitions for notational simplicity, but the results may be extended to other partitioning schemes. We briefly discuss how to derive a direct plug-in rule for selecting the value of J_n using this expansion, which provides an alternative to the cross-validation procedures discussed in Györfi et al. (2002, Chapters 8, 13).

We impose the following additional assumption.

Assumption 2. (a) $\sigma^2(x) = \mathbb{V}[Y|X = x]$ and $f(x)$ are continuous on $\mathcal{X}$.
(b) $\mu(x)$ is $(S + 1)$-times continuously differentiable on (an extension of) $\mathcal{X}$.

These additional smoothness conditions allow us to characterize the leading constants in the asymptotic IMSE expansion, as opposed to giving bounds in the rates of convergence. Assumption 2(b) is a slight strengthening of Assumption 1(e). Let $\mathrm{vol}(\mathcal{X})$ denote the volume of the support, with $|\mathcal{X}_\ell|$ denoting the length of the interval $\mathcal{X}_\ell$ for $\ell = 1, 2, \ldots, d$, and set $\mathbf{X}_n = (X_1, \ldots, X_n)'$. To save some notation, we also assume $K = S + 1$.

Theorem 3. *Suppose the conditions of Theorem 1 and Assumption 2 hold. If $w(x)$ is continuous on $\mathcal{X}$, then:*

$$\int_{\mathcal{X}} \mathbb{E}\left[\left(\partial^m \hat{\mu}(x) - \partial^m \mu(x)\right)^2 \mid \mathbf{X}_n\right] w(x)dx = \frac{J_n^{d+2[m]}}{n}[\mathscr{V}_{K,d,m} + o_p(1)] + J_n^{-2(K-[m])}[\mathscr{B}_{K,d,m} + o_p(1)],$$

where $\mathscr{V}_{K,d,m}$ and $\mathscr{B}_{K,d,m}$ are given in Eqs. (A.4) and (A.5) in the Appendix.

This result gives a general conditional IMSE expansion valid for any dimension d, any order K, and any derivative m. Under similar conditions, analogous results restricted to $d > 1$ with $[m] = 0$ or $d = 1$ with $[m] = m \geq 0$ are given by Ruppert and Wand (1994) for conventional local polynomial estimators and for $d = 1$ with $[m] = m \geq 0$ by Huang (2003) and Zhou and Wolfe (2000) for regression splines.

We leave the exact expressions of the constants $\mathscr{V}_{K,d,m}$ and $\mathscr{B}_{K,d,m}$ for the general case in the Appendix as they are notationally cumbersome. These expressions simplify considerably for interesting special cases. Specifically, consider estimating $\mu(x)$, i.e., $[m] = 0$. While $\mathscr{B}_{K,d,0}$ remains cumbersome (see Eq. (A.6)), the variance constant reduces to

$$\mathscr{V}_{K,d,0} = \frac{\dim(R(\cdot))}{\mathrm{vol}(\mathcal{X})} \int_{\mathcal{X}} \frac{\sigma^2(x)}{f(x)} w(x)dx,$$

for any d and K. If, in addition, we restrict attention to the univariate case,

$$\mathscr{B}_{K,1,0} = \frac{\mathrm{vol}(\mathcal{X})^{2K}}{2^{2K+1}(K!)^2}\left(\frac{2}{1+2K} - P_K' Q_K^{-1} P_K\right) \times \int_{\mathcal{X}} \left(\partial^K \mu(x)\right)^2 w(x)dx,$$

where $P_K = \int_{-1}^{1} R(x)x^K dx$ and $Q_K = \int_{-1}^{1} R(x)R(x)'dx$. Alternatively, for the piecewise constant fit in the multivariate case, we obtain the tidy expression

$$\mathscr{B}_{1,d,0} = \frac{1}{12}\sum_{\ell=1}^{d} |\mathcal{X}_\ell|^2 \int_{\mathcal{X}} \left(\frac{\partial \mu(x)}{\partial x_\ell}\right)^2 w(x)dx.$$

In all possible cases, minimization of the general asymptotic IMSE obtained in Theorem 3 with respect to J_n gives the optimal choice

$$J_n^* = \left\langle (\mathscr{C}_{K,d,m}\, n)^{\frac{1}{d+2K}} \right\rangle, \quad \mathscr{C}_{K,d,m} = \frac{2(K - [m])\ \mathscr{B}_{K,d,m}}{(d + 2[m])\ \mathscr{V}_{K,d,m}},$$

and $\langle\cdot\rangle$ denotes the nearest integer. A feasible plug-in rule can be easily constructed by using preliminary estimators for the unknown objects in $\mathscr{C}_{K,d,m}$.

4. Bahadur representation and asymptotic normality

This section studies the asymptotic behavior of partitioning-based estimators of linear functionals of the regression function. We establish a uniform Bahadur representation, asymptotic normality, and consistency of a suitable standard-error estimator, for both regular and irregular (not root-n estimable) estimands. The estimand of interest is given by $\theta = \theta(\mu)$ and we consider the simple plug-in estimator $\hat{\theta} = \theta(\hat{\mu})$. The following assumption characterizes the class of functionals considered.

Assumption 3. $\theta(\tilde{\mu}) \in \mathbb{R}$ is linear, and $|\theta(\tilde{\mu})| \leq C \max_{[m]\leq s} \|\partial^m \tilde{\mu}\|_\infty$, for some $C > 0$.

This assumption restricts the class of functionals to be linear and bounded (i.e., continuous) in the appropriate uniform norm. It is not difficult to extend the results presented here to cover non-linear functionals, although this extension is omitted to conserve space.[6] Many interesting econometric applications are covered by linear functionals of the regression function. For concreteness, consider the following three examples.[7]

Example (*Pointwise Inference*). $\theta_{1,m}(\mu) = \partial^m \mu(x), m \in \mathbb{Z}_+^d$, $[m] < K$, where differentiation is defined in Eq. (2). This irregular estimand is useful for nonparametric inference for the regression function and its derivatives. □

Example (*Partial and Full Means*). $\theta_{2,\delta}(\mu) = \int_{\times_{\ell=1}^{\delta} \mathcal{X}_\ell} \mu(x) f(x_1, \ldots, x_\delta)dx_1 \cdots dx_\delta$, $\delta \leq d$, where components of $x \in \mathcal{X}$ not integrated over are held fixed at some value (we assume the x_ℓ are ordered such that integration is over the first δ covariates). Estimating partial and full means is an important problem in econometrics (see, e.g., Newey, 1994). It is well known that $\theta_{2,\delta}(\hat{\mu})$ will not be $\sqrt{n}$-consistent unless $\delta = d$, though the convergence rate increases as more regressors are integrated out. □

Example (*Weighted Average Derivative*). $\theta_{3,m}(\mu) = -\int_{\mathcal{X}} \mu(x)(\partial^m w(x))dx$, $[m] = 1$, where $w(x)$ is a continuously differentiable weighting (trimming) function that vanishes outside a compact subset of $\mathcal{X}$. The functional in this example corresponds to the indirect weighted average derivative (integration by parts gives $\theta_{3,m}(\mu) = \int_{\mathcal{X}} (\partial^m \mu(x)) w(x)dx$), and leads to a simpler estimator based on the regression function directly. Estimating weighted average derivatives is a well-studied problem (see, e.g., Stoker, 1986). The conditions on the weighting function $w(x)$ are essential to eliminate the influence of the boundary of the regressors' support, and achieve $\sqrt{n}$-consistency. □

The first result in this section establishes a uniform Bahadur-type representation for $\theta(\hat{\mu})$. Specifically, we show that the estimator may be represented as an average of independent, conditionally mean-zero random variables forming a triangular array based on certain smoothing weights, plus a remainder that enjoys a particular rate of convergence. This representation facilitates verification of a variety of properties of semiparametric estimators employing the partitioning estimator as a preliminary step.[8]

[6] This extension is achieved by a standard "linearization" argument: first the functional is assumed to be differentiable in the appropriate sense (e.g. Frechet differentiable with respect to an appropriate norm), and then rate restrictions are imposed so that the linearization error is asymptotically negligible.

[7] For other examples of linear (and non-linear) functionals of interest see, e.g., Andrews (1991), Newey (1997), Chen (2007), Ichimura and Todd (2007), and references therein.

[8] For a recent detailed discussion of the applicability of the Bahadur representation to semiparametric inference, and such a result for kernel-based local polynomials, see Kong et al. (2010).

To describe the result, define $\varepsilon_i = Y_i - \mu(X_i)$, $i = 1, \ldots, n$, and $q_j = \mathbb{P}[X \in P_j]$, $j = 1, \ldots, J_n^d$. Because $q_j \asymp J_n^{-d}$ by Assumption 1(d), q_j captures the rate of convergence of each cell (as well as the local behavior of $f(x)$ in each cell). The Bahadur representation of the partitioning-based estimator is then given by:

$$\begin{aligned} \theta(\hat{\mu}) - \theta(\mu) &= \frac{1}{n}\sum_{i=1}^{n} \Psi_n(X_i)\varepsilon_i + \theta(\nu_n), \\ \Psi_n(z) &= \sum_{j=1}^{J_n^d} \Theta_j' \Omega_j^{-1} R_j(z)/q_j, \end{aligned} \tag{3}$$

with $\Theta_j = (\theta([R_j(\cdot)]_1), \ldots, \theta([R_j(\cdot)]_{\dim(R(\cdot))}))'$, where $[\cdot]_g$ denotes the gth element of a vector, and $\Omega_j = \mathbb{E}\left[R_j(X)R_j(X)'\right]/q_j$.

The smoothing weight $\Psi_n(x)$ is a nonrandom function which varies with n only through the partitioning scheme. By linearity of the functional, the Bahadur representation for $\hat{\mu}$ automatically yields the result for $\theta(\hat{\mu})$ in Eq. (3). To be concrete, in the Appendix we first write $\hat{\mu}(x) - \mu(x) = \sum_{i=1}^{n} \psi_n(x, X_i)\varepsilon_i/n + \nu_n(x)$ with $\psi_n(x, z) = \sum_{j=1}^{J_n^d} R_j(x)'\Omega_j^{-1}R_j(z)/q_j$ and a remainder $\nu_n(x)$, and then obtain the smoothing weight and remainder in Eq. (3) by applying the functional $\theta(\cdot)$ to ψ_n and ν_n, respectively: $\Psi_n(z) = \theta(\psi_n(\cdot, z))$ and $\theta(\nu_n)$. The following theorem characterizes the uniform convergence rate of $\theta(\nu_n)$.[9]

Theorem 4. *Let Assumption 3 hold with $s \leq S \wedge (K-1)$, and consider the representation in Eq. (3). If the conditions of Theorem 2 hold, then:*

$$\theta(\nu_n) = O_p\left(\frac{J_n^{(2-\xi/2)d+s}\log(J_n^d)^{1+\xi/2}}{n^{3/2}} + \frac{J_n^{d+s}}{n} + J_n^{-((S+\alpha)\wedge K - s)}\right).$$

Before stating the asymptotic normality result, it is helpful to first discuss an asymptotic variance formula, which also captures the rate of convergence in general. To this end, define

$$V_n = \mathbb{E}\left[\Psi_n(X)^2\sigma^2(X)\right] = \sum_{j=1}^{J_n^d} \Theta_j'\Omega_j^{-1}\Gamma_j\Omega_j^{-1}\Theta_j/q_j, \tag{4}$$

with $\Gamma_j = \mathbb{E}\left[R_j(X)R_j(X)'\sigma^2(X)\right]/q_j$. Since a linear least squares estimate is computed within each cell, the asymptotic variance is of the Huber–Eicker–White heteroskedasticity robust form. A plug-in sample analogue of V_n is given by

$$\begin{aligned} \hat{V}_n &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\Psi}_n(X_i)\hat{\varepsilon}_i)^2 \\ &= \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}\Theta_j'\hat{\Omega}_j^{-1}\hat{\Gamma}_j\hat{\Omega}_j^{-1}\Theta_j/q_j, \quad \hat{\varepsilon}_i = Y_i - \hat{\mu}(X_i) \\ \hat{\Omega}_j &= \frac{1}{n}\sum_{i=1}^{n} R_j(X_i)R_j(X_i)'/q_j, \\ \hat{\Gamma}_j &= \frac{1}{n}\sum_{i=1}^{n} R_j(X_i)R_j(X_i)'\hat{\varepsilon}_i^2/q_j. \end{aligned} \tag{5}$$

Notice that q_j is artificially introduced to take explicit account for the convergence rate of each sample (and population) average. These quantities are unknown, but they exactly cancel out in the formulation above, leading to a feasible estimator of the large-sample variance.

Theorem 5. *Suppose the conditions of Theorem 4 hold with $\eta \geq 0$, that $\sigma^2(x)$ is bounded away from zero on $\mathcal{X}$, and $\theta(\nu_n) = o_p(\sqrt{V_n}/\sqrt{n})$.*

(a) *For $\eta > 0$, if $0 < \|\Psi_n\|_2 \to \infty$ and $\|\Psi_n\|_{2+\eta}/\|\Psi_n\|_2 = o(n^{\eta/(4+2\eta)})$, then:*

$$\frac{\sqrt{n}(\theta(\hat{\mu}) - \theta(\mu))}{\sqrt{V_n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{V_n}} + o_p(1) \to_d \mathcal{N}(0, 1).$$

If, in addition, $\|\hat{\mu} - \mu\|_\infty = o_p(1)$, then $\hat{V}_n/V_n \to_p 1$.

(b) *If $\|\Psi_n - \Psi\|_2 \to 0$, $0 < \|\Psi\|_2 < \infty$, and $\theta(\mu) = \mathbb{E}[\Psi(X)\mu(X)]$, then:*

$$\frac{\sqrt{n}(\theta(\hat{\mu}) - \theta(\mu))}{\sqrt{V_n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\Psi(X_i)\varepsilon_i}{\sqrt{V}} + o_p(1) \to_d \mathcal{N}(0, 1),$$

and $V_n \to V = \mathbb{E}[\Psi(X)^2\sigma^2(X)]$. If, in addition, $\|\hat{\mu} - \mu\|_\infty = o_p(1)$, then $\hat{V}_n/V_n \to_p 1$.

This result gives simple and intuitive sufficient conditions for asymptotic normality of a partitioning-based plug-in estimator of $\theta = \theta(\mu)$, and for consistency of a suitable standard-error estimator. The theorem is divided in two parts, which are mutually exclusive, depending on the asymptotic behavior of the smoothing weights in the Bahadur representation. This approach is similar in spirit to the central limit theorems of Newey (1997) for series estimators (compare to his Assumptions 6, and 7), but using the Bahadur representation we put simple sufficient conditions directly on the smoothing weights. These results automatically apply to vector-valued estimands, although we restrict θ to be scalar for simplicity.

The distinctive feature separating the cases is mean-square continuity of the functional $\theta(\cdot)$ and its Riesz representation (see, e.g., van der Vaart, 1991). These conditions are not imposed in Theorem 5(a), so the estimand is irregular, and the CLT is obtained by directly exploiting the triangular array structure of the Bahadur representation. In contrast, in Theorem 5(b) these conditions imply that the estimand is $\sqrt{n}$-consistent and asymptotically linear with influence function $\psi_i = \Psi(X_i)\varepsilon_i$, which permits an easy characterization of the asymptotic variance. This case is important because it gives easy-to-verify sufficient conditions for asymptotic linearity.

The high-level conditions in Theorem 5 need to be verified in each application (i.e. for a particular $\theta(\cdot)$). We demonstrate the applicability of this theorem by returning to the three examples introduced above, and giving simple primitive conditions under which the high-level conditions hold for the partitioning plug-in estimator.

Example (*Pointwise Inference (Continued)*)**.** Suppose the conditions of Theorem 2 hold with $\eta > 0$ and the partition satisfies $J_n^{(2-\xi)d}\log(J_n^d)^{1+\xi/2} = o(n)$ and $\sqrt{n}J_n^{-d/2-(S+\alpha)\wedge K} \to 0$. Then, for $[m] < K$ the conditions of Theorem 5(a) are met, as $\|\Psi_n\|_p^p \asymp J_n^{(p-1)d+p[m]}$ and $V_n \asymp J_n^{d+2[m]}$. Therefore, $\partial^m\hat{\mu}(x) = \partial^m\mu(x) + O_p(J_n^{d/2+[m]}/\sqrt{n})$. The rate restrictions are quite mild in this example. Negligibility of the remainder term requires the "variance" condition $J_n^{(3/2-\xi/2)d}\log(J_n^d)^{1+\xi/2} = o(n)$; standard error estimation necessitates the only slightly stronger restriction above. In the case $\xi = \eta = 1$ (in Theorem 2), the two coincide, giving $J_n^d\log(J_n^d)^{3/2} = o(n)$, and only three bounded moments are assumed. As a comparison, the central limit theorem of Newey (1997) for regression splines requires the analogue of $J_n^{2d}/n \to 0$ and $\sqrt{n}J_n^{-(S+\alpha)\wedge K} \to 0$, and assumes four bounded moments. These improvements are due to the fact that we are able to exactly characterize the convergence rate of V_n, and to the faster rates of convergence and weaker rate restrictions obtained in the previous section for partitioning estimators. □

[9] An almost sure version of this theorem is available in the Appendix.

Example (*Partial and Full Means (Continued)*)**.** Begin with the irregular case ($\delta < d$). Suppose the conditions of Theorem 2 hold with $\eta > 0$ and the partition satisfies $J_n^{[(3-\xi)d+\delta]/2}\log(J_n^d)^{1+\xi/2} = o(n)$ and $\sqrt{n}J_n^{-(d-\delta)/2-(S+\alpha)\wedge K} \to 0$. The conditions of Theorem 5(a) are met as $\|\Psi_n\|_p^p \asymp J_n^{(p-1)(d-\delta)}$ and hence $V_n \asymp J_n^{d-\delta}$. For some values of δ and ξ, this may imply $\|\hat{\mu} - \mu\|_\infty \to_p 0$, otherwise the exponent on J_n must be (slightly) increased to $(2-\xi)d + \delta/2$. These rate restrictions are strengthened by $J_n^{\delta/2}$ compared to the pointwise case, exactly the decrease in the order of the variance. As δ increases to d, the rate of the variance decreases, leading to the rate of convergence $\theta_{2,\delta}(\hat{\mu}) = \theta_{2,\delta}(\mu) + O_p(J_n^{(d-\delta)/2}/\sqrt{n})$, which shows that the estimator is $\sqrt{n}$-consistent only in the full mean case. In this case, $\theta_{2,d}(\mu) = \int_{\mathcal{X}} \mu(x)f(x)dx = \mathbb{E}[\Psi(X)\mu(X)]$, with $\Psi(x) = 1$. Moreover, $\Psi_n(x) = \sum_{j=1}^{J_n^d} e_1' R_j(x) = 1$, and hence $\|\Psi_n - \Psi\|_2 = 0$, which verifies the conditions in Theorem 5(b). □

Example (*Weighted Average Derivative (Continued)*)**.** Suppose the conditions of Theorem 2 hold and the partition satisfies $J_n^{(2-\xi/2)d}\log J^{1+\xi/2} = o(n)$ and $\sqrt{n}J_n^{-(S+\alpha)\wedge K} \to 0$. Then, the conditions of Theorem 5(b) hold and uniform consistency of $\hat{\mu}(x)$ is implied. Specifically, note that $\theta_{3,m}(\mu) = \int_{\mathcal{X}} \mu(x)(\partial^m w(x))dx = \mathbb{E}[\Psi(X)\mu(X)]$, with $\Psi(x) = -f(x)^{-1}\partial^m w(x)$, and hence $\Psi_n(x) = \sum_{j=1}^{J_n^d} R_j(x)'\Omega_j^{-1}\mathbb{E}[R_j(X)\Psi(X)]/q_j$. Under an appropriate smoothness assumption, there will exist $\{\gamma_j^0\}$ such that $\max_{1\le j\le J_n^d}\|\mathbb{1}_{P_j}(\cdot)\Psi(\cdot) - R_j(\cdot)'\gamma_j^0\|_\infty = o(1)$, yielding the mean-square convergence condition. Hence $\theta_{3,m}$ will be $\sqrt{n}$-consistent. □

It is important to mention that Theorems 4 and 5 (and the examples discussed above) are established using uniform norms, which leads to the simple and general sufficient conditions above. In some examples, however, it is possible to improve on these sufficient conditions by relying on the (weaker) L_2 norm. For instance, if the linear functional is continuous with respect to the L_2 norm (and hence regular), then it is possible to improve on the rate restrictions of Theorem 5 by relying on sharper rates on the remainder of the Bahadur representation. In the specific case of partitioning estimators, because of the sharp uniform rates obtained in this paper the difference between the mean-square and uniform convergence rates is only a slow-varying function (i.e., $\log(J_n^d)$) under appropriate moment assumptions, and hence using the stronger uniform norm is not too restrictive.

5. Monte Carlo evidence

We report a subset of the results from an extensive Monte Carlo study that we conducted to explore the finite-sample performance of the partitioning estimator in comparison to local polynomials and regression splines. We focused on estimating the regression function $\mu(x)$, and examined two measures of global accuracy, root integrated mean-square error (MSE) and integrated mean absolute error (MAE), as well as root mean-square error (RMSE) at interior and boundary points of $\mathcal{X}$. The full set of results from our simulation study is available in the online supplement, and includes different sample sizes, dimensions and distributions for the covariates, regression functions, and levels of variability. In addition, as a complement to the nonparametric results presented here, Cattaneo and Farrell (2011) report another extensive simulation study employing the partitioning estimator as a preliminary estimator in semiparametric treatment effect estimation.

We generated 5000 simulated data sets according to $Y_i = \mu(X_{i,1}, X_{i,2}) + \varepsilon_i$, $i = 1, \ldots, n$, with $\varepsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$. The covariates are independently distributed as truncated Beta(B, B) distributions. We set $\sigma^2 = 1$ and consider both $B = 1$ (uniform) and $B = 1/2$ (which places more mass at the boundary), and truncate to [0.05, 0.95]. We discuss only four different specifications for the regression function $\mu(x_1, x_2)$ in this section:

$$\begin{aligned}
&\text{Model 1: } \mu(x_1,x_2)\\
&\quad = 0.7\exp\{-3((4x_1-2+0.8)^2+8(x_2-1/2)^2)\}\\
&\qquad + \exp\left\{-3\left((4x_1-2-0.8)^2+8(x_2-1/2)^2\right)\right\},\\
&\text{Model 2: } \mu(x_1,x_2) = \sin(5x_1)\sin(10x_2),\\
&\text{Model 3: } \mu(x_1,x_2) = \left((1-(4x_1-2)^2)^2\right)(\sin(5x_2)/5),\\
&\text{Model 4: } \mu(x_1,x_2)\\
&\quad = \mathbb{1}\{(4x_1-2)\in[-2,1]\}((4x_1-2)^7-19)/20\\
&\qquad - \mathbb{1}\{(4x_1-2)\in(-1,0]\}(4x_1-2)^2\\
&\qquad + \mathbb{1}\{(4x_1-2)\in(0,1/2]\}(4x_1-2)^4/2\\
&\qquad + \mathbb{1}\{(4x_1-2)\in(1/2,1]\}(4x_1-2)^5\\
&\qquad + \mathbb{1}\{(4x_1-2)\in(1,2]\}(2-(4x_1-2)^3)\\
&\qquad + 4.26\,(\exp(-3x_2) - 4\exp(-6x_2) + 3\exp(-9x_2)).
\end{aligned}$$

These bivariate regression functions are graphed in Fig. 1. These models are adapted from Fan and Gijbels (1996, Chapter 7.5), Braun and Huang (2005) and Eggermont and LaRiccia (2009, Chapter 22.1) to the simulation setup consider here. Model 4 is discontinuous, but we nonetheless include it as another potentially interesting case for comparison.

For all three nonparametric estimators, we use linear and cubic polynomials (i.e. $K = 2$ and $K = 4$ in our notation). We employ a product Epanechnikov kernel with a common bandwidth for local polynomials, and a tensor product of B-splines for regression splines. The tuning parameters are chosen as follows: for local polynomials, the bandwidth is chosen to minimize the asymptotic conditional IMSE derived by Ruppert and Wand (1994); for the partitioning estimator, we use J_n^* defined above; and for regression splines, we use $J_n^* + 1$ knots for each covariate, also placed uniformly in the support. The final choice may not be optimal for regression splines, but is made for two reasons. First, it permits a direct comparison between partitioning and splines, highlighting the role of the inherent discontinuities of the partitioning estimator. Second, direct plug-in rules for splines are available only for special cases. We use both infeasible and feasible tuning parameters selectors, where the data-driven selectors were implemented by extending the procedure outlined by Fan and Gijbels (1996, Section 4.2) to $d = 2$. The infeasible tuning parameter formula is invalid for Model 4, and thus we set $h_n^* = 1/3$ for local polynomials and $J_n^* = 3$ for partitioning and regression splines. We employed the feasible selectors for all models (e.g., ignoring the lack of continuity in Model 4).

We report here only the case $n = 1000$, presented in Table 1. The top half shows results for $B = 1/2$, followed by the uniform case. The first two columns show the infeasible tuning parameter and the rounded mean feasible choice across simulations. In general, no estimator dominates the others and hence an absolute ranking does not emerge from this simulation study. The partitioning estimator is on par with the other two estimators in many cases, by any of the accuracy measures. In general the global measures are not particularly useful to rank these estimators, and although it appears that local polynomials perform better "on average", the differences are usually small. The partitioning estimator often outperforms the others at the point $(x_1, x_2) = (0.1, 0.1)$, indicating good boundary performance.

The discontinuities of the partitioning estimator require further discussion. By comparing to B-splines, we observe that according to the global accuracy measures these discontinuities do not appear to have a deleterious effect: the partitioning estimator is often on par with splines, and occasionally more accurate. The discontinuities are measure zero, so this may not be surprising,

Table 1
Error comparisons for local polynomials, *B*-splines, and partitioning estimators.

Degree:	Tuning parameter		Root integrated MSE		Ingetrated MAE		Point estimation RMSE					
	Linear	Cubic	Linear	Cubic	Linear	Cubic	(0.5, 0.5)		(0.1, 0.5)		(0.1, 0.1)	
							Linear	Cubic	Linear	Cubic	Linear	Cubic
Model 1, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible estimation												
Local polynomial	0.17	0.24	0.160	0.190	0.123	0.147	0.168	0.188	0.151	0.186	0.178	0.216
B-splines	9	4	0.174	0.172	0.134	0.133	0.215	0.235	0.111	0.162	0.161	0.167
Partitioning	9	4	0.179	0.205	0.138	0.156	0.180	0.662	0.137	0.388	0.151	0.169
Feasible estimation												
Local polynomial	0.28	0.27	0.199	0.201	0.146	0.155	0.082	0.148	0.095	0.114	0.083	0.095
B-splines	4	1	0.167	0.176	0.125	0.137	0.416	0.244	0.160	0.124	0.139	0.163
Partitioning	4	1	0.177	0.174	0.133	0.133	0.317	0.256	0.240	0.183	0.136	0.159
Model 2, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible estimation												
Local polynomial	0.12	0.18	0.204	0.232	0.160	0.180	0.203	0.214	0.190	0.267	0.182	0.321
B-splines	16	4	0.209	0.172	0.165	0.132	0.411	0.151	0.270	0.155	0.176	0.169
Partitioning	16	4	0.252	0.217	0.196	0.165	0.573	0.824	0.383	0.396	0.166	0.171
Feasible estimation												
Local polynomial	0.31	0.23	0.487	0.489	0.390	0.387	0.702	0.748	0.602	0.604	0.277	0.284
B-splines	4	4	0.343	0.172	0.275	0.132	0.171	0.151	0.146	0.155	0.226	0.169
Partitioning	4	4	0.373	0.217	0.304	0.165	0.869	0.824	0.298	0.396	0.308	0.171
Model 3, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible estimation												
Local polynomial	0.15	0.37	0.167	0.156	0.128	0.119	0.142	0.107	0.162	0.132	0.186	0.169
B-splines	16	4	0.176	0.158	0.136	0.121	0.284	0.151	0.220	0.155	0.175	0.167
Partitioning	16	4	0.233	0.204	0.180	0.155	0.433	0.681	0.305	0.388	0.174	0.170
Feasible estimation												
Local polynomial	0.33	0.27	0.259	0.265	0.164	0.172	0.113	0.119	0.233	0.242	0.179	0.190
B-splines	4	4	0.187	0.166	0.144	0.128	0.220	0.151	0.142	0.145	0.138	0.166
Partitioning	4	4	0.199	0.195	0.155	0.148	0.214	0.550	0.250	0.329	0.124	0.177
Model 4, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible estimation												
Local polynomial	0.33	0.33	0.666	0.318	0.478	0.242	0.140	0.176	0.307	0.148	0.200	0.194
B-splines	9	9	0.726	0.340	0.575	0.268	0.223	0.276	0.166	0.181	0.198	0.191
Partitioning	9	9	0.652	0.347	0.511	0.266	0.151	0.234	0.205	0.236	0.177	0.210
Feasible estimation												
Local polynomial	0.19	0.26	0.636	0.537	0.523	0.442	0.401	0.409	0.661	0.367	0.595	0.897
B-splines	4	1	0.794	0.627	0.616	0.505	0.692	0.364	0.171	0.180	0.165	0.208
Partitioning	4	1	0.784	0.621	0.603	0.497	0.748	0.436	0.208	0.216	0.161	0.177
Model 1, $X_{i,\ell} \sim \beta(1, 1)$												
Infeasible estimation												
Local polynomial	0.16	0.23	0.166	0.191	0.126	0.144	0.149	0.160	0.175	0.195	0.258	0.275
B-splines	9	4	0.186	0.175	0.146	0.133	0.212	0.203	0.129	0.181	0.216	0.246
Partitioning	9	4	0.183	0.205	0.142	0.156	0.164	0.570	0.150	0.398	0.193	0.236
Feasible estimation												
Local polynomial	0.26	0.26	0.215	0.213	0.165	0.170	0.059	0.127	0.104	0.121	0.099	0.108
B-splines	4	1	0.173	0.187	0.133	0.146	0.397	0.239	0.204	0.150	0.173	0.234
Partitioning	4	1	0.181	0.184	0.139	0.143	0.341	0.266	0.287	0.205	0.159	0.208
Model 2, $X_{i,\ell} \sim \beta(1, 1)$												
Infeasible estimation												
Local polynomial	0.12	0.18	0.204	0.227	0.156	0.171	0.174	0.174	0.208	0.266	0.263	0.400
B-splines	16	4	0.205	0.171	0.159	0.128	0.369	0.129	0.286	0.177	0.249	0.247
Partitioning	16	4	0.252	0.216	0.196	0.164	0.515	0.690	0.412	0.407	0.223	0.238
Feasible estimation												
Local polynomial	0.28	0.23	0.505	0.508	0.411	0.411	0.677	0.704	0.568	0.570	0.315	0.324
B-splines	4	4	0.327	0.171	0.257	0.128	0.168	0.129	0.171	0.177	0.329	0.247
Partitioning	4	4	0.362	0.216	0.292	0.164	0.755	0.690	0.318	0.407	0.428	0.238
Model 3, $X_{i,\ell} \sim \beta(1, 1)$												
Infeasible estimation												
Local polynomial	0.16	0.38	0.161	0.148	0.120	0.110	0.113	0.091	0.175	0.138	0.263	0.206
B-splines	9	4	0.176	0.159	0.136	0.118	0.115	0.130	0.148	0.177	0.220	0.246
Partitioning	9	4	0.194	0.203	0.150	0.154	0.103	0.572	0.160	0.399	0.196	0.239
Feasible estimation												
Local polynomial	0.33	0.27	0.211	0.218	0.133	0.140	0.103	0.103	0.216	0.220	0.166	0.171
B-splines	4	4	0.172	0.164	0.128	0.123	0.174	0.129	0.162	0.165	0.170	0.241
Partitioning	4	4	0.182	0.190	0.137	0.142	0.181	0.451	0.238	0.333	0.139	0.231

(*continued on next page*)

Table 1 (*continued*)

Degree:	Tuning parameter		Root integrated MSE		Ingetrated MAE		Point estimation RMSE					
	Linear	Cubic	Linear	Cubic	Linear	Cubic	(0.5, 0.5)		(0.1, 0.5)		(0.1, 0.1)	
							Linear	Cubic	Linear	Cubic	Linear	Cubic
Model 4, $X_{i,\ell} \sim \beta(1, 1)$												
Infeasible estimation												
Local polynomial	0.33	0.33	0.613	0.322	0.424	0.246	0.136	0.160	0.194	0.149	0.225	0.228
B-splines	9	9	0.677	0.347	0.508	0.275	0.139	0.256	0.152	0.190	0.296	0.276
Partitioning	9	9	0.619	0.350	0.465	0.269	0.134	0.189	0.182	0.249	0.236	0.281
Feasible estimation												
Local polynomial	0.19	0.26	0.603	0.512	0.484	0.408	0.353	0.359	0.554	0.329	0.708	0.942
B-splines	4	1	0.727	0.543	0.527	0.418	0.515	0.304	0.254	0.195	0.253	0.319
Partitioning	4	1	0.723	0.524	0.522	0.393	0.557	0.471	0.304	0.293	0.257	0.259

Notes. Tuning parameters are local polynomial bandwidth and the number of cells for partitioning estimation and *B*-splines, as described in the text. Feasible tuning parameters reported are the (rounded) mean of all estimated values. Integrated MSE and MAE are estimated by averaging over the design points in each simulated data set.

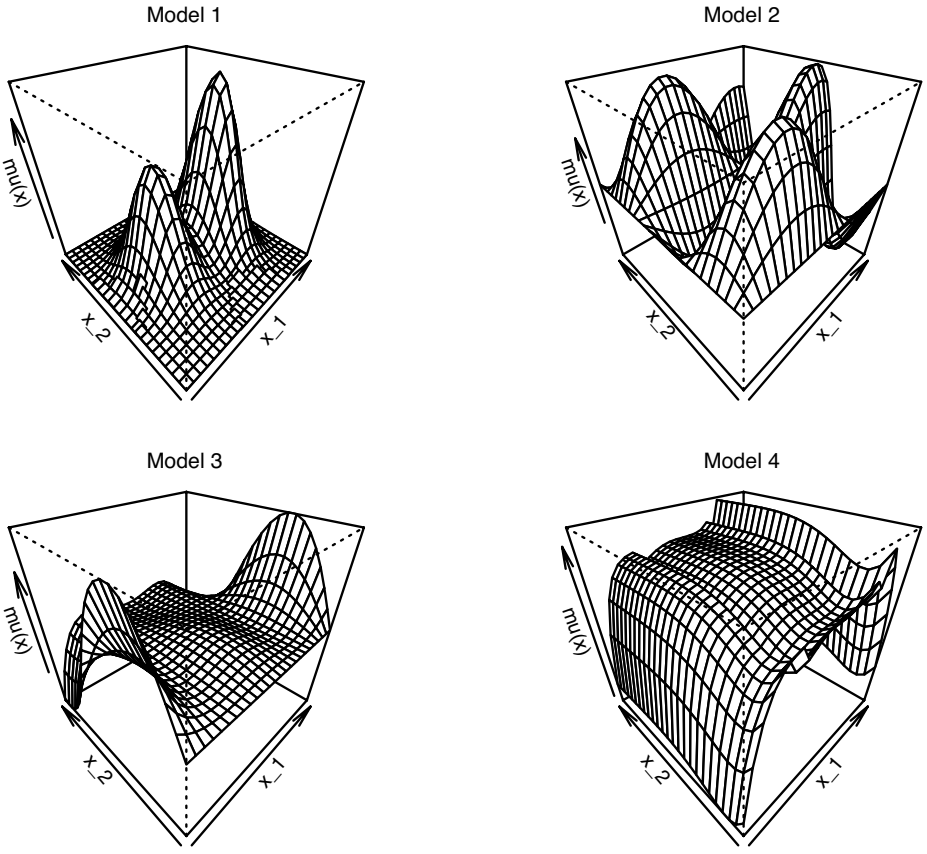

Fig. 1. Regression functions for simulations.

but it shows that the asymptotics provide a good finite-sample approximation in this case. Local polynomials tend to slightly outperform both *B*-splines and partitioning estimators in terms of these global measures, but not in all cases. The pointwise results are also mixed, depending on the data generating process and the evaluation point considered. For instance, in Model 2 at the point $(x_1, x_2) = (1/2, 1/2)$ with $B = 1/2$, the partitioning estimator performs poorly (though the feasible local polynomial estimator is hardly better), except for the infeasible linear fit. This suggests that a practitioner interested in inference at a particular point should not place a cell boundary at that point. Nonetheless, in other cases the partitioning estimator outperforms its continuous counterpart, even in point estimation. The partitioning estimator performs very well in the discontinuous Model 4, even though cell boundaries are not placed at the discontinuities.

6. Conclusion

This paper aimed to give a thorough asymptotic treatment of partitioning estimators of the regression function, its derivatives, and functionals thereof. We established (optimal) mean-square and uniform rates of convergence, a general conditional IMSE expansion and an optimal plug-in rule to select the number of cells, and finally a Bahadur representation and asymptotic normality for linear functionals of the estimator, with valid standard-error estimates. We also showed how these results apply to a few examples, and performed an extensive simulation study.

This estimation strategy appears to have some advantages and disadvantages when compared to other popular nonparametric procedures. Indeed, one goal of this paper was to provide a comprehensive analysis of the partitioning estimators to permit

formal comparison to different nonparametric procedures, as we discussed along the manuscript. While the partitioning estimator is simple and intuitive (and has been proposed in the econometric literature before), it has the perhaps unappealing feature of being discontinuous in finite samples. This property is also shared by the popular nearest-neighbor estimator (e.g., Györfi et al., 2002, Chapter 6), but not by the conventional series- and kernel-based estimators in usual cases. Thus, while we view this estimator as potentially useful in applications (e.g., as a preliminary exploratory device), it is important to highlight that it does ignore the underlying smoothness of the regression function when constructing the estimate. From a theoretical perspective, it is nonetheless interesting that imposing such smoothness is not needed to construct a nonparametric regression estimator that achieves the usual optimal rates of convergence. Moreover, this result shows that the partitioning estimator is not overfitting, even though it enjoys more degrees-of-freedom by not imposing smoothness restrictions as other estimators do (e.g., regression splines).

Appendix. Proof of theorems

Complete technical details may be found in the online supplement. Let C denote a generic positive constant that may take different values in different places. We use $\bigotimes$ for Kronecker products and $\prod$ for usual multiplication. Matrix inequalities are in the positive definite sense. Consecutive uses of the symbol $\asymp$ are interpreted pairwise. For a multi-index k, we define the additional notation: $k! = k_1! \cdots k_d!$, $k \leq \tilde{k} \Leftrightarrow k_1 \leq \tilde{k}_1, \ldots, k_d \leq \tilde{k}_d$, and $\sum_{[k]\leq K} = \sum_{L=0}^{K} \sum_{[k]=L}$ for $K \geq 0$.

Without loss of generality we take the basis to be centered at the midpoint of each cell and scaled by the length of the cell. Observe that centering the polynomial basis around the center of each cell avoids issues of differentiability at the boundary of each cell and the support $\mathcal{X}$. Define the one-to-one function $g(k) : \mathbb{Z}_+^d \to \mathbb{N}$ that gives the index position of $R(x)$ corresponding to entry x^k. Let $g^* = \max_k\{g(k) : k \in \mathbb{Z}_+^d,\ [k] \leq K-1\}$. For a generic cell P_j, let p_{j*}, $\overline{p}_j$, and p_j^* be the vectors in $\mathbb{R}^d$ giving the start, midpoint, and end of the cell, respectively, and let $\overline{p}_{\ell,j} = (p_{\ell,j} + p_{\ell,j-1})/2 \in \mathbb{R}$ be the midpoint of each interval. Define the matrix functions $D(a)$ to be the $K \times K$ diagonal matrix with entries given by $a^{-(v-1)}$, $v = 1, \ldots, K$ and $L(b)$ to be the $K \times K$ lower triangular matrix with typical element $\binom{u-1}{v-1}(-b)^{u-v}$, $(u, v) \in \{1, \ldots, K : u \geq v\}$. We then take the (rotated) polynomial basis to be given by $\tilde{R}_j(x) \equiv \mathbb{1}_{P_j}(x)\tilde{R}(x) = \mathbb{1}_{P_j}(x) S_K \bigotimes_{\ell=1}^{d} \left\{D\left(p_{\ell,j} - \overline{p}_{\ell,j}\right) L(\overline{p}_{\ell,j}) r(x_\ell)\right\}$, where S_K is a $g^* \times K^d$ selection matrix which removes terms of degree exceeding $K-1$. Finally let $\tilde{R}_j = (\tilde{R}_j(X_1), \ldots, \tilde{R}_j(X_n))'$ and (globally) redefine $\Omega_j = \mathbb{E}[\tilde{R}_j(X)\tilde{R}_j(X)']/q_j$ and $\hat{\Omega}_j = \tilde{R}_j'\tilde{R}_j/(nq_j)$, maintaining the same notation for the latter two for simplicity.

A.1. Preliminary lemmas

Several intermediate lemmas are required before proving the main results. These lemmas establish properties of partitioning estimators which may be of independent interest for other applications.

Lemma A.1. *Under Assumption 1(b), the basis satisfies:* $\max_{1\leq j\leq J_n^d} \max_{[m]\leq s} \|\partial^m \tilde{R}_j(\cdot)\|_\infty = O(J_n^s)$, *for* $s \leq K-1$.

Proof. By construction of the partition, for $x \in P_j$, $|x - \overline{p}_j| \leq |p_j^* - \overline{p}_j| \asymp J_n^{-1}$. For fixed $x \in \mathcal{X}$ and a multi-index m such that $[m] \leq K-1$:

$$\left|\partial^m \tilde{R}_j(x)\right|^2 = \frac{1}{(p_j^* - \overline{p}_j)^{2m}} \mathbb{1}_{P_j}(x) \sum_{[k]\leq K-1} \mathbb{1}\{m \leq k\} \times \left\{\frac{k!}{(k-m)!} \frac{(x-\overline{p}_j)^{k-m}}{(p_j^* - \overline{p}_j)^{k-m}}\right\}^2 = O\left(J_n^{2[m]}\right),$$

uniformly in $1 \leq j \leq J_n^d$, $x \in P_j$, and $\{m : [m] \leq K-1\}$, and in particular for those satisfying $[m] \leq s \leq K-1$, for any such s. □

Lemma A.2. *Define* $\mu_j(x) \equiv \mathbb{1}_{P_j}(x)\mu(x)$, *and following the definition in Eq. (2),* $\partial^m \mu_j(x) \equiv \mathbb{1}_{P_j}(x)\partial^m \mu(x)$. *Under Assumption 1(b) and (e), there is a nonrandom vector* β_j^0, *depending only on K and j, such that for* $s \leq S \wedge (K-1)$: $\max_{1\leq j\leq J_n^d} \max_{[m]\leq s} \|\partial^m \mu_j(\cdot) - \partial^m \tilde{R}_j(\cdot)'\beta_j^0\|_\infty = O(J_n^{-((S+\alpha)\wedge K - s)})$.

Proof. Assumption 1(e) implies that $\partial^m \mu_j(x)$ satisfies the Taylor expansion for $x \in P_j$ given by:

$$\partial^m \mu_j(x) = \sum_{[k]\leq S\wedge(K-1)-[m]} \frac{1}{k!}\left(\partial^{k+m}\mu_j(\overline{p}_j)\right)\left(x - \overline{p}_j\right)^k + O\left(\left|x - \overline{p}_j\right|^{(S+\alpha)\wedge K - [m]}\right), \tag{A.1}$$

with constants which can be made uniform in the multi-index m, s, and j. For $k \in \mathbb{Z}_+^d$ define the function $\beta_j^0(k) = \frac{1}{k!}\left(\partial^k \mu_j(\overline{p}_j)\right)(p_j^* - \overline{p}_j)^k$ and the coefficient vector β_j^0 as the $g^* \times 1$ vector with entry e equal to $\beta_j^0(g^{-1}(e))$. Therefore:

$$\partial^m \tilde{R}_j(x)'\beta_j^0 = \sum_{[k]\leq S\wedge(K-1)} \mathbb{1}\{m \leq k\} \frac{\left(x - \overline{p}_j\right)^{k-m}}{(k-m)!} \partial^k \mu_j(\overline{p}_j) = \sum_{[\tilde{k}+m]\leq S\wedge(K-1)} \frac{\left(x - \overline{p}_j\right)^{\tilde{k}}}{\tilde{k}!} \partial^{\tilde{k}+m}\mu_j(\overline{p}_j).$$

This matches the Taylor series, hence subtracting from Eq. (A.1) completes the proof. □

Lemma A.3. *Under Assumption 1,* $\Omega_j \asymp I_{g^*}$, *the identity matrix, uniformly in j.*

Proof. By Assumption 1(d) and the construction of the partition, $q_j \asymp J_n^{-d}$. Applying this result and Assumption 1(d) again, we have: $\Omega_j \asymp J_n^d \int_{\mathcal{X}} \tilde{R}_j(x)\tilde{R}_j(x)'dx$. Now, by Assumption 1(b), properties of the Kronecker product, and the construction of the transformed basis,

$$\Omega_j \asymp J_n^d S_K \bigotimes_{\ell=1}^{d} \left\{\int_{p_{\ell,j-1}}^{p_{\ell,j}} r\left(\frac{x_\ell - \overline{p}_{\ell,j}}{p_{\ell,j} - \overline{p}_{\ell,j}}\right) \times r\left(\frac{x_\ell - \overline{p}_{\ell,j}}{p_{\ell,j} - \overline{p}_{\ell,j}}\right)' dx_\ell\right\} S_K'.$$

Let H denote the Hilbert matrix of order K, which is positive definite. Changing variables $z = (x_\ell - \overline{p}_{\ell,j})/(p_{\ell,j} - \overline{p}_{\ell,j})$, applying $\left|p_{\ell,j} - p_{\ell,j-1}\right| \asymp J_n^{-1}$, changing variables $t = (z+1)/2$, gives

$$\Omega_j \asymp S_K \left\{\bigotimes_{\ell=1}^{d} \int_{-1}^{1} r(z)r(z)'dz\right\} S_K' \asymp S_K \left\{\bigotimes_{\ell=1}^{d} [D(2)L(-1)]^{-1} H \left[L(-1)D(2)\right]^{-1}\right\} S_K' \asymp I_{g^*}. \quad \square$$

Lemma A.4. *Let $a_n = n^{-1}J_n^d \log(J_n^d)$. Under the conditions of Theorem 1: $\max_{1\le j\le J_n^d} |\hat{\Omega}_j - \Omega_j|^2 = O_p(a_n)$. If, in addition, $J_n^d \asymp (n/\log(n))^\gamma$, $\gamma \in (0,1)$, the same is true almost surely.*

Proof. For $k, \tilde{k} \in k \in \mathbb{Z}_+^d : [k] \le K-1$, let the $(g(k), g(\tilde{k}))$ element of $(\hat{\Omega}_j - \Omega_j)$ be denoted $\sum_{i=1}^n W_{ij}(k,\tilde{k})/(nq_j)$, where $W_{ij}(k,\tilde{k}) = [\tilde{R}_j(X_i)\tilde{R}_j(X_i)']_{g(k),g(\tilde{k})} - [\mathbb{E}[\tilde{R}_j(X_i)\tilde{R}_j(X_i)']]_{g(k),g(\tilde{k})}$. By Lemma A.1 and the triangle inequality, $|W_{ij}(k,\tilde{k})| \le C$ and $\mathbb{E}[W_{ij}(k,\tilde{k})^2] \le Cq_j$. Thus by Boole's inequality, K being fixed, Bernstein's inequality, and $q_j \asymp J_n^{-d}$:

$$\begin{aligned}
&\mathbb{P}\left[\max_{1\le j\le J_n^d}\left|\hat{\Omega}_j - \Omega_j\right| > (a_n)^{1/2}\varepsilon\right] \\
&\quad\le CJ_n^d \max_{1\le j\le J_n^d}\max_{[k],[\tilde{k}]\le K-1}\mathbb{P}\left[\left|\sum_{i=1}^n W_{ij}(k,\tilde{k})\right| > q_j\sqrt{nJ_n^d\log(J_n^d)}\varepsilon\right] \\
&\quad\le CJ_n^d \max_{1\le j\le J_n^d}\max_{[k],[\tilde{k}]\le K-1}\exp\left\{-C\frac{q_j^2 nJ_n^d\log(J_n^d)\varepsilon^2}{nq_j + q_j\sqrt{nJ_n^d\log(J_n^d)}\varepsilon}\right\},
\end{aligned}$$

which is arbitrarily small for ε large enough by the rate restriction of Theorem 1. When $J_n^d \asymp (n/\log(n))^\gamma$, the conclusion holds with probability one by the Borel–Cantelli Lemma. □

Lemma A.5. *Let the conditions of Theorem 2 hold, and for ξ therein let $r_n^2 = n^{-1}J_n^{d(2-\xi)}\log(J_n^d)^\xi$. Then for $G = (\mu(X_1),\ldots,\mu(X_n))'$, we have $\max_{1\le j\le J_n^d}|\tilde{R}_j'(Y-G)/(nq_j)| = O_p(r_n)$. If, in addition, $J_n^d \asymp (n/\log(n))^\gamma$, $\gamma\in(0,1)$, and $\eta > 2(1+\xi\gamma)/(1-\xi\gamma)$, the same is true almost surely.*

Proof. With the convention $0/0 = 0$, define $t_n = J_n^{d\xi/\eta}\log(J_n^d)^{-\xi/\eta}$. Following the same notation as in Lemma A.4, let $H_{ij}(k) = \mathbb{1}_{P_j}(X_i)[\tilde{R}_j(X_i)]_{g(k)}(Y_i\mathbb{1}\{Y_i \le t_n\} - \mathbb{E}[Y_i\mathbb{1}\{Y_i\le t_n\}|X_i])$ and $T_{ij}(k) = \mathbb{1}_{P_j}(X_i)[\tilde{R}_j(X_i)]_{g(k)}(Y_i\mathbb{1}\{Y_i > t_n\} - \mathbb{E}[Y_i\mathbb{1}\{Y_i > t_n\}|X_i])$. For the truncated term, since $|H_{ij}(k)| \le t_n$ and $\mathbb{E}[H_{ij}(k)^2] \le Cq_j$, Bernstein's inequality and $q_j \asymp J_n^{-d}$ give, for fixed $k \in \mathbb{Z}_+^d$:

$$\begin{aligned}
&J_n^d \max_{1\le j\le J_n^d}\mathbb{P}\left[\left|\sum_{i=1}^n H_{ij}(k)\right| > nq_jr_n\varepsilon\right] \\
&\quad\le C\exp\left\{\log(J_n^d)\left[1 - C\frac{nr_n^2(J_n^d\log(J_n^d))^{-1}\varepsilon^2}{1+t_nr_n\varepsilon}\right]\right\}.
\end{aligned}$$

For the tails, by Markov's inequality, $\mathbb{E}[T_{ij}(k)] = 0$, Lemma A.1, Assumption 1(c), and $q_j \asymp J_n^{-d}$:

$$\begin{aligned}
&J_n^d \max_{1\le j\le J_n^d}\mathbb{P}\left[\left|\sum_{i=1}^n T_{ij}(k)\right| > nq_jr_n\varepsilon\right] \\
&\quad\le C\frac{J_n^d}{nr_n^2t_n^\eta\varepsilon^2}\max_{1\le j\le J_n^d}\frac{1}{q_j^2}\mathbb{E}\left[\mathbb{1}_{P_j}(X_i)\mathbb{E}\left[|Y_i|^{2+\eta}|X_i\right]\right] \\
&\quad\le C\frac{J_n^{2d}}{nr_n^2t_n^\eta\varepsilon^2}.
\end{aligned}$$

These two bounds do not depend on k, and hence by Boole's inequality and K constant,

$$\begin{aligned}
&\mathbb{P}\left[\max_{1\le j\le J_n^d}\left|\tilde{R}_j'(Y-G)/(nq_j)\right| > r_n\varepsilon\right] \\
&\quad\le CJ_n^d\max_{1\le j\le J_n^d}\max_{[k]\le K-1}\mathbb{P}\left[\left|\sum_{i=1}^n H_{ij}(k)\right| > nq_jr_n\varepsilon\right] \\
&\quad + CJ_n^d\max_{1\le j\le J_n^d}\max_{[k]\le K-1}\mathbb{P}\left[\left|\sum_{i=1}^n T_{ij}(k)\right| > nq_jr_n\varepsilon\right],
\end{aligned}$$

which is arbitrarily small for ε large enough by $\xi \in [0,1]$, the rate restriction of the Theorem, and the definition of t_n. The conclusion holds with probability one by the Borel–Cantelli lemma if $J_n^d \asymp (n/\log(n))^\gamma$ and $t_n = n^\tau$ for $(1+\xi\gamma)/\eta < \tau < (1-\xi\gamma)/2$. □

A.2. Convergence rates

Proof of Theorem 1. Define $\mathbb{1}_{n,j} = \mathbb{1}\{\lambda_{\min}(\hat{\Omega}_j) \ge C\}$ for some positive constant C, where $\lambda_{\min}(\hat{\Omega}_j)$ is the smallest eigenvalue, and take $\hat{\mu}(x) = \sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\tilde{R}_j(x)'\hat{\beta}_j$ (cf. Eq. (1)). As $\min_{1\le j\le J_n^d}\mathbb{1}_{n,j} = 1$ w.p.a. 1, this distinction vanishes asymptotically. First:

$$\begin{aligned}
&\max_{[m]\le s}\left\|\partial^m\hat{\mu} - \partial^m\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\mu_j\right\|_2^2 \\
&\quad\le \max_{[m]\le s}3\sum_{j=1}^{J_n^d}\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))'\hat{\Omega}_j^{-1}\tilde{R}_j'(Y-G)/(nq_j)\|_2^2 \qquad (T_{n1}) \\
&\qquad + \max_{[m]\le s}3\sum_{j=1}^{J_n^d}\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))'\hat{\Omega}_j^{-1}\tilde{R}_j'(G-\tilde{R}_j\beta_j^0)/(nq_j)\|_2^2 \qquad (T_{n2}) \\
&\qquad + \max_{[m]\le s}3\sum_{j=1}^{J_n^d}\|\mathbb{1}_{n,j}[(\partial^m\tilde{R}_j(\cdot))'\beta_j^0 - \partial^m\mu_j(\cdot)]\|_2^2. \qquad (T_{n3})
\end{aligned}$$

By properties of the trace, Assumption 1(c), $\tilde{R}_j(\tilde{R}_j'\tilde{R}_j)^{-1}\tilde{R}_j'$ idempotent, K fixed, and $q_j \asymp J_n^{-d}$,

$$\begin{aligned}
&\mathbb{E}\left[|\mathbb{1}_{n,j}\hat{\Omega}_j^{-1/2}\tilde{R}_j'(Y-G)/(nq_j)|^2|\{X_i\}\right] \\
&\quad= \frac{\mathbb{1}_{n,j}}{nq_j}\mathrm{tr}\left\{\tilde{R}_j(\tilde{R}_j'\tilde{R}_j)^{-1}\tilde{R}_j'\mathbb{E}\left[(Y-G)(Y-G)'|\{X_i\}\right]\right\} \\
&\quad\le C\frac{\mathbb{1}_{n,j}}{nq_j}\mathrm{tr}\left\{\tilde{R}_j\left(\tilde{R}_j'\tilde{R}_j\right)^{-1}\tilde{R}_j'\right\} \le \frac{C}{nq_j} \le \frac{CJ_n^d}{n}.
\end{aligned}$$

Hence, $T_{n1} \le O_p(J_n^{2s})\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\left|\hat{\Omega}_j^{-1/2}\tilde{R}_j'(Y-G)/nq_j\right|^2\int_{P_j}f(x)dx = O_p\left(J_n^{d+2s}/n\right)$, by Markov's inequality and Lemmas A.1 and A.4.

By Boole's and Bernstein's inequality and the condition of Theorem 1:

$$\begin{aligned}
&\mathbb{P}\left[\max_{1\le j\le J_n^d}\sum_{i=1}^n(\mathbb{1}_{P_j}(X_i) - q_j) > nq_j\varepsilon\right] \\
&\quad\le C\exp\left\{\log(J_n^d)\left[1 - C\frac{n}{J_n^d\log(J_n^d)}\frac{\varepsilon^2}{1+\varepsilon}\right]\right\} \to 0. \qquad (A.2)
\end{aligned}$$

Therefore, by $\tilde{R}_j(\tilde{R}_j'\tilde{R}_j)^{-1}\tilde{R}_j'$ idempotent and Lemma A.2:

$$\begin{aligned}
&\max_{1\le j\le J_n^d}\left|\mathbb{1}_{n,j}\hat{\Omega}_j^{-1/2}\tilde{R}_j'(G-\tilde{R}_j\beta_j^0)/(nq_j)\right|^2 \\
&\quad\le \max_{1\le j\le J_n^d}\left|(G-\tilde{R}_j\beta_j^0)'(G-\tilde{R}_j\beta_j^0)/(nq_j)\right| \\
&\quad\le \max_{1\le j\le J_n^d}\left\|\mathbb{1}_{P_j}(\cdot)(\mu(\cdot) - \tilde{R}_j(\cdot)'\beta_j^0)\right\|_\infty^2 \\
&\qquad\times \max_{1\le j\le J_n^d}\frac{1}{nq_j}\sum_{i=1}^n\mathbb{1}_{P_j}(X_i) = O_p\left(J_n^{-2((S+\alpha)\wedge K)}\right). \qquad (A.3)
\end{aligned}$$

Applying Lemmas A.1 and A.4, and $\sum_{j=1}^{J_n^d}\int_{P_j} f(x)dx = 1$, we have $T_{n2} = O_p(J_n^{-2((S+\alpha)\wedge K-s)})$.

Finally, Lemma A.2 immediately gives: $T_{n3} = O(J_n^{-2((S+\alpha)\wedge K-s)})$. □

Proof of Theorem 2. First:

$$\begin{aligned}
&\max_{[m]\le s}\left\|\partial^m\hat{\mu} - \partial^m\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\mu_j\right\|_\infty^2 \\
&\le \max_{1\le j\le J_n^d}\max_{[m]\le s} 3\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))'\hat{\Omega}_j^{-1}\tilde{R}_j'(Y-G)/(nq_j)\|_\infty^2 \\
&\quad + \max_{1\le j\le J_n^d}\max_{[m]\le s} 3\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))' \\
&\quad \times \hat{\Omega}_j^{-1}\tilde{R}_j'(G-\tilde{R}_j\beta_j^0)/(nq_j)\|_\infty^2 \\
&\quad + \max_{1\le j\le J_n^d}\max_{[m]\le s} 3\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))'\beta_j^0 - \partial^m\mu_j(\cdot)\|_\infty^2 \\
&= O\left(J_n^{2s}\right)O_p\left(\frac{J_n^{d(2-\xi)}\log(J_n^d)^\xi}{n}\right) + O_p\left(J_n^{-2((S+1)\wedge K-s)}\right),
\end{aligned}$$

where we apply Lemmas A.1, A.4 and A.5 for the first term; Lemmas A.1 and A.4 and Eq. (A.3) for the second; and Lemma A.2 for the third. The result follows as $\min_{1\le j\le J_n^d}\mathbb{1}_{n,j} = 1$ w.p.a. 1. □

We now demonstrate a version of Theorem 2 that holds with probability one.

Theorem A.1. *Suppose the conditions of Theorem 1 hold. If, in addition, for some $\xi \in [0, 1\wedge\eta]$ the partition satisfies $J_n^d \asymp (n/\log(n))^\gamma$, $\gamma \in (0,1)$ and $\eta > 2(1+\xi\gamma)/(1-\xi\gamma)$, then for $s \le S\wedge(K-1)$:*

$$\begin{aligned}
&\max_{[m]\le s}\left\|\partial^m\hat{\mu} - \partial^m\mu\right\|_\infty^2 \\
&= O_{as}\left(\frac{J_n^{(2-\xi)d+2s}\log(J_n^d)^\xi}{n} + J_n^{-2((S+\alpha)\wedge K-s)}\right).
\end{aligned}$$

Proof of Theorem A.1. The rate restriction on J_n implies that of Theorem 2, whose proof may thus be strengthened to hold with probability one using Eq. (A.2). □

A.3. Asymptotic mean-square error

We first give three lemmas necessary for results. The first two are straightforward, and Lemma A.8 follows identically to Lemma A.4. Proofs may be found in the supplemental appendix.

Lemma A.6. *Let the conditions of Theorem 1 hold and $g(\cdot)$ be continuous on $\mathcal{X}$. Then for $h_j(x) = \mathbb{1}_{P_j}(x)h(x)$, with remainder uniform in $1\le j\le J_n^d$: $\int_{P_j} h(z)g(z)dz = g(\overline{p}_j)\int_{P_j} h(z)dz + \max_{1\le j\le J_n^d}\|h_j(\cdot)\|_\infty (o(J_n^{-d}))$.*

Lemma A.7. *Let the conditions of Theorem 1 hold. If $g(\cdot)$ is continuous on $\mathcal{X}$, then: $\sum_{j=1}^{J_n^d} g(\overline{p}_j)\mathrm{vol}(P_j) = \int_{\mathcal{X}} g(z)dz + o(1)$.*

Lemma A.8. *Under the conditions of Theorem 3, for Γ_j defined Eq. (4) and any $k\in\mathbb{Z}_+^d$:*

$$\begin{aligned}
&\text{(a)}\ \max_{1\le j\le J_n^d}\left|\frac{1}{nq_j}\sum_{i=1}^n \tilde{R}_j(X_i)\tilde{R}_j(X_i)'\sigma^2(X_i) - \Gamma_j\right|^2 \\
&= O_p\left(\frac{J_n^d\log(J_n^d)}{n}\right);
\end{aligned}$$

$$\begin{aligned}
&\text{(b)}\ \max_{1\le j\le J_n^d}\left|\frac{1}{nq_j}\sum_{i=1}^n \tilde{R}_j(X_i)\frac{(X_i-\overline{p}_j)^k}{(p_j^*-\overline{p}_j)^k}\right. \\
&\quad \left. - \frac{1}{q_j}\mathbb{E}\left[\tilde{R}_j(X)\frac{(X_i-\overline{p}_j)^k}{(p_j^*-\overline{p}_j)^k}\right]\right|^2 = O_p\left(\frac{J_n^d\log(J_n^d)}{n}\right).
\end{aligned}$$

Proof of Theorem 3. We first give some notation and facts used repeatedly throughout. With a slight abuse notation, let $|\mathcal{X}|^k = \prod_{\ell=1}^d|\mathcal{X}_\ell|^{k_\ell}$. Let $\mathcal{U} = \times_{\ell=1}^d[-1,1]$. We frequently use the change of variables $z_\ell = (x_\ell - \overline{p}_{\ell,j})/(p_{\ell,j}-\overline{p}_{\ell,j})$, $\ell = 1,\ldots,d$, the Jacobian of which is $\prod_{\ell=1}^d(p_{\ell,j}-\overline{p}_{\ell,j}) = 2^{-d}\mathrm{vol}(P_j) = (2J_n)^{-d}\mathrm{vol}(\mathcal{X})$. For any $k\in\mathbb{Z}_+^d$: $(p_j^*-\overline{p}_j)^k = (2J_n)^{-[k]}|\mathcal{X}|^k$.

Using Lemmas A.1 and A.6 and the change of variables above, we get the following results, which also hold for $w(x) = f(x)$ or $m = 0$:

$$\begin{aligned}
&\text{(a)}\ \int_{\mathcal{X}}(\partial^m\tilde{R}_j(x))(x-\overline{p}_j)^{k-m}w(x)dx \\
&= 2^{-d}w(\overline{p}_j)(p_j^*-\overline{p}_j)^{k-2m}\mathrm{vol}(P_j) \\
&\quad\times\int_{\mathcal{U}}(\partial^m R(z))z^k dz + o(J_n^{-d-K});
\end{aligned}$$

$$\text{(b)}\ \Omega_j = \frac{2^{-d}}{q_j}f(\overline{p}_j)\mathrm{vol}(P_j)\int_{\mathcal{U}}R(z)R(z)'dz + o(J_n^{-d});$$

$$\begin{aligned}
&\text{(c)}\ \int_X(\partial^m\tilde{R}_j(x))(\partial^m\tilde{R}_j(x))'w(x)dx \\
&= \frac{2^{-d}w(\overline{p}_j)\mathrm{vol}(P_j)}{(p_j^*-\overline{p}_j)^{2m}}\int_{\mathcal{U}}\left(\partial^m R(z)\right) \\
&\quad\times\left(\partial^m R(z)\right)'dz + o(J_n^{-d-2[m]}).
\end{aligned}$$

First consider the conditional variance term: $\int_{\mathcal{X}}\mathbb{V}[\partial^m\hat{\mu}(x)\mid\mathbf{X}_n]w(x)dx$. By Lemma A.6, $\Gamma_j = \sigma^2(\overline{p}_j)\Omega_j + o(J_n^{-d})$. Applying this result and Lemmas A.1 and A.4 and A.8(a), we have:

$$\begin{aligned}
&\mathbb{V}\left[\sum_{j=1}^{J_n^d}(\partial^m\tilde{R}_j(x))'\mathbb{1}_{n,j}\hat{\Omega}_j^{-1}\tilde{R}_jY/(nq_j)|\mathbf{X}_n\right] \\
&= \sum_{j=1}^{J_n^d}\frac{1}{nq_j}(\partial^m\tilde{R}_j(x))'\Omega_j^{-1}\Gamma_j\Omega_j^{-1} \\
&\quad\times(\partial^m\tilde{R}_j(x)) + o_p\left(\frac{J_n^{d+2[m]}}{n}\right) \\
&= \sum_{j=1}^{J_n^d}\frac{1}{nq_j}\sigma^2(\overline{p}_j)\mathrm{tr}\left\{\Omega_j^{-1}(\partial^m\tilde{R}_j(x))(\partial^m\tilde{R}_j(x))'\right\} \\
&\quad + o_p\left(\frac{J_n^{d+2[m]}}{n}\right).
\end{aligned}$$

Integrating the above expression, applying Lemma A.6, the above facts and change of variables, and Lemma A.7 (under Assumption 2(a)), we have:

$$\begin{aligned}
&\sum_{j=1}^{J_n^d}\frac{1}{nq_j}\sigma^2(\overline{p}_j)\mathrm{tr}\left\{\Omega_j^{-1}\int_X\left(\partial^m\tilde{R}_j(x)\right)\left(\partial^m\tilde{R}_j(x)\right)'w(x)dx\right\} \\
&\quad + o_p\left(\frac{J_n^{d+2[m]}}{n}\right) \\
&= \frac{J_n^{d+2[m]}}{n}\frac{2^{2[m]}}{|\mathcal{X}|^{2m}\mathrm{vol}(\mathcal{X})}\left(\int_{\mathcal{X}}\frac{\sigma^2(x)}{f(x)}w(x)dx\right)
\end{aligned}$$

$$\times \operatorname{tr}\left\{\left(\int_{\mathcal{U}} R(z)R(z)'dz\right)^{-1}\int_{\mathcal{U}}\left(\partial^m R(z)\right)\left(\partial^m R(z)\right)'dz\right\}$$
$$\times\left[1+o(1)\right]+o_p\left(J_n^{d+2[m]}/n\right).$$

Next consider the bias portion of the expansion: $\int_{\mathcal{X}}(\mathbb{E}[\hat{\mu}(x)|\mathbf{X}_n]-\mu(x))^2 w(x)dx$. Define $T_{K,j,m}(x)=\sum_{k:[k]=K}\left(\partial^k\mu_j(\overline{p}_j)\right)(x-\overline{p}_j)^{k-m}/(k-m)!$, so that under Assumption 2(b), $\partial^m\mu_j(x)=T_{K,j,m}(x)+o(J_n^{-(K-[m])})$ uniformly in $1\le j\le J_n^d$. Then by Lemmas A.4 and A.8,

$$\begin{aligned}
&\sum_{j=1}^{J_n^d}\partial^m\tilde{R}_j(x)'\mathbb{1}_{n,j}(\tilde{R}_j'\tilde{R}_j)^{-1}\sum_{i=1}^{n}\tilde{R}_j(X_i)\mu(X_i)\\
&\quad-\sum_{j=1}^{J_n^d}\partial^m\mu_j(x)=\sum_{j=1}^{J_n^d}\Bigg(\partial^m\tilde{R}_j(x)'\mathbb{1}_{n,j}(\tilde{R}_j'\tilde{R}_j)^{-1}\\
&\quad\times\left(\sum_{i=1}^{n}\tilde{R}_j(X_i)\tilde{R}_j(X_i)'\right)\beta_j^0-\partial^m\mu_j(x)\Bigg)\\
&\quad+\sum_{j=1}^{J_n^d}\partial^m\tilde{R}_j(x)'\mathbb{1}_{n,j}(\tilde{R}_j'\tilde{R}_j)^{-1}\\
&\quad\times\sum_{i=1}^{n}\tilde{R}_j(X_i)\left(T_{K,j,0}(X_i)+o(J_n^{-K})\right)\\
&=-\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\mathbb{1}_{P_j}(x)T_{k,j,m}(x)+\sum_{j=1}^{J_n^d}\partial^m\tilde{R}_j(x)'\mathbb{1}_{n,j}\\
&\quad\times(\tilde{R}_j'\tilde{R}_j)^{-1}\sum_{i=1}^{n}\tilde{R}_j(X_i)T_{K,j,0}(X_i)+o_p\left(J_n^{-(K-[m])}\right)\\
&=\sum_{[k]=K}\sum_{j=1}^{J_n^d}\mathbb{1}_{P_j}(x)\left(\partial^k\mu_j(\overline{p}_j)\right)\\
&\quad\times\left(\frac{\partial^m\tilde{R}_j(x)'}{k!q_j}\Omega_j^{-1}\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^k\right]-\frac{(x-\overline{p}_j)^{k-m}}{(k-m)!}\right)\\
&\quad+o_p\left(J_n^{-(K-[m])}\right).
\end{aligned}$$

Then since $\min_{1\le j\le J_n^d}\mathbb{1}_{n,j}=1$ w.p.a. 1 by Lemma A.4, the integrated, squared bias becomes:

$$\begin{aligned}
&\sum_{j=1}^{J_n^d}\sum_{\substack{k,\ \tilde{k}\\ [k]=[\tilde{k}]=K}}\left(\partial^k\mu_j(\overline{p}_j)\right)\left(\partial^{\tilde{k}}\mu_j(\overline{p}_j)\right)\\
&\quad\times\Bigg\{\frac{1}{(k-m)!(\tilde{k}-m)!}\int_{P_j}(x-\overline{p}_j)^{k+\tilde{k}-2m}w(x)dx\\
&\quad+\frac{1}{k!\tilde{k}!}\frac{1}{q_j^2}\int_{P_j}\partial^m\tilde{R}_j(x)'\Omega_j^{-1}\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^k\right]\\
&\quad\times\mathbb{E}\left[(X-\overline{p}_j)^{\tilde{k}}\tilde{R}_j(X)'\right]\Omega_j^{-1}\partial^m\tilde{R}_j(x)w(x)dx\\
&\quad-\frac{1}{k!(\tilde{k}-m)!}\frac{1}{q_j}\int_{P_j}(x-\overline{p}_j)^{\tilde{k}-m}\partial^m\tilde{R}_j(x)'w(x)dx\Omega_j^{-1}\\
&\quad\times\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^k\right]-\frac{1}{\tilde{k}!(k-m)!}\frac{1}{q_j}\\
&\quad\times\int_{P_j}(x-\overline{p}_j)^{k-m}\partial^m\tilde{R}_j(x)'w(x)dx\Omega_j^{-1}\\
&\quad\times\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^{\tilde{k}}\right]\Bigg\}+o_p\left(J_n^{-2(K-[m])}\right)\\
&=\sum_{j=1}^{J_n^d}\sum_{\substack{k,\ \tilde{k}\\ [k]=[\tilde{k}]=K}}\left(\partial^k\mu_j(\overline{p}_j)\right)\left(\partial^{\tilde{k}}\mu_j(\overline{p}_j)\right)\{B_1+B_2-B_3-B_4\}\\
&\quad+o_p\left(J_n^{-2(K-[m])}\right),
\end{aligned}$$

where the final equality defines the terms B_1–B_4. Applying Lemma A.6 and the change of variables above, and discarding a remainder of order $o(J_n^{-d})O(J_n^{-2(K-[m])})$, these terms are:

$$\begin{aligned}
B_1&=\frac{w(\overline{p}_j)}{(k-m)!(\tilde{k}-m)!}\int_{P_j}(x-\overline{p}_j)^{k+\tilde{k}-2m}dx\\
&=\frac{(p_j^*-\overline{p}_j)^{k+\tilde{k}-2m}w(\overline{p}_j)\mathrm{vol}(P_j)}{2^d(k-m)!(\tilde{k}-m)!}\int_{\mathcal{U}}z^{k+\tilde{k}-2m}dz;\\
B_2&=\frac{1}{k!\tilde{k}!}\frac{1}{q_j^2}\int_{P_j}\operatorname{tr}\{(\partial^m\tilde{R}_j(x))'\Omega_j^{-1}\mathbb{E}[\tilde{R}_j(X)(X-\overline{p}_j)^k]\\
&\quad\times\mathbb{E}[(X-\overline{p}_j)^{\tilde{k}}\tilde{R}_j(X)']\Omega_j^{-1}(\partial^m\tilde{R}_j(x))\}w(x)dx\\
&=\frac{(p_j^*-\overline{p}_j)^{k+\tilde{k}-2m}w(\overline{p}_j)\mathrm{vol}(P_j)}{2^d k!\tilde{k}!}\\
&\quad\times\operatorname{tr}\Bigg\{\left(\int_{\mathcal{U}}R(z)R(z)'dz\right)^{-1}\int_{\mathcal{U}}R(z)z^k dz\\
&\quad\times\int_{\mathcal{U}}R(z)'z^{\tilde{k}}dz\left(\int_{\mathcal{U}}R(z)R(z)'dz\right)^{-1}\\
&\quad\times\int_{\mathcal{U}}(\partial^m R(z))(\partial^m R(z))'dz\Bigg\};\\
B_3&=\frac{(p_j^*-\overline{p}_j)^{k+\tilde{k}-2m}w(\overline{p}_j)\mathrm{vol}(P_j)}{2^d k!(\tilde{k}-m)!}\\
&\quad\times\int_{\mathcal{U}}(\partial^m R(z))'z^{\tilde{k}-m}dz\left(\int_{\mathcal{U}}R(z)R(z)'dz\right)^{-1}\\
&\quad\times\int_{\mathcal{U}}R(z)z^k dz;
\end{aligned}$$

and finally B_4 is identical to B_3 except with k and $\tilde{k}$ reversed.

All four terms have the common factor $(p_j^*-\overline{p}_j)^{k+\tilde{k}-2m}w(\overline{p}_j)\mathrm{vol}(P_j)$, which contains all dependence on the partition. By Lemma A.7, the facts at the outset, and that $[k]=[\tilde{k}]=K$: $\sum_{j=1}^{J_n^d}(\partial^k\mu_j(\overline{p}_j))(\partial^{\tilde{k}}\mu_j(\overline{p}_j))(p_j^*-\overline{p}_j)^{k+\tilde{k}-2m}w(\overline{p}_j)\mathrm{vol}(P_j)=(2J_n)^{-2(K-[m])}|\mathcal{X}|^{k+\tilde{k}-2m}\int_{\mathcal{X}}(\partial^k\mu_j(x))(\partial^{\tilde{k}}\mu_j(x))w(x)dx[1+o(1)]$.

Define:

$$\begin{aligned}
\mathcal{V}_{K,d,m}&=\frac{2^{2[m]}}{\mathrm{vol}(\mathcal{X})}\left(\prod_{\ell=1}^{d}|\mathcal{X}_\ell|^{-2m_\ell}\right)\left(\int_{\mathcal{X}}\frac{\sigma^2(x)}{f(x)}w(x)dx\right)\\
&\quad\times\operatorname{tr}\left\{\left(\int_{\mathcal{U}}R(z)R(z)'dz\right)^{-1}\right.\\
&\quad\left.\times\int_{\mathcal{U}}\left(\partial^m R(z)\right)\left(\partial^m R(z)\right)'dz\right\};
\end{aligned}\tag{A.4}$$

$$\begin{aligned}\mathscr{B}_{K,d,m} &= \frac{1}{2^{2(K+d-[m])}} \sum_{\substack{k,\,\tilde{k} \\ [k]=[\tilde{k}]=K}} \left(\prod_{\ell=1}^{d} |\mathcal{X}_\ell|^{k_\ell+\tilde{k}_\ell-2m_\ell}\right) \\ &\times \left(\int_{\mathcal{X}} \left(\partial^k \mu(x)\right)\left(\partial^{\tilde{k}} \mu(x)\right) w(x)dx\right) \\ &\times \left\{\frac{1}{(k-m)!(\tilde{k}-m)!}\int_{\mathcal{U}} z^{k+\tilde{k}-2m}dz \right. \\ &+ \frac{1}{k!\tilde{k}!}\mathrm{tr}\left[\left(\int_{\mathcal{U}} R(z)R(z)'dz\right)^{-1}\int_{\mathcal{U}} R(z)z^k dz \right. \\ &\times \int_{\mathcal{U}} R(z)'z^{\tilde{k}}dz\left(\int_{\mathcal{U}} R(z)R(z)'dz\right)^{-1} \\ &\times \left.\int_{\mathcal{U}} \left(\partial^m R(z)\right)\left(\partial^m R(z)\right)' dz\right] - \frac{1}{k!(\tilde{k}-m)!} \\ &\times \int_{\mathcal{U}} (\partial^m R(z))' z^{\tilde{k}-m}dz\left(\int_{\mathcal{U}} R(z)R(z)'dz\right)^{-1} \\ &\times \int_{\mathcal{U}} R(z)z^k dz - \frac{1}{\tilde{k}!(k-m)!}\int_{\mathcal{U}} (\partial^m R(z))' \\ &\times \left. z^{k-m}dz\left(\int_{\mathcal{U}} R(z)R(z)'dz\right)^{-1}\int_{\mathcal{U}} R(z)z^{\tilde{k}}dz\right\}. \end{aligned} \tag{A.5}$$

Combining all the above steps we obtain the final result, with $\min_{1\le j\le J_n^d} \mathbb{1}_{n,j} = 1$ w.p.a. 1. □

Finally, we note that for $[m] = 0$:

$$\begin{aligned}\mathscr{B}_{K,d,0} &= \frac{1}{2^{2K+d}} \sum_{\substack{k,\,\tilde{k} \\ [k]=[\tilde{k}]=K}} \frac{1}{k!\tilde{k}!}\left(\prod_{\ell=1}^{d} |\mathcal{X}_\ell|^{k_\ell+\tilde{k}_\ell}\right) \\ &\times \left\{\int_{\mathcal{X}} \left(\partial^k \mu(x)\right)\left(\partial^{\tilde{k}} \mu(x)\right) w(x)dx\right\} \\ &\times \left\{\int_{\mathcal{U}} z^{k+\tilde{k}}dz - \int_{\mathcal{U}} R(z)'z^{\tilde{k}}dz \right. \\ &\times \left.\left(\int_{\mathcal{U}} R(z)R(z)'dz\right)^{-1}\int_{\mathcal{U}} R(z)z^k dz\right\}. \end{aligned} \tag{A.6}$$

A.4. Bahadur representation and asymptotic normality

Proof of Theorem 4. Using the linearity condition on $\theta(\cdot)$, we express the remainder in Eq. (3) as $\theta(\nu_n) = T_{n1} + T_{n2} + T_{n3} + T_{n4}$, where $T_{n1} = \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j}\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\hat{\Omega}_j^{-1}\tilde{R}_j'(Y-G)/(nq_j)$, $T_{n2} = \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j}\hat{\Omega}_j^{-1}\tilde{R}_j'(G - \tilde{R}_j\beta_j^0)/(nq_j)$, $T_{n3} = \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}(\Theta_j'\beta_j^0 - \theta(\mu_j))$, and $T_{n4} = \sum_{j=1}^{J_n^d}(\mathbb{1}_{n,j} - 1)[\theta(\mu_j) + \Theta_j'\Omega_j^{-1}\tilde{R}_j'(Y-G)/(nq_j)]$.

Further, we can write $T_{n1} = T_{n11} - T_{n12}$, with $T_{n11} = \sum_{i=1}^{n} \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j}\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\hat{\Omega}_j^{-1}\tilde{R}_j(X_i)\varepsilon_i/(nq_j)$ and $T_{n12} = \sum_{i=1}^{n}\sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j}\Omega_j^{-1}(\hat{\Omega}_j - \Omega_j)\Omega_j^{-1}\tilde{R}_j(X_i)\varepsilon_i/(nq_j)$. Applying linearity and then continuity of the functional $\theta(\cdot)$ from Assumption 3, followed by Lemmas A.1 and A.3–A.5 we have the following bound on $|T_{n11}|$:

$$\begin{aligned}|T_{n11}| &= \left|\theta\left(\sum_{j=1}^{J_n^d}(\tilde{R}_j(\cdot))'\mathbb{1}_{n,j}\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\Omega_j^{-1} \right.\right. \\ &\times \left.\left.(\Omega_j - \hat{\Omega}_j)\hat{\Omega}_j^{-1}\frac{\tilde{R}_j'(Y-G)}{nq_j}\right)\right| \\ &\le C\max_{[m]\le s}\left\|\sum_{j=1}^{J_n^d}(\partial^m \tilde{R}_j(\cdot))'\mathbb{1}_{n,j}\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\Omega_j^{-1} \right. \\ &\times \left.(\Omega_j - \hat{\Omega}_j)\hat{\Omega}_j^{-1}\frac{\tilde{R}_j'(Y-G)}{nq_j}\right\|_\infty \\ &\le C\left(\max_{1\le j\le J_n^d}\max_{[m]\le s}\|\partial^m \tilde{R}_j(\cdot)\|_\infty\right)\left(\max_{1\le j\le J_n^d}\left|\Omega_j - \hat{\Omega}_j\right|^2\right) \\ &\times \left(\max_{1\le j\le J_n^d}\left|\mathbb{1}_{n,j}\hat{\Omega}_j^{-1}\right|\right)\left(\max_{1\le j\le J_n^d}\left|\Omega_j^{-1}\right|^2\right) \\ &\times \left(\max_{1\le j\le J_n^d}\left|\frac{\tilde{R}_j'(Y-G)}{nq_j}\right|\right) \\ &= O_p\left(\frac{J_n^{(2-\xi/2)d+s}\log(J_n^d)^{1+\xi/2}}{n^{3/2}}\right). \end{aligned}$$

For T_{n12}, begin by defining $W_j(i,l) = \mathbb{1}_{n,j}\Omega_j^{-1}(\tilde{R}_j(X_i)\tilde{R}_j(X_i)' - \mathbb{E}[\tilde{R}_j(X_i)\tilde{R}_j(X_i)'])\Omega_j^{-1}\tilde{R}_j(X_l)\varepsilon_l$, so that we write $T_{n12} = \sum_{j=1}^{J_n^d}\sum_{i=1}^{n}\sum_{l=1}^{n}\Theta_j'W_j(i,l)/(n^2q_j^2)$. Observe that $\mathbb{E}[T_{n12}] = 0$ and that unless $i = h$ and $l = m$, $\mathbb{E}\left[W_j(i,l)W_j(h,m)\right] = 0$. By Lemmas A.1 and A.3, Assumption 1(c), and $q_j \asymp J_n^{-d}$, we have: $\max_{1\le j\le J_n^d}\mathbb{E}[W_j(i,i)W_j(i,i)'] \le C(\max_{1\le j\le J_n^d}|\Omega_j^{-1}|^4)(|\tilde{R}_j(\cdot)|_\infty^6)(\sup_{x\in\mathcal{X}}\sigma^2(x))\max_{1\le j\le J_n^d}\mathbb{E}[\mathbb{1}_{P_j}(X_i)] = O(J_n^{-d})$. Similarly $\max_{1\le j\le J_n^d}\mathbb{E}\left[W_j(i,l)W_j(i,l)'\right] = O(J_n^{-2d})$. Further, Assumption 3 and Lemma A.1 give that: $\max_{1\le j\le J_n^d}|\Theta_j| \le C\max_{1\le j\le J_n^d}\max_{[m]\le s}\|\partial^m\tilde{R}_j(\cdot)\|_\infty = O(J_n^s)$. Therefore the variance of T_{n2} is $O_p(J_n^{2d+2s}/n^2)$ because

$$\begin{aligned}\mathbb{E}[T_{n2}^2] &= \sum_{j=1}^{J_n^d}\frac{1}{(nq_j)^4}\sum_{i=1}^{n}\sum_{l=1}^{n}\Theta_j'\mathbb{E}\left[W_j(i,l)W_j(i,l)'\right]\Theta_j \\ &\le \frac{CJ_n^{4d}}{n^4}\left(\max_{1\le j\le J_n^d}|\Theta_j|\right)\left(\max_{1\le j\le J_n^d} n\mathbb{E}[W_j(i,l)W_j(i,l)'] \right. \\ &+ \left. n(n-1)\mathbb{E}[W_j(i,l)W_j(i,l)']\right) \\ &\times \left(\max_{[m]\le s}\max_{1\le j\le J_n^d}\sup_{x\in P_j}(\partial^m\tilde{R}_j(\cdot))\right), \end{aligned}$$

using $q_j \asymp J_n^{-d}$, linearity and continuity of $\theta(\cdot)$, and Lemma A.1. Hence $|T_{n2}| = O_p\left(J_n^{d+s}/n\right)$, by Markov's inequality.

Similar steps as employed for T_{n11} give $|T_{n2}| = O_p(J_n^{-((S+\alpha)\wedge K-s)})$ and $|T_{n3}| = O_p(J_n^{-((S+\alpha)\wedge K-s)})$, additionally applying Lemma A.2. Finally, from $\min_{1\le j\le J_n^d}\mathbb{1}_{n,j} = 1$ w.p.a. 1 it follows that T_{n4} is smaller order than the other terms. This completes the proof. □

We now demonstrate a version of Theorem 4 that holds with probability one.

Theorem A.2. *Let Assumption 3 hold with $s \le S \wedge (K-1)$, and consider the representation in Eq. (3). If the conditions of Theorem A.1 hold, then:*

$$\theta(\nu_n) = O_{as}\left(\frac{J_n^{(3/2-\xi/2)d+s}\log(J_n^d)^{(1+\xi)/2}}{n} + J_n^{-((S+\alpha)\wedge K-s)}\right).$$

Proof of Theorem A.2. Use the same expansion as in the proof of Theorem 4. Remainders T_{n2}, T_{n3}, and T_{n4} are handled identically, applying the almost sure versions of the same steps, but T_{n1} is bounded directly, using the same steps as for T_{n11} above. □

Proof of Theorem 5(a). By assumption $\sigma^2(x)$ is bounded away from zero, so under Assumption 1(c) we have $\Gamma_j \asymp \Omega_j$. Further by $q_j \asymp J_n^{-d}$ and Lemma A.3 we have:

$$V_n \asymp \mathbb{E}[\Psi_n(X)^2] = \|\Psi_n\|_2^2, \quad \text{and}$$
$$V_n \asymp \sum_{j=1}^{J_n^d} \Theta_j' \Omega_j^{-1} \Theta_j / q_j \asymp J_n^d \sum_{j=1}^{J_n^d} |\Theta_j|^2. \tag{A.7}$$

The condition that $\theta(\nu_n) = o_p(\sqrt{V_n}/\sqrt{n})$ and the result of Theorem 4 immediately give the triangular array representation of the theorem. By construction, $\mathbb{E}[\Psi_n(X_i)\varepsilon_i/\sqrt{nV_n}] = 0$ and $\sum_{i=1}^n \mathbb{E}[(\Psi_n(X_i)\varepsilon_i/\sqrt{nV_n})^2] = 1$. It remains to verify the Lindeberg condition. For any $\delta > 0$, by the Hölder and Markov's inequalities, Assumption 1(c), $V_n \asymp \|\Psi_n\|_2^2$ by Eq. (A.7), and the conditions of the theorem,

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left[\left(\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right)^2 \mathbb{1}\left\{\left|\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right| > \delta\right\}\right] \\
&\quad \le n\left[\mathbb{E}\left[\left(\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right)^{2+\eta}\right]\right]^{\frac{2}{2+\eta}} \left[\mathbb{P}\left[\left|\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right| > \delta\right]\right]^{\frac{\eta}{2+\eta}} \\
&\quad \le \frac{1}{\delta^\eta} \frac{\mathbb{E}\left[|\Psi_n(X_i)|^{2+\eta}\mathbb{E}[|\varepsilon_i|^{2+\eta} \mid X_i]\right]}{n^{\eta/2} V_n^{1+\eta/2}} \\
&\quad = O\left(\left(\frac{\|\Psi_n\|_{2+\eta}}{n^{\eta/(4+2\eta)}\|\Psi_n\|_2}\right)^{2+\eta}\right) \to 0.
\end{aligned}$$

Convergence in distribution follows by the Lindeberg–Feller central limit theorem.

For the second conclusion, observe that by $\mathbb{1}_{n,j} = 1$ w.p.a. 1, uniformly in j, we have $\hat{V}_n/V_n - 1 = T_{n1} + T_{n2} + T_{n3} + o_p(1)$, where $T_{n1} = V_n^{-1}\hat{V}_n - V_n^{-1}\sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}\Theta_j'\hat{\Omega}_j^{-1}\tilde{\Gamma}_j\hat{\Omega}_j^{-1}\Theta_j/q_j$, $T_{n2} = V_n^{-1}\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\Theta_j'(\hat{\Omega}_j^{-1} + \Omega_j^{-1})\tilde{\Gamma}_j(\hat{\Omega}_j^{-1} - \Omega_j^{-1})\Theta_j/q_j$, $T_{n3} = V_n^{-1}\sum_{j=1}^{J_n^d}\Theta_j'\Omega_j^{-1}(\tilde{\Gamma}_j - \Gamma_j)\Omega_j^{-1}\Theta_j/q_j$, and $\tilde{\Gamma}_j = \sum_{i=1}^n \tilde{R}_j(X_i)\tilde{R}_j(X_i)'\varepsilon_i^2/(nq_j)$. First, expanding the squared terms, T_{n1} can be split into two terms, and upon applying Lemmas A.1 and A.4, $q_j \asymp J_n^{-d}$, Eqs. (A.2) and (A.7), and the condition of the theorem, we find that

$$\begin{aligned}
T_{n1} &= V_n^{-1}\sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}\Theta_j'\hat{\Omega}_j^{-1} \\
&\quad \times \left(\frac{1}{nq_j}\sum_{i=1}^n \tilde{R}_j(X_i)\tilde{R}_j(X_i)'(\hat{\mu}(X_i) - \mu(X_i))^2\right) \\
&\quad \times \hat{\Omega}_j^{-1}\Theta_j/q_j - V_n^{-1}\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\Theta_j'\hat{\Omega}_j^{-1} \\
&\quad \times \left(\frac{1}{nq_j}\sum_{i=1}^n \tilde{R}_j(X_i)\tilde{R}_j(X_i)'2\varepsilon_i(\hat{\mu}(X_i) - \mu(X_i))\right) \\
&\quad \times \hat{\Omega}_j^{-1}\Theta_j/q_j \\
&\le \left(\max_{1\le j\le J_n^d} \mathbb{1}_{n,j}|\hat{\Omega}_j^{-1}|^2\right)\left(\max_{1\le j\le J_n^d}\|\tilde{R}_j(\cdot)\|_\infty^2\right)\left(\|\hat{\mu} - \mu\|_\infty\right) \\
&\quad \times \left\{\|\hat{\mu} - \mu\|_\infty \frac{J_n^d}{V_n}\sum_{j=1}^{J_n^d}|\Theta_j|^2 \frac{1}{nq_j}\sum_{i=1}^n \mathbb{1}_{P_j}(X_i)\right. \\
&\quad \left. + \frac{J_n^d}{V_n}\sum_{j=1}^{J_n^d}|\Theta_j|^2\frac{1}{nq_j}\sum_{i=1}^n \mathbb{1}_{P_j}(X_i)|\varepsilon_i|\right\} \\
&= O_p\left(\|\hat{\mu} - \mu\|_\infty\right)\left\{o_p(1)O(1)O_p(1) + O_p(1)\right\} \\
&= o_p(1),
\end{aligned}$$

where the final line additionally uses Assumption 1(c) and the final relation of Eq. (A.7) to give:

$$\begin{aligned}
&\mathbb{E}\left[\frac{J_n^d}{V_n}\sum_{j=1}^{J_n^d}|\Theta_j|^2\frac{1}{nq_j}\sum_{i=1}^n \mathbb{1}_{P_j}(X_i)|\varepsilon_i|\right] \\
&\quad \le C\frac{J_n^d}{V_n}\sum_{j=1}^{J_n^d}|\Theta_j|^2\frac{\mathbb{E}\left[\mathbb{1}_{P_j}(X_i)\mathbb{E}[|\varepsilon_i| \mid X_i]\right]}{q_j} = O(1).
\end{aligned}$$

By Lemma A.1 and otherwise identical steps to the above, we get: $\mathbb{E}[V_n^{-1}\sum_{j=1}^{J_n^d}|\Theta_j|^2|\tilde{\Gamma}_j|/q_j] = O(1)$. Therefore, applying Lemmas A.3 and A.4: $|T_{n2}| \le C(\max_{1\le j\le J_n^d}(\mathbb{1}_{n,j}|\hat{\Omega}_j^{-1}|^3 \vee |\Omega_j^{-1}|^3)) \times (\max_{1\le j\le J_n^d}|\hat{\Omega}_j - \Omega_j|)V_n^{-1}\sum_{j=1}^{J_n^d}|\Theta_j|^2|\tilde{\Gamma}_j|/q_j = o_p(1)$.

Finally, referring to the definitions in Eq. (3), observe that $T_{n3} = \sum_{i=1}^n T_{n3}(i)/n$, where $T_{n3}(i) = V_n^{-1}(\Psi_n(X_i)^2\varepsilon_i^2 - \mathbb{E}[\Psi_n(X_i)^2\varepsilon_i^2])$, so that $\mathbb{E}[T_{n3}(i)] = 0$. Consider two cases. First, suppose $\eta < 2$. Then by Burkholder's inequality, the fact that for $\delta \in (0,1)$, $(a+b)^{(1+\delta)/2} \le a^{(1+\delta)/2} + b^{(1+\delta)/2}$, the c_r inequality, Jensen's inequality, Assumption 1(c), and Eq. (A.7):

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n T_{n3}(i)\right|^{1+\eta/2}\right] \\
&\quad \le \frac{C}{n^{1+\eta/2}}\mathbb{E}\left[\left|\sum_{i=1}^n T_{n3}(i)^2\right|^{(1+\eta/2)/2}\right] \\
&\quad \le \frac{C}{n^{1+\eta/2}}\mathbb{E}\left[\sum_{i=1}^n |T_{n3}(i)|^{1+\eta/2}\right] \\
&\quad \le \frac{C}{n^{\eta/2}}\frac{\mathbb{E}\left[|\Psi_n(X_i)|^{2+\eta}\,\mathbb{E}\left[|\varepsilon_i|^{2+\eta} \mid X\right]\right] + \left(\mathbb{E}\left[\Psi_n(X_i)^2\sigma^2(X)\right]\right)^{1+\eta/2}}{V_n^{1+\eta/2}} \\
&\quad = O\left(\left(\frac{\|\Psi_n\|_{2+\eta}}{n^{\eta/(4+2\eta)}\|\Psi_n\|_2}\right)^{2+\eta}\right) \to 0.
\end{aligned}$$

Next, for the case of $\eta \ge 2$ we utilize only the fourth moment to find that: $\mathbb{E}\left[\left(\sum_{i=1}^n T_{n3}(i)/n\right)^2\right] \le V_n^{-2}\mathbb{E}\left[\Psi_n(X_i)^4\varepsilon_i^4\right]/n = O\left(\|\Psi_n\|_4^4 n^{-1}\|\Psi_n\|_2^{-4}\right) \to 0$, again using Jensen's inequality, Assumption 1(c), and the first relation of Eq. (A.7). In either case, $T_{3n} = o_p(1)$ by Markov's inequality. □

Proof of Theorem 5(b). By Assumption 1(c), the Cauchy–Schwarz and triangle inequalities, and the conditions of the theorem: $V_n - V = \mathbb{E}[(\Psi_n(X)^2 - \Psi(X)^2)\sigma^2(X)] \le C\mathbb{E}[(\Psi_n(X) - \Psi(X))^2]^{1/2}\mathbb{E}[(\Psi_n(X) - \Psi(X) + 2\Psi(X))^2]^{1/2} \le C\|\Psi_n - \Psi\|_2(\|\Psi_n - \Psi\|_2 + 2\|\Psi\|_2) \to 0$, whence the second conclusion.

Convergence in distribution follows under the assumed moment condition on $\Psi(X)$ and a standard central limit theorem, because $\sqrt{n}(\theta(\hat{\mu}) - \theta(\mu))/\sqrt{V_n} - \sum_{i=1}^n \Psi(X_i)\varepsilon_i/(\sqrt{nV}) = \sum_{i=1}^n[(\Psi_n(X_i) - \Psi(X_i))\varepsilon_i/(\sqrt{nV}) + \Psi_n(X_i)\varepsilon_i/(\sqrt{nV})(\sqrt{V/V_n} - 1)] + \sqrt{n}\theta(\nu_n)/\sqrt{V_n} = o_p(1)$ using the above result, the assumed mean-square convergence of $\Psi_n(X)$, and the remainder condition of the theorem.

For the final conclusion, as in the proof of Theorem 5(a) write $\hat{V}_n/V_n - 1 = T_{n1} + T_{n2} + T_{n3} + o_p(1)$, for T_{n1}, T_{n2}, and T_{n3} defined there. As above, $T_{n1} = o_p(1)$ and $T_{n2} = o_p(1)$. Next, $T_{n3} = (V_n^{-1} - V^{-1})\sum_{i=1}^n \Psi_n(X_i)^2\varepsilon_i^2/n + \sum_{i=1}^n[\Psi_n(X_i)^2 - \Psi(X_i)^2]\varepsilon_i^2/(nV) + \sum_{i=1}^n(\Psi(X_i)^2\varepsilon_i^2 - V)/(nV)$, where the first two terms are $o_p(1)$ as in the second conclusion and Markov's inequality, and the third by the law of large numbers. □

References

Andrews, D.W.K., 1991. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. Econometrica 59, 307–345.

Banerjee, A.N., 2007. A method of estimating the average derivative. Journal of Econometrics 136, 65–88.

Belloni, A., Chen, X., Chernozhukov, V., Kato, K., 2012. On the asymptotic theory for least squares series: pointwise and uniform results. Arxiv Preprint arXiv:1212.0442.

Braun, W.J., Huang, L.-S., 2005. Kernel spline regression. The Canadian Journal of Statistics 33, 259–278.

Calonico, S., Cattaneo, M.D., Titiunik, R., 2012. Robust data-driven inference in the regression-discontinuity design. Working Paper, University of Michigan.

Cattaneo, M.D., Farrell, M.H., 2011. Efficient estimation of the dose response function under ignorability using subclassification on the covariates. In: Drukker, D. (Ed.), Advances in Econometrics: Missing Data Methods, Vol. 27A. Emerald Group Publishing Limited, pp. 93–127.

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

Chen, X., Huang, J., 2003. Sup norm convergence rate and asymptotic normality for a class of linear sieve estimators. Shared in personal communication (in preparation).

de Jong, R.M., 2002. A note on ''Convergence rates and asymptotic normality for series estimators'': uniform convergence rates. Journal of Econometrics 11, 1–9.

Eggermont, P.P.B., LaRiccia, V.N., 2009. Maximum Penalized Likelihood Estimation, Volume II: Regression. Springer.

Fama, E.F., French, K.R., 2008. Dissecting anomalies. The Journal of Finance 63, 1653–1678.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman and Hall, London.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H., 2002. A Distribution-Free Theory of Nonparametric Regression. Springer-Verlag.

Huang, J.Z., 2003. Local asymptotics for polynomial spline regression. Annals of Statistics 31, 1600–1635.

Ichimura, H., Todd, P.E., 2007. Implementing nonparametric and semiparametric estimators. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics. In: Handbook of Econometrics, vol. 6B. Elsevier (Chapter 74).

Imbens, G., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142, 615–635.

Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47, 5–86.

Kohler, M., Krzyżak, A., Walk, H., 2006. Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. Journal of Multivariate Analysis 97, 311–323.

Kohler, M., Krzyżak, A., Walk, H., 2009. Optimal global rates of convergence for nonparametric regression with unbounded data. Journal of Statistical Planning and Inference 139, 1286–1296.

Kong, E., Linton, O., Xia, Y., 2010. Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. Econometric Theory 26, 1529–1564.

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

Newey, W.K., 1994. Kernel estimation of partial means and a general variance estimator. Econometric Theory 10, 233–253.

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

Ruppert, D., Wand, M.P., 1994. Multivariate locally weighted least squares regression. Annals of Statistics 22, 1346–1370.

Stoker, T., 1986. Consistent estimation of scaled coefficients. Econometrica 54, 1461–1481.

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

Tukey, J.W., 1947. Non-parametric estimation II. Statistically equivalent blocks and tolerance regions—the continuous case. Annals of Mathematical Statistics 18, 529–539.

van der Vaart, A., 1991. On differentiable functionals. Annals of Statistics 19, 178–204.

Zhou, S., Wolfe, D.A., 2000. On derivative estimation in spline regression. Statistica Sinica 10, 93–108.