

# Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators\*

Matias D. Cattaneo<sup>†</sup>                      Max H. Farrell  
Department of Economics              Department of Economics  
University of Michigan                  University of Michigan

October 5, 2011

## 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. An asymptotic mean-square error approximation is also obtained in a special case, 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.

**Keywords:** nonparametric estimation, partitioning, subclassification, convergence rates, Bahadur representation, asymptotic normality.

**JEL Classification:** C14, C21.

---

\*The authors thank Guido Imbens, Michael Jansson, and Jim Powell for comments on an early version of this paper. Yves Atchade, Lutz Kilian, seminar participants at Berkeley and Michigan, and conference participants at the 2010 Advances in Econometrics Conference have also provided useful comments. We also thank the co-Editor, Peter Robinson, an Associate Editor and four reviewers for detailed comments and suggestions that improved the paper.

<sup>†</sup>Corresponding address: 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

## 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 conditional expectation, its derivative, or a functional thereof is the main object of interest, while in other cases their nonparametric estimators are employed as a first step in a semiparametric procedure. With the growing availability of large data sets and affordable computing power, more complex nonparametric and semiparametric methods are becoming increasingly popular. The implementation of nonparametric estimators, however, requires suitable large sample properties, including sufficiently rapid rates of convergence and known asymptotic distributions with valid standard-error estimators. Series-based and kernel-based methods are typical 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 finer with the sample size, and then within each cell the unknown regression function (and its derivatives) is approximated by parametric regression; a natural choice being linear least squares using a fixed-order polynomial basis. Consistent estimation of the unknown function 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, Kohler, Krzyżak, and Walk (2002, Chapter 4).

The partitioning estimator, although simple and intuitive, has not received a thorough treatment in the econometrics or statistics literatures. The available results typically concern mean-square rates for special cases (see, e.g., Kohler, Krzyżak, and Walk (2006) and references therein). The main goal of this paper is to provide a general asymptotic treatment of partitioning estimators of the regression function and its derivatives. 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 estimators are established and shown to be optimal. More generally, the uniform convergence rate explicitly highlights a natural trade-off between moment assumptions and rate restrictions. Second, in the piecewise constant case, we are further able to characterize the leading terms in an asymptotic mean-square error expansion and

therefore obtain an optimal plug-in selector for the underlying smoothing parameter. Third, we derive a uniform Bahadur representation of linear functionals of the partitioning estimator. Fourth, we establish asymptotic normality of linear functionals of the partitioning estimator, as well as consistency of a suitable standard-error estimator, under simple and intuitive conditions. We cover both regular and irregular estimands. The applicability of the new results is illustrated in three simple examples: (i) derivative of the regression function at a point, (ii) partial and full means, and (iii) weighted average derivatives.

Our motivation to work on this problem is twofold. First, in our view, it is of theoretical interest to understand the asymptotic statistical properties of partitioning estimators, and how these compare to other nonparametric estimators. To facilitate this, after the necessary notation and assumptions are introduced, Section 2.2 provides a detailed comparison between partitioning estimators and other nonparametric estimators. In that section we also outline how our main results contribute to the existent literature.

A second, equally important motivation for studying the partitioning estimators stems from their role in empirical work. Perhaps originating with the regressogram of Tukey (1947), partitioning-based statistical procedures have been informally suggested in the literature and are commonly used in applications, despite their formal properties being unknown in most cases. See, e.g., Cochran (1968) and Rosenbaum and Rubin (1983, 1984) for two well-known examples in the program evaluation literature.<sup>1</sup> Two other recent examples in the econometrics literature where a partitioning-based inference strategy has been proposed are Banerjee (2007) for average derivative estimation and Imbens and Lemieux (2008) in the context of the regression discontinuity design. The results in our paper can be used to formalize the asymptotic validity of these inference procedures and, more generally, can also be useful in other semiparametric and nonparametric settings where the infinite dimensional parameter is a regression function, its derivative, or a functional thereof.

The paper proceeds as follows. In Section 2 we state our main assumptions, describe the partitioning estimator, and provide a comparison to other nonparametric estimators. Section 3 presents the rates of convergence as well as an asymptotic mean-square error expansion. Section 4 develops the Bahadur representation for linear functionals of the estimator, and obtains asymptotic normality. The results of a small Monte Carlo study are presented in Section 5. Proofs are gathered in a technical appendix, and a more detailed supplemental

---

<sup>1</sup>See Imbens and Wooldridge (2009) for a recent survey of the program evaluation literature, which also includes other examples of partitioning-based procedures. See also Cattaneo and Farrell (2011) for further discussion on partitioning in the context of program evaluation, and for an example where some of the results obtained in the present paper are employed in the context of semiparametric inference.

appendix is available upon request.

## 2 THE PARTITIONING ESTIMATOR AND THE LITERATURE

### 2.1 THE PARTITIONING ESTIMATOR

The following conditions are imposed on the data-generating process throughout the paper.

**Assumption 1.**

- (a)  $(Y_1, X'_1), \dots, (Y_n, X'_n)$  is an i.i.d. sample from  $(Y, X')$ , with  $X \in \mathcal{X}$  absolutely continuously distributed.
- (b)  $\mathcal{X} \subset \mathbb{R}^d$  is a Cartesian product of compact, convex intervals.
- (c)  $\mathbb{E}[|Y|^{2+\eta} | X]$  is bounded for some  $\eta \geq 0$ .
- (d) The Lebesgue density of  $X$ ,  $f(x)$ , is bounded and bounded away from zero on  $\mathcal{X}$ .
- (e)  $\mu(x)$  is  $S$ -times continuously differentiable on (an extension of)  $\mathcal{X}$ .

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

**Remarks on Assumption 1.**

- i. 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.
- ii. 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 in nonparametric series estimators (Newey 1997) and nonparametric local polynomials (Fan and Gijbels 1996), but it can be relaxed in the context of 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, Krzyżak, and Walk (2009) show that it is possible to attain Stone (1982)'s optimal  $L_2$  convergence rate even without compactness as long as certain moment conditions hold, and Kohler, Krzyżak, and Walk (2006) show that a cleverly constructed special partitioning estimator attains this rate. In the case of the uniform convergence rate, however, it appears to be an open question whether Stone's (1982)'s bound is achievable without compactness under reasonable regularity conditions.

- iii. Part (c) allows for the case of  $\eta = 0$  (i.e., bounded second conditional moment), and the generality will be useful in the derivation of the uniform convergence rate.
- iv. 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.
- v. Part (e) is a classical smoothness condition controlling the amount of bias reduction possible, when coupled with an appropriate choice of (basis and) order of the polynomial ( $K$ ) employed within each cell.

To describe the nonparametric procedure, we first give a precise description of the partitioning scheme. Following Assumption 1(b), denote  $\mathcal{X} = \prod_{\ell=1}^d \mathcal{X}_\ell$ , with  $\mathcal{X}_\ell \subset \mathbb{R}$ , compact and convex. For a sequence  $J_n \rightarrow \infty$  as  $n \rightarrow \infty$ , partition each  $\mathcal{X}_\ell$  into the  $J_n$  disjoint intervals  $[p_{\ell,j-1}, p_{\ell,j})$ ,  $j = 1, \dots, J_n - 1$ , 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  $J_n^d$  sets formed as Cartesian products of all such intervals. Let  $P_j \subset \mathbb{R}^d$  denote a generic cell of the partition,  $j = 1, \dots, J_n^d$ , 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 and enable nonparametric estimation, we require that  $|p_{\ell,j} - p_{\ell,j-1}| \asymp J_n^{-1}$  for all  $\ell = 1, \dots, d$  and  $j = 1, \dots, J_n$ , 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  $j = 1, \dots, J_n$  nor  $n$ . Hence, by construction the partition satisfies:  $\text{vol}(P_j) \asymp J_n^{-d}$  or equivalently  $C_* J_n^{-d} \leq \min_{1 \leq j \leq J_n^d} \text{vol}(P_j) \leq \max_{1 \leq j \leq J_n^d} \text{vol}(P_j) \leq C_* J_n^{-d}$ , 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  $J_n$ . More general, complex partitioning

schemes are possible, although the one considered here appears to be parsimonious while providing adequate flexibility to cover many applications of interest.

Within each cell the unknown conditional expectation is approximated by solving a least squares problem. For fixed  $K \in \mathbb{N}$ , let  $r(x_\ell) = (1, x_\ell, x_\ell^2, \dots, x_\ell^{K-1})'$  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 order  $K - 1$  formed as the Kronecker product of the  $r(x_\ell)$ , discarding duplicates and terms with order exceeding  $K - 1$ . That is, using multi-index notation,<sup>2</sup> for  $x = (x_1, \dots, x_d)' \in \mathbb{R}^d$  and  $k = (k_1, \dots, k_d)' \in \mathbb{Z}_+^d$ , a typical element of  $R(x)$  is given by  $x^k$  for some  $k \in \{k \in \mathbb{Z}_+^d : |k| \leq K - 1\}$ . We assume  $R(x)$  is ordered ascendingly in  $k \in \mathbb{Z}_+^d$  and  $\ell = 1, \dots, d$ . To fix ideas, consider two simple cases: if  $K = 1$ , then  $R(x) = (1)$  and sample means are fitted in each cell, while if  $K = 2$ ,  $R(x) = (1, x_1, \dots, x_d)'$ , corresponding to ordinary linear least squares. This construction is explicitly meant to cover the general, unrestricted case, although in applications other (restricted) bases may be of interest. For example, if  $\mu(x)$  additively separable, then the interactions between covariates may be excluded from the basis  $R(x)$ , leading to a simpler least squares problem. This additional flexibility is useful, for example, in the context of estimation via control functions. As formalized in the Appendix (Lemma A.2), the ultimate goal of this construction is to ensure that  $R(x)$  is flexible enough to remove bias up to the appropriate order.

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. 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  $J_n \rightarrow \infty$  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

---

<sup>2</sup> The multi-index  $k$  is a  $d$ -tuple of nonnegative integers. We adopt the following (standard) notational conventions:  $x^k = x_1^{k_1} \cdots x_d^{k_d}$ ;  $|k| = k_1 + \cdots + k_d$ ;  $k! = k_1! \cdots k_d!$ ;  $k \leq k' \Leftrightarrow k_1 \leq k'_1, \dots, k_d \leq k'_d$ ;  $\sum_{|k| \leq K-1} = \sum_{m=0}^{K-1} \sum_{k_1=0}^m \cdots \sum_{k_d=0}^m$ ; and  $\partial^k \mu(x) = \frac{\partial^{|k|} \mu(x)}{\partial^{k_1} x_1 \cdots \partial^{k_d} x_d}$ .

notation the partitioning regression estimator (of order  $K - 1$ ) is given by:

$$\begin{aligned} \hat{\mu}(x) &= \sum_{j=1}^{J_n^d} R_j(x)' \hat{\beta}_j, & \hat{\beta}_j &= (R_j' R_j)^{-} R_j' Y, \\ R_j &= (R_j(X_1), \dots, R_j(X_n))', & Y &= (Y_1, \dots, Y_n)', \end{aligned} \quad (1)$$

where  $A^{-}$  denotes any generalized symmetric inverse. Under the 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 in the Appendix for details), and the standard inverse will exist. The structure given in Eqn. (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” are 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  $|m|$ . A natural and intuitive estimator of  $\partial^m \mu(x)$  is given by

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

which we take as the definition throughout. In words,  $\partial^m \mu(x)$  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 the least squares problem within each cell is unaffected, but  $\partial^m R(x)$  corresponds to a lower degree polynomial (i.e. has zeros in some components), the resulting estimator is based on a lower polynomial of approximation within each cell. This intuitively corresponds to estimating the rougher function  $\partial^m \mu(x)$ . As shown in the appendix, because the same polynomial of best approximation can serve for all  $m \geq 0$  (provided  $K$  is large enough), the main results of this paper also naturally cover the estimation of derivatives of the regression function.

## 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 sieve estimation perspective, the partitioning estimator may be recast as a series estimator. Define  $\mathbf{R}_n(x) = (R_1(x)', \dots, R_{J_n^d}(x))'$  by collecting the bases over all  $J_n^d$  cells, and

set  $\mathbf{R}_n = [\mathbf{R}_n(X_1), \dots, \mathbf{R}_n(X_n)]'$ . The partition regression estimator can then be written as

$$\hat{\mu}(x) = \mathbf{R}_n(x)' \hat{\mathbf{B}}_n, \quad \hat{\mathbf{B}}_n = (\mathbf{R}_n' \mathbf{R}_n)^{-1} \mathbf{R}_n' Y = (\hat{\beta}'_1, \dots, \hat{\beta}'_{J_n^d})'.$$

This representation implies that results available from the sieve estimation literature are in principle applicable to the partitioning estimator (see, Chen (2007)). However, as shown in the following sections, exploiting the specific structure of the partition (i.e., the form of its basis) we are able to improve on rate restrictions and obtain faster uniform convergence rates, under simple primitive conditions, when compared to the results available in the general series estimation literature (e.g., Newey (1997) and later refined by de Jong (2002)). Furthermore, we are able to obtain new results, such as derivative estimation and a Bahadur representation, that do not appear to be otherwise available in general.

Polynomial regression splines are a special kind of series estimators for which improved results are available (e.g., 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, but a key distinctive characteristic of splines is that at the cell boundaries (called “knots”) the spline estimate is forced to be smooth whenever possible: a spline of degree  $K$  is  $(K - 1)$ -times differentiable at each knot. It is precisely this condition which make splines 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, while regression splines are only known to have suboptimal uniform rates (de Jong 2002). It seems plausible that an optimal uniform rate could also be established using results from Huang (2003), but we are unable to find an explicit statement in the literature. Moreover, we also obtain the optimal rate for general derivative estimation, as well as other results, as discussed in the following sections.

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 ap-

proaches 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 local polynomial estimators with a 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). These estimation procedures solve the following weighted least-squares problem:

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

where  $W_n(X_i, x)$  is a non-negative weighting function and  $B(X_i)$  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  $h_n \rightarrow 0$ . Moreover, the local polynomials in Fan and Gijbels (1996) are obtained by choosing  $B(X) = R(X - x)$  and setting  $\hat{\mu}(x) = e_1' \hat{\beta}_n(x)$  with  $e_1 = (1, 0, 0, \dots, 0)'$ , while the local polynomial estimator discussed in Eggermont and LaRiccia (2009, Chapter 16) employ  $B(X) = R(X)$  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  $B(X) = R(X)$ , and set  $\hat{\mu}(x) = R(x)' \hat{\beta}_n(x)$ . Therefore, one can not directly apply results for either local polynomial method to partitioning estimators.

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 the current 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 consider 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  $h_n J_n \rightarrow \infty$  (the proof does not directly apply if  $\overline{\lim}_{n \rightarrow \infty} h_n J_n < \infty$ ). 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.

### 3 CONVERGENCE RATES

This section studies the rates of convergence of partitioning estimators under Assumption 1 given above. For scalars  $a$  and  $b$ , let  $a \wedge b = \min\{a, b\}$ . 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  $L_\infty$  norms; function arguments are suppressed if there is no confusion. Rates for derivative estimation are given in terms of  $\max_{|m| \leq s} \|\partial^m h\|_p$ , where the maximum is taken over all multi-indexes  $m$  such that  $|m| \leq s$ .

#### 3.1 MEAN-SQUARE 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  $s \leq S \wedge (K - 1)$ :*

$$\max_{|m| \leq s} \|\partial^m \hat{\mu} - \partial^m \mu\|_2^2 = O_p \left( \frac{J_n^{d+2s}}{n} + J_n^{-2((S+1) \wedge K - s)} \right).$$

This theorem shows that, by setting  $J_n^d$  proportional to  $n^{d/(2(S+1)+d)}$  and  $K \geq S + 1$ , the partitioning estimator achieves the optimal rate of  $n^{-(S+1-s)/(2(S+1)+d)}$  as given by Stone (1982). Therefore, this estimator has the same  $L_2$  rate-optimality properties as other series-based and kernel-based estimators.

Because the partitioning estimator can be recast as a series estimator, the conclusion in Theorem 1 (for the regression function) 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  $J_n^d$ ), and exactly the same one required in the special case of multivariate polynomial splines (e.g., Huang (2003)). This rate restriction controls the granularity of the partition, and guarantees that all cells are asymptotically “full” of observations. For example,  $R_j' R_j = \sum_{i=1}^n \mathbb{1}_{P_j}(X_i) = N_j$  when  $K = 1$ , where  $N_j$  denotes

the number of observations in  $P_j$ , and the rate restriction ensures that for  $n$  large enough this is positive, uniformly over the cells.

This theorem also contributes to the literature in two additional ways. First, existing results for partitioning estimators of  $\mu$  only yield the optimal rate when  $Y$  is bounded, and otherwise give suboptimal rates (see, Györfi, Kohler, Krzyżak, and Walk (2002, Corollaries 19.3 and 11.2)). Theorem 1 therefore improves on these results. 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).

### 3.2 UNIFORM CONVERGENCE

We consider next the  $L_\infty$  convergence rate of the (derivatives of the) partitioning estimator.

**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  $s \leq S \wedge (K - 1)$ :*

$$\max_{|m| \leq s} \|\partial^m \hat{\mu} - \partial^m \mu\|_\infty^2 = O_p \left( \frac{J_n^{(2-\xi)d+2s} \log(J_n^d)^\xi}{n} + J_n^{-2((S+1) \wedge K - s)} \right).$$

The parameter  $\xi$  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 in the Appendix,  $\xi$  allows for greater or lesser weight placed on the tails of the (conditional) distribution of the outcome variable  $Y$ , which in turn provides a trade-off between the rate restriction imposed (on  $J_n$ ) in the theorem and the actual (possibly suboptimal) rate of convergence of the estimator.

The choice of  $\xi$  explicitly accounts for the underlying moment assumptions imposed. The convergence rate in Theorem 2 will be optimal if  $\eta \geq 1$ , allowing for  $\xi = 1$ , provided the rate restrictions are satisfied. In this case,  $J_n^d$  may be chosen to attain Stone's (1982) bound of  $(\log(n)/n)^{(S+1-s)/(2(S+1)+d)}$ ; e.g., if  $\mathbb{E}[Y^4|X] < \infty$  (i.e.,  $\eta = 2$ ), then the additional requirement of Theorem 2 is  $J_n^{2d} = O(n)$ , requiring essentially  $(S+1) \wedge K \geq d/2$ . This result improves on known uniform convergence rates for general series estimation: de Jong (2002) improved on Newey's (1997) uniform convergence rate under the assumptions of a bounded fourth moment and the slightly stronger rate restriction  $J_n^{2d} = o(n)$ , but was still unable to

attain the optimal rate. (As mentioned above, it seems plausible that polynomial regression splines also have an optimal convergence rate under appropriate conditions, but such a result does not appear to be readily available in the literature.) When  $\eta = 0$ , which implies that  $\xi = 0$ , only bounded conditional variance is assumed. In this case, Theorem 2 gives the uniform convergence rate  $J_n^{2(d+s)}/n$ , and convergence implies the other rate restrictions. If, in addition,  $s = 0$  then the convergence rate coincides with the one obtained by Newey (1997) for regression splines, but under weaker rate restrictions. In general, when  $0 < \xi < 1$  both rate restrictions in Theorem 2 must hold, since neither implies the other for certain values of  $\xi$  and  $\eta$ .

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 (e.g., model assumptions and/or tuning/smoothing parameter restrictions). In many semiparametric applications forcing the preliminary nonparametric estimator to achieve the optimal rate requires restrictive moment assumptions and/or rate restrictions; Theorem 2 shows that these conditions may be ameliorated by an appropriate choice of  $\xi$ . To give an example of how this idea works, consider the standard sufficient condition for asymptotic linearity of a semiparametric estimator that uses a first-step nonparametric estimate of the regression function:  $\sqrt{n} \|\hat{\mu} - \mu\|_\infty^2 = o_p(1)$ . (See Newey and McFadden (1994), Chen (2007), and Ichimura and Todd (2007), and references therein.) This condition can still be satisfied under Theorem 2 with  $\xi = 0$ , thereby removing the need for more than two moments. Moreover, in some situations  $0 < \xi < 1$  may be the best choice if the goal is to minimize the moment assumptions and/or the rate restrictions on the tuning parameter. See Cattaneo and Farrell (2011) for a semiparametric estimator that employs such a choice.

Finally, it is also shown in the appendix that if, in addition to the conditions of Theorem 2 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 the same conclusion holds almost surely.

### 3.3 ASYMPTOTIC MEAN-SQUARE ERROR FOR $K = 1$

Here we show that in the particular case of a piecewise constant fit ( $K = 1$ ) the leading constants in the (expected)  $L_2$  approximation can be computed explicitly, leading to a simple (integrated) mean-square error approximation for the partitioning estimator. We then employ this result to derive a plug-in rule for selecting the value of  $J_n$ , thereby providing an alternative to the cross-validation procedures discussed in Györfi, Kohler, Krzyżak, and Walk (2002, Chapters 8, 13).

The special structure of the constant-fit partitioning estimator appears to be crucial in the derivation. The main complication in extending this result to  $K > 1$  is in handling the random “denominator”  $(R'_j R_j)^{-1}$ , which is difficult to deal with in general. When  $K = 1$ ,  $R'_j R_j = N_j$  is a binomial random variable whose inverse moments can be calculated or approximated (see Lemma A.6 in the Appendix). An alternative approach to the one presented here is to consider a conditional mean-square expansion, which in principle can be computed for any choice of  $K$  but leads to an untidy result due to the partitioning structure.

The following theorem present the integrated mean-square approximation for the special case of evenly spaced partitions. This result can be extended to other partitioning schemes, but we focus on this special leading case for notational simplicity. Let  $\text{vol}(\mathcal{X})$  denote the volume of the support and  $|\mathcal{X}_\ell|$  denote the length of the interval  $\mathcal{X}_\ell$ ,  $\ell = 1, 2, \dots, d$ .

**Theorem 3.** *Suppose the conditions of Theorem 1 hold with  $S = 1$  and  $K = 1$ . If  $\sigma^2(x)$  and  $f(x)$  are continuous on  $\mathcal{X}$ , where  $\sigma^2(X) = \mathbb{E}[(Y - \mu(X))^2|X]$ , then for an evenly spaced partition:*

$$\int_{\mathcal{X}} \mathbb{E} [(\hat{\mu}(x) - \mu(x))^2] f(x) dx = \frac{J_n^d}{n} [\mathcal{V} + o(1)] + \frac{1}{J_n^2} [\mathcal{B} + o(1)],$$

where

$$\mathcal{V} = \frac{1}{\text{vol}(\mathcal{X})} \mathbb{E} \left[ \frac{\sigma^2(X)}{f(X)} + \frac{\mu(X)^2}{f(X)} - \mu(X)^2 \right], \quad \mathcal{B} = \frac{1}{12} \mathbb{E} [\nabla \mu(X)' D_{\mathcal{X}} \nabla \mu(X)],$$

with  $D_{\mathcal{X}}$  the  $d \times d$  diagonal matrix with entries  $|\mathcal{X}_\ell|^2$ ,  $\ell = 1, \dots, d$ , and  $\nabla \mu(x) = \partial \mu(x) / \partial x$ .

Minimization of the expected  $L_2$  approximation gives an optimal plug-in choice  $J_n^* = C^* n^{-1/(d+2)}$  with  $C^* = (d\mathcal{V}/(2\mathcal{B}))^{1/(d+2)}$ . A feasible plug-in rule can be easily constructed by using preliminary estimators for the unknown functions in  $C^*$ ; we do not spell out the details to conserve space.

## 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 of the estimator, asymptotic normality, and consistency of a suitable standard-error estimator. The results allow for both regular and irregular (not root- $n$  estimable) estimands. The estimand of interest is given by  $\theta = \theta(\mu)$ . The following assumption characterizes the class of functionals considered.

**Assumption 2.**  $\theta(\tilde{\mu}) \in \mathbb{R}$  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. This section considers the simple plug-in estimator  $\hat{\theta} = \theta(\hat{\mu})$ . It is not difficult to extend the results presented here to cover non-linear functionals, although this extension is omitted to conserve space.<sup>3</sup> Many interesting econometric applications are covered by linear functionals of the regression function. For concreteness, consider the following three examples.<sup>4</sup>

**Example: Pointwise Inference.**  $\theta_{1,m}(\mu) = \partial^m \mu(x)$ ,  $m \in \mathbb{Z}_+$ ,  $|m| < K$ , where differentiation is defined in (2). ■

This example is useful for nonparametric inference for the regression function and its derivatives. This estimand is known to be irregular.

**Example: Partial and Full Means.**  $\theta_{2,\delta}(\mu) = \int_{\prod_{\ell=1}^{\delta} \mathcal{X}_\ell} \mu(x) f(x_1, \dots, x_\delta) dx_1 \cdots dx_\delta$ ,  $\delta \in \{1, \dots, d\}$ , where components of  $x \in \mathcal{X}$  not integrated over are held fixed at some value. ■

This second example corresponds to the important problem in econometrics of estimating partial and full means (see, e.g., Newey (1994)). It is well known that  $\theta_{2,\delta}(\hat{\mu})$  will not be  $\sqrt{n}$ -consistent unless  $\delta = d$ , although the convergence rate improves as more regressors are integrated out (i.e., as  $\delta$  increases to  $d$ ).

**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  $\mathcal{X}$ . ■

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 hence achieve  $\sqrt{n}$ -consistency. 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. The corresponding plug-in estimator requires weaker rate

<sup>3</sup>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.

<sup>4</sup>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.

restrictions than an estimator based on the direct weighted average derivative functional, which involves the derivative of the regression function.

The first result in this section establishes a uniform Bahadur representation for  $\theta(\hat{\mu})$ . Specifically, the result shows 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 useful for applications. This representation is very useful to establish the asymptotic normality of the estimators. In addition, this representation also facilitates verification of a variety of properties of semiparametric estimators employing the nonparametric partitioning estimator as a preliminary step.<sup>5</sup>

To describe the result, define  $\varepsilon_i = Y_i - \mu(X_i)$ ,  $i = 1, \dots, n$ , and  $q_j = \mathbb{P}[X \in P_j]$ ,  $j = 1, \dots, J_n^d$ . Because  $q_j \asymp J_n^{-d}$  by Assumption 1(d),  $q_j$  captures the rate of convergence of each individual 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:

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

with  $\Theta_j = (\theta([R_j(\cdot)]_1), \dots, \theta([R_j(\cdot)]_{\dim(R(\cdot))}))'$ , where  $[v]_g$  denotes the  $g^{\text{th}}$  element of the vector  $v$ , and  $\Omega_j = \mathbb{E}[R_j(X)R_j(X)'] / q_j$ .

The smoothing weight  $\Psi_n(x)$  is a nonrandom function which varies with  $n$  only through the partitioning scheme. It follows by linearity of the functional  $\theta(\cdot)$  that the representation for the level of the function automatically yields the result for the estimand of interest. That is, the smoothing weight may be also represented as  $\Psi_n(z) = \theta(\psi_n(\cdot, z))$ , for the function  $\psi_n(x, z) = \sum_{j=1}^{J_n^d} R_j(x)' \Omega_j^{-1} R_j(z) / q_j$  from the representation of the level of the function at a particular point:  $\hat{\mu}(x) - \mu(x) = \sum_{i=1}^n \psi_n(x, X_i) \varepsilon_i / n + \nu_n(x)$ . The exact form of the remainder  $\theta(\nu_n)$  is given in the appendix, and is derived by applying the functional  $\theta$  to the random function  $\nu_n$ , the remainder in the Bahadur representation of the level of the function. The following theorem characterizes the uniform convergence rate of  $\theta(\nu_n)$ .

**Theorem 4.** *Let Assumption 2 hold with  $s \leq S \wedge (K - 1)$ , and consider the representation*

---

<sup>5</sup>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, Linton, and Xia (2010).

in (3). If the conditions of Theorem 2 hold, then:

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

This result provides the rate of convergence of the remainder in the Bahadur representation of  $\theta(\hat{\mu})$ , which for most purposes is the main information needed to employ the representation. (An almost sure version of this theorem is available in the appendix.)

An asymptotic variance formula is needed to describe the large sample distribution of the estimator, which also captures its rate of convergence in general. To this end, define

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

with  $\Gamma_j = \mathbb{E} [R_j(X) R_j(X)' \sigma^2(X)] / q_j$ . Intuitively, 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

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

Notice that  $q_j$  is artificially introduced to take explicit account for the convergence rate of each sample (and population) average. The  $q_j$ 's are unknown quantities, 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  $0 < \|\Psi_n\|_2 \rightarrow \infty$  and  $\|\Psi_n\|_{2+\eta} / \|\Psi_n\|_2 = o(n^{\eta/(4+2\eta)})$ , then:*

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

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

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

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

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

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  $\theta$  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.

Finally, in both cases, Theorem 5 establishes consistency of  $\hat{V}_n$  under the additional high-level condition that  $\|\hat{\mu} - \mu\|_\infty = o_p(1)$ , thereby providing feasible asymptotic inference for  $\theta$ . Uniform consistency of the regression function estimator is a mild additional requirement, which in many examples will be implied by the conditions already imposed on the Bahadur representation remainder.

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  $\sqrt{n}J_n^{-d/2-(S+1)\wedge K} \rightarrow 0$ . 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  $J_n^{2d}/n \rightarrow 0$  and  $\sqrt{n}J_n^{-(S+1)\wedge K} \rightarrow 0$ , 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. ■

**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  $\sqrt{n}J_n^{-(d-\delta)/2-(S+1)\wedge K} \rightarrow 0$ . 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  $\delta$  and  $\xi$ , this may imply  $\|\hat{\mu} - \mu\|_\infty \rightarrow_p 0$ , 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  $\delta$  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  $J_n^{(2-\xi/2)d} \log(J_n^d)^{1+\xi/2} = o(n)$  and  $\sqrt{n}J_n^{-(S+1)\wedge K} \rightarrow 0$ . 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  $\max_{1 \leq j \leq J_n^d} \|\mathbb{1}_{P_j}(\cdot)\Psi(\cdot) - R_j(\cdot)'\gamma_j^0\|_\infty = o(1)$ , yielding the mean-square convergence condition  $\|\Psi_n - \Psi\|_2^2 \rightarrow 0$ . Hence  $\theta_{3,m}$  will be  $\sqrt{n}$ -consistent, with limiting variance as given in the theorem. ■

It is important to mention that Theorems 4 and 5 (and the examples discussed above) are established using the uniform norm  $\|\cdot\|_\infty$ , an approach 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  $\|\cdot\|_2$ . For instance, if the

linear functional is continuous with respect to  $\|\cdot\|_2$  (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, and 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. A similar discussion applies to functionals continuous with respect to  $\max_{|m|\leq s} \|\partial^m \cdot\|_\infty$ .

## 5 MONTE CARLO

In this section we report the results of a small Monte Carlo study, aimed at analyzing some of the pointwise finite sample properties of the partitioning estimator and other nonparametric estimators. Specifically, our goal is to compare the bias, variance and mean-square error properties of the partitioning estimator, classical local polynomials, and polynomial regression splines, when estimating  $\mu(x)$  for a fixed point  $x \in \mathcal{X}$ . Cattaneo and Farrell (2011) report results from an extensive simulation study in the context of semiparametric estimation where alternative preliminary nonparametric estimators, including partitioning, series and nearest neighbor, are employed. Those findings greatly complement the results presented here.

We consider only a bivariate model with three alternative data-generating processes (DGPs) for brevity. The DGPs considered here are taken from Fan and Gijbels (1996, Chapter 4), preserving as many features of the originals as possible while adapting to the bivariate setting.<sup>6</sup> For each DGP,  $Y_i = \mu(X_{1,i}, X_{2,i}) + \varepsilon_i$ , for  $X_{\ell,i} \sim U[-2, 2]$ ,  $\ell = 1, 2$  and  $\varepsilon_i \sim N(0, 1)$ . To describe the three DGPs, define for a scalar  $x_\ell$  the functions  $\mu_1(x_\ell) = x_\ell + 2 \exp\{-16x_\ell^2\}$  and  $\mu_2(x_\ell) = x_\ell(1 - x_\ell) \sin\{2\pi/(x_\ell + 3)\}$ . Let  $\mu(X_1, X_2) = \mu_{m_1}(X_1) + \mu_{m_2}(X_2) + \mu_{m_1}(X_1)\mu_{m_2}(X_2)$ , with the three DGPs given by, respectively: (1)  $m_1 = m_2 = 1$ ; (2)  $m_1 = 1, m_2 = 2$ ; and (3)  $m_1 = m_2 = 2$ . We conduct 2,000 simulations of each model, with sample sizes of 500, 1,000, and 2,000.

We constructed the piecewise constant, linear, and quadratic partitioning estimators (i.e.  $K=1, 2$ , and 3) over an evenly spaced partition for several choices of  $J_n$ . For comparison, we also compute  $\hat{\mu}(x)$  using classical local polynomials and cubic regression splines (see Section

---

<sup>6</sup>The models here are adapted from Examples 1 and 5 of Fan and Gijbels (1996, Chapter 4), see their Tables 4.1 and 4.3. At present we are not concerned with noise, so we set  $\mathbb{E}[\varepsilon^2] = 1$  throughout. Example 5 is originally used on a uniform fixed design:  $x_i = i/n \in [0, 1]$ , hence we slightly change the functional form to accommodate the present setting.

2.2 for discussion). Local polynomial estimation is implemented with the Epanechnikov kernel with bandwidths set to  $4/J_n$  for each  $J_n$  (where  $4 = |\mathcal{X}_\ell|$ ) and the same list of polynomial degrees (where, e.g.,  $K = 2$  is local linear regression). For regression splines we set the number of knots to  $J_n + 1$  in each dimension. Choosing smoothing and tuning parameters is a delicate matter, and our simulations are not exhaustive. Our heuristic goal with these particular selections was to ensure that each nonparametric procedure had access to roughly the same data (locally to  $x$ ).

The results do not differ qualitatively across the sample sizes 500, 1,000, and 2,000, and so we only report the results for 1,000. Only the range of “acceptable” tuning parameters changes with the sample size. For lower sample sizes and higher  $J_n$  (and larger  $K$ ) there will be an increasing number of cells with insufficient data, so  $(R'_j R_j)$  and its local polynomial and spline equivalents are (nearly) singular, and the estimates are numerically unstable. This effect is already evident in Tables 1–3, where estimation with  $K = 3, J_n = 7$  is not stable. It is beyond the scope of this work to study a formal trimming procedure.

Tables 1–3 present the results for DGPs 1–3, respectively. We estimate  $\mu(x)$  for three choices of  $x$ :  $x_1 = (0, 0)$ ,  $x_2 = (-1, 0)$ , and  $x_3 = (-1.9, 0)$ , i.e. two interior points and one meant to illustrate the boundary behavior of the three estimators. For the partitioning estimator in particular, the results bear out that increasing  $K$  provides bias reduction. However, the piecewise linear fit seems to perform well and allow for a larger  $J_n$ , whereas further increases in  $K$  exacerbate the above numerical instability. The broad conclusion from the comparisons is that no method dominates the others. For example, for DGP 1, local polynomials outperform partitioning and regression splines at  $x_2$ , but partitioning appears more accurate at  $x_3$ . However, this is not a general boundary property: the order is reversed for DGP 3.

## A PROOF OF THEOREMS

Let  $C$  denote a generic positive constant that may take different values in different places. For scalars, vectors, or matrices, let  $|\cdot|$  be the Euclidean norm. Matrix inequalities are understood to be in the positive definite sense. Consecutive uses of the symbol  $\asymp$  are to be interpreted pairwise. All results below hold for the partitioning schemes considered here, as described in the text. For a generic cell  $P_j$ , let  $p_{j*}$ ,  $\bar{p}_j$ , and  $p_j^*$  be the vectors in  $\mathbb{R}^d$  giving the start, mid-point, and end of the cell, respectively, defined in distance to the origin.

Prior to proving the main results, it is convenient to take a nonsingular linear transformation of the polynomial basis. The estimator  $\hat{\mu}(x)$  is invariant to such rotations, thus without loss of generality we may 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  $\mathcal{X}$ . Recall that  $R(x)$  is ordered ascendingly in  $k \in \mathbb{Z}_+^d$  and  $\ell = 1, \dots, d$ . Define the one-to-one function  $g(k) : \mathbb{Z}_+^d \rightarrow \mathbb{N}$  that gives the index position of  $R(x)$  corresponding to entry  $x^k$ . Let  $g^* = \max_k \{g(k) : k \in \mathbb{Z}_+^d, |k| \leq K-1\}$ . Then  $R(x)$  is a  $g^* \times 1$  vector with element  $g(k)$  equal to  $x^k$  for all  $\{k \in \mathbb{Z}_+^d : |k| \leq K-1\}$ . As  $R(x)$  excludes duplicates and terms with order exceeding  $K-1$ , it follows that  $g^* \leq K^d$ .

Recall from the text that the interval endpoints  $p_{\ell,j-1}$  and  $p_{\ell,j}$ , for  $j = 1, \dots, J_n$ , define the partition of the  $\ell$ -dimension of  $\mathcal{X}$ , and let  $\bar{p}_{\ell,j} = (p_{\ell,j} + p_{\ell,j-1})/2 \in \mathbb{R}$  be the midpoint of each interval. Define the matrix functions  $D(a)$  to be the  $K \times K$  diagonal matrix with entries given by  $a^{-(v-1)}$ ,  $v = 1, \dots, K$  and  $L(b)$  to be the  $K \times K$  lower triangular with typical element  $\binom{u-1}{v-1} (-b)^{u-v}$ ,  $(u, v) \in \{1, \dots, K : u \geq v\}$ . We then take the polynomial basis to be given by

$$R_j(x) \equiv \mathbb{1}_{P_j}(x)R(x) = \mathbb{1}_{P_j}(x)S_K \bigotimes_{\ell=1}^d \{D(p_{\ell,j} - \bar{p}_{\ell,j}) L(\bar{p}_{\ell,j})r(x_\ell)\},$$

where  $\bigotimes_{\ell=1}^d$  represents the (repeated) Kronecker product. Each element of the product  $L(\bar{p}_{\ell,j})r(x_\ell)$  is (the binomial expansion of)  $(x_\ell - \bar{p}_{\ell,j})^{k_\ell}$ ,  $0 \leq k_\ell \leq K-1$ , and premultiplication by  $D(p_{\ell,j} - \bar{p}_{\ell,j})$  rescales appropriately. The matrix  $S_K$  is a  $g^* \times K^d$  selection matrix which removes duplicates and terms of order exceeding  $K-1$ . By properties of the selection matrix and Kronecker product this is equivalent to transforming the entire polynomial basis.

### 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), for  $s \leq K - 1$  the polynomial basis satisfies:*

$$\max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \|\partial^m R_j(\cdot)\|_\infty = O(J_n^s).$$

*Proof.* By construction of the partition, for  $x \in P_j$ ,  $|x - \bar{p}_j| \leq |p_j^* - \bar{p}_j| \asymp J_n^{-1}$ , therefore for a multi-index  $m$  such that  $|m| \leq K - 1$ :

$$\begin{aligned} |\partial^m R_j(x)|^2 &= \mathbf{1}_{P_j}(x) \sum_{|k| \leq K-1} \left\{ \partial^m \frac{(x - \bar{p}_j)^k}{(p_j^* - \bar{p}_j)^k} \right\}^2 \\ &= \mathbf{1}_{P_j}(x) \sum_{|k| \leq K-1} \mathbf{1}\{m \leq k\} \left\{ \frac{k!}{(k-m)!} \frac{(x - \bar{p}_j)^{k-m}}{(p_j^* - \bar{p}_j)^k} \right\}^2 \\ &= \left( \frac{1}{(p_j^* - \bar{p}_j)^m} \right)^2 \mathbf{1}_{P_j}(x) \sum_{|k| \leq K-1} \mathbf{1}\{m \leq k\} \left\{ \frac{k!}{(k-m)!} \frac{(x - \bar{p}_j)^{k-m}}{(p_j^* - \bar{p}_j)^{k-m}} \right\}^2 \\ &\leq C \left( \frac{1}{(p_j^* - \bar{p}_j)^m} \right)^2 = O(J_n^{2|m|}), \end{aligned}$$

uniformly in  $\{m : |m| \leq K - 1\}$ , and in particular for  $|m| \leq s \leq K - 1$ .  $\square$

**Lemma A.2.** *Define  $\mu_j(x) \equiv \mathbf{1}_{P_j}(x)\mu(x)$ , and following the definition in Eqn. (2),  $\partial^m \mu_j(x) \equiv \mathbf{1}_{P_j}(x)\partial^m \mu(x)$ . Under Assumptions 1(b) and 1(e), there is a non-random vector  $\beta_j^0$ , depending only on  $K$  and  $j$ , such that for  $s \leq S \wedge (K - 1)$ :*

$$\max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \|\partial^m \mu_j(x) - \partial^m R_j(x)' \beta_j^0\|_\infty = O(J_n^{-((S+1) \wedge K - s)}).$$

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

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

with constants which can be made uniform in the multi-index  $m$ ,  $s$ , and  $j$ . The terms of the summation are assumed to be ordered ascendingly in  $g(k)$ . It remains to construct  $\beta_j^0$  appropriately so that  $\partial^m R_j(x)' \beta_j^0$  is the Taylor approximation given in (A.1). Recall the notational conventions defined in footnote 2. For fixed  $m \in \mathbb{Z}_+^d$ ,  $|m| \leq s$ , any entry of  $\partial^m R_j(x)$  with  $k \leq m$  is zero. Thus, entry  $g(k)$  of  $\partial^m R_j(x)$  is given by:

$$\mathbf{1}\{m \leq k\} \frac{k!}{(k-m)!} \frac{(x - \bar{p}_j)^{k-m}}{(p_j^* - \bar{p}_j)^k}.$$

Next, for  $k \in \mathbb{Z}_+^d$  define the function  $\beta_j^0(k)$  as  $\beta_j^0(k) = \partial^k \mu_j(\bar{p}_j) (p_j^* - \bar{p}_j)^k / k!$ . As  $g(k)$  is one-to-one and returns the index of the entry corresponding to multi-index  $k$ , we can define the coefficient vector  $\beta_j^0$  as the  $g^* \times 1$  vector with entry  $e$  equal to  $\beta_j^0(g^{-1}(e))$ , for all entries  $e = 1, \dots, g^*$ . Therefore:

$$\begin{aligned} \partial^m R_j(x)' \beta_j^0 &= \sum_{|k| \leq S \wedge (K-1)} \mathbb{1}\{m \leq k\} \frac{k!}{(k-m)!} \frac{(x - \bar{p}_j)^{k-m}}{(p_j^* - \bar{p}_j)^k} \frac{1}{k!} \left( \partial^k \mu_j(\bar{p}_j) \right) (p_j^* - \bar{p}_j)^k \\ &= \sum_{|k| \leq S \wedge (K-1)} \mathbb{1}\{m \leq k\} \frac{1}{(k-m)!} (x - \bar{p}_j)^{k-m} \partial^k \mu_j(\bar{p}_j). \end{aligned}$$

By definition, the multi-index satisfies  $|k + k'| = |k| + |k'|$ , and so re-indexing the above sum by changing variables  $k' = k - m$ , we obtain

$$\partial^m R_j(x)' \beta_j^0 = \sum_{|k'+m| \leq S \wedge (K-1)} \frac{1}{k'!} \left( \partial^{k'+m} \mu_j(\bar{p}_j) \right) (x - \bar{p}_j)^{k'}.$$

This matches the Taylor series, hence subtracting from Eqn. (A.1) gives:

$$\begin{aligned} \max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \left\| \partial^m \mu_j(x) - \partial^m R_j(x)' \beta_j^0 \right\|_\infty &= O \left( \max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \sup_{x \in P_j} |x - \bar{p}_j|^{(S+1) \wedge K - |m|} \right) \\ &= O \left( J_n^{-((S+1) \wedge K - |s|)} \right), \end{aligned}$$

completing the proof.  $\square$

**Lemma A.3.** Recall that  $q_j = \mathbb{P}[X \in P_j]$  and  $\Omega_j = \mathbb{E}[R_j(X)R_j(X)'] / q_j$ . 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, followed by Assumption 1(b), properties of the Kronecker product, and the construction of the transformed basis, gives:

$$\begin{aligned} \Omega_j &= \frac{1}{q_j} \int_{\mathcal{X}} R_j(x) R_j(x)' f(x) dx \asymp J_n^d \int_{\mathcal{X}} R_j(x) R_j(x)' dx \\ &\asymp J_n^d S_K \bigotimes_{\ell=1}^d \left\{ \int_{p_{\ell,j-1}}^{p_{\ell,j}} r \left( \frac{x_\ell - \bar{p}_{\ell,j}}{p_{\ell,j} - \bar{p}_{\ell,j}} \right) r \left( \frac{x_\ell - \bar{p}_{\ell,j}}{p_{\ell,j} - \bar{p}_{\ell,j}} \right)' dx_\ell \right\} S_K' \\ &\asymp J_n^d \left( \prod_{\ell=1}^d |p_{\ell,j} - \bar{p}_{\ell,j}| \right) S_K \left\{ \bigotimes_{\ell=1}^d \int_{-1}^1 r(w) r(w)' dw \right\} S_K' \\ &\asymp S_K \left\{ \bigotimes_{\ell=1}^d \int_{-1}^1 r(w) r(w)' dw \right\} S_K', \end{aligned}$$

where we have used a change of variables and  $|p_{\ell,j} - p_{\ell,j-1}| \asymp J_n^{-1}$ . Changing variables again using the inversion of the centering and scaling performed by the matrices  $L(\cdot)$  and  $D(\cdot)$  yields

$$\begin{aligned}\Omega_j &\asymp S_K \left\{ \bigotimes_{\ell=1}^d \int_0^1 [D(2)L(-1)]^{-1} r(t)r(t)' [L(-1)D(2)]^{-1} dt \right\} S'_K \\ &\asymp S_K \left\{ \bigotimes_{\ell=1}^d [D(2)L(-1)]^{-1} H [L(-1)D(2)]^{-1} \right\} S'_K \asymp I_{g^*},\end{aligned}$$

where  $H$  denotes the Hilbert matrix of order  $K$ , which is positive definite.  $\square$

**Lemma A.4.** *Let  $a_n = n^{-1} J_n^d \log(J_n^d)$ , and recall  $\hat{\Omega}_j = R'_j R_j / (nq_j)$ . Under Assumption 1 and the rate restriction of Theorem 1:  $\max_{1 \leq j \leq J_n^d} |\hat{\Omega}_j - \Omega_j|^2 = O_p(a_n)$ . If, in addition,  $J_n^d \asymp (n/\log(n))^\gamma$ ,  $\gamma > 0$ , the same is true almost surely.*

*Proof.* For  $k, k' \in \mathbb{Z}_+^d$  with  $|k|, |k'| \leq K-1$ , let the  $(g(k), g(k'))$  element of  $(\hat{\Omega}_j - \Omega_j)$  be denoted  $\sum_{i=1}^n W_{ij}(k, k') / (nq_j)$ , where  $W_{ij}(k, k') = [R_j(X_i)R_j(X_i)']_{g(k), g(k')} - [\mathbb{E}[R_j(X_i)R_j(X_i)']]_{g(k), g(k')}$ . By Lemma A.1 (taking  $s = 0$ ),  $|W_{ij}(k, k')| < C$  and  $\mathbb{E}[W_{ij}(k, k')^2] \leq Cq_j$ , for any  $k, k'$ . Thus by Boole's inequality,  $K$  being fixed, Bernstein's inequality, and  $q_j \asymp J_n^{-d}$ :

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

which is arbitrarily small for  $\varepsilon$  large enough by the rate restriction of Theorem 1.

When  $J_n^d = C(n/\log(n))^\gamma$ , we use the above bound to write:  $\sum_{n=1}^\infty \mathbb{P}[\max_{1 \leq j \leq J_n^d} |\hat{\Omega}_j - \Omega_j| > (a_n)^{1/2} \varepsilon] \leq \sum_{n=1}^\infty C(n/\log(n))^{\gamma - C\gamma\varepsilon^2/(1+\sqrt{a_n}\varepsilon)} < \infty$ , where summability is ensured by choosing  $\varepsilon$  large enough and  $a_n \rightarrow 0$  by the rate restriction in Theorem 1. The conclusion follows by the Borel-Cantelli Lemma.  $\square$

**Lemma A.5.** *Let the conditions of Theorem 2 hold, and for  $\xi$  therein let  $r_n^2 = n^{-1} J_n^{d(2-\xi)} \log(J_n^d)^\xi$ . Then for  $G = (\mu(X_1), \dots, \mu(X_n))'$ , we have  $\max_{1 \leq j \leq J_n^d} |R'_j(Y-G)/(nq_j)| = O_p(r_n)$ . If, in addition,  $J_n^d \asymp (n/\log(n))^\gamma$ ,  $\gamma > 0$ , 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  $H_{ij}(k) = \mathbb{1}_{P_j}(X_i) [R_j(X_i)]_{g(k)} (Y_i \mathbb{1}\{Y_i \leq t_n\} - \mathbb{E}[Y_i \mathbb{1}\{Y_i \leq t_n\} | X_i])$  and  $T_{ij}(k) = \mathbb{1}_{P_j}(X_i) [R_j(X_i)]_{g(k)} (Y_i \mathbb{1}\{Y_i > t_n\} - \mathbb{E}[Y_i \mathbb{1}\{Y_i > t_n\} | X_i])$ .

For the truncated term, since  $|H_{ij}(k)| \leq t_n$  by construction and  $\mathbb{E}[H_{ij}(k)^2] \leq Cq_j$ , applying Bernstein's inequality and  $q_j \asymp J_n^{-d}$  we find that for fixed  $k \in \mathbb{Z}_+^d$ :

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

By  $\xi \in [0, 1]$  and the rate restriction of the Theorem, the above probability can be made arbitrarily small for  $\varepsilon$  large enough, as:

$$\frac{n}{J_n^d \log(J_n^d)} r_n^2 = \frac{J_n^{d(1-\xi)}}{\log(J_n^d)^{1-\xi}} \geq 1, \text{ and } \frac{t_n}{r_n} \frac{J_n^d \log(J_n^d)}{n} = \left( \frac{J_n^{d\xi(1+2/\eta)} \log(J_n^d)^{2-\xi(1+2/\eta)}}{n} \right)^{1/2} = O(1).$$

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

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

This is arbitrarily small for large enough  $\varepsilon$ , since:

$$\frac{J_n^{2d}}{nr_n^2 t_n^\eta} = \frac{J_n^{2d}}{n} \frac{n}{J_n^{d(2-\xi)} \log(J_n^d)^\xi} \frac{\log(J_n^d)^\xi}{J_n^{d\xi}} = 1.$$

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

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

is arbitrarily small for  $\varepsilon$  large enough.

The conclusion will hold almost surely by the Borel-Cantelli Lemma if we find sequences  $r_n \rightarrow 0$  and  $t_n \rightarrow \infty$  such that  $(r_n^2 n)/(J_n^d \log(J_n^d)) \not\rightarrow 0$ ,  $(t_n/r_n)(J_n^d \log(J_n^d))/n \not\rightarrow \infty$ , and,  $\sum_{n=1}^{\infty} J_n^{2d}/(nr_n^2 t_n^\eta) < \infty$ . For  $r_n$  in the statement of the Lemma, the first requirement is satisfied as above. For  $J_n^d \asymp (n/\log(n))^\gamma$  and  $t_n = n^\tau$ ,  $\tau > 0$ , the second and third conditions above require  $(1 + \xi\gamma)/\eta < \tau < (1 - \xi\gamma)/2$ . This interval is nonempty since by assumption  $\eta > 2 \left( \frac{1+\xi\gamma}{1-\xi\gamma} \right)$ .  $\square$

## A.2 CONVERGENCE RATES

*Proof of Theorem 1.* Define  $\mathbb{1}_{n,j} = \mathbb{1}\{\lambda_{\min}(\hat{\Omega}_j) \geq C\}$  for some positive constant  $C$ , where  $\lambda_{\min}(\hat{\Omega}_j)$  is the smallest eigenvalue. In the proofs that follow we will redefine the notation  $\hat{\mu}(x) = \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j} R_j(x)' \hat{\beta}_j$  (cf. Eqn. (1)). As  $\min_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} = 1$  w.p.a. 1 by Lemma A.4, this causes no problem in the proof or formation of the estimator. For  $\beta_j^0$  as in Lemma A.2 and  $G = (\mu(X_1), \dots, \mu(X_n))'$ :

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

The proof proceeds by bounding  $T_{n1}$ – $T_{n3}$ . To begin, observe that by properties of the trace operator, Assumption 1(c),  $R_j(R_j' R_j)^{-1} R_j'$  idempotent,  $K$  fixed, and  $q_j \asymp J_n^{-d}$ ,

$$\begin{aligned} \mathbb{E} \left[ \left| \mathbb{1}_{n,j} \hat{\Omega}_j^{-1/2} R_j' (Y - G) / (nq_j) \right|^2 \middle| \{X_i\} \right] &= \frac{\mathbb{1}_{n,j}}{nq_j} \text{tr} \left\{ \mathbb{E} \left[ (Y - G)' R_j (R_j' R_j)^{-1} R_j' (Y - G) \middle| \{X_i\} \right] \right\} \\ &= \frac{\mathbb{1}_{n,j}}{nq_j} \text{tr} \left\{ R_j (R_j' R_j)^{-1} R_j' \mathbb{E} \left[ (Y - G)(Y - G)' \middle| \{X_i\} \right] \right\} \\ &\leq C \frac{\mathbb{1}_{n,j}}{nq_j} \text{tr} \left\{ (R_j' R_j)^{-1} R_j' R_j \right\} \leq \frac{C}{nq_j} \leq \frac{C J_n^d}{n}. \quad (A.2) \end{aligned}$$

This bound is uniform in  $1 \leq j \leq J_n^d$ , and hence:

$$\mathbb{E} \left[ \sum_{j=1}^{J_n^d} q_j \mathbb{1}_{n,j} \left| \frac{\hat{\Omega}_j^{-1/2} R_j' (Y - G)}{nq_j} \right|^2 \right] \leq \max_{1 \leq j \leq J_n^d} \mathbb{E} \left[ \mathbb{1}_{n,j} \left| \frac{\hat{\Omega}_j^{-1/2} R_j' (Y - G)}{nq_j} \right|^2 \right] \sum_{j=1}^{J_n^d} q_j = O \left( J_n^d / n \right).$$

Hence by Markov's inequality, that  $q_j = \int_{P_j} f(x)dx$ , Lemmas A.1 and A.4, and because the differentiation only affects the basis at the point of evaluation, we have the following bound:

$$\begin{aligned} T_{n1} &\leq \left( \max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \|\partial^m R_j(\cdot)\|_\infty^2 \right) \left( \max_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} \left| \hat{\Omega}_j^{-1} \right| \right) \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j} \left| \hat{\Omega}_j^{-1/2} \frac{R_j'(Y-G)}{nq_j} \right|^2 \int_{P_j} f(x)dx \\ &= O(J_n^{2s}) O_p(1) O_p(J_n^d/n) = O_p(J_n^{d+2s}/n). \quad (\text{A.3}) \end{aligned}$$

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

$$\begin{aligned} \mathbb{P} \left[ \max_{1 \leq j \leq J_n^d} \sum_{i=1}^n (\mathbb{1}_{P_j}(X_i) - q_j) > nq_j \varepsilon \right] &\leq C J_n^d \max_{1 \leq j \leq J_n^d} \exp \left\{ -C \frac{nq_j \varepsilon^2}{1 + \varepsilon} \right\} \\ &\leq C \exp \left\{ \log(J_n^d) \left[ 1 - C \frac{n}{J_n^d \log(J_n^d)} \frac{\varepsilon^2}{1 + \varepsilon} \right] \right\} \rightarrow 0. \quad (\text{A.4}) \end{aligned}$$

Therefore, by  $R_j(R_j'R_j)^{-1}R_j'$  idempotent and Lemma A.2:

$$\begin{aligned} &\max_{1 \leq j \leq J_n^d} \left| \mathbb{1}_{n,j} \hat{\Omega}_j^{-1/2} R_j' (G - R_j \beta_j^0) / (nq_j) \right|^2 \\ &= \max_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} \left| (G - R_j \beta_j^0)' R_j (R_j'R_j)^{-1} R_j' (G - R_j \beta_j^0) / (nq_j) \right| \\ &\leq \max_{1 \leq j \leq J_n^d} \left| (G - R_j \beta_j^0)' (G - R_j \beta_j^0) / (nq_j) \right| \\ &= \max_{1 \leq j \leq J_n^d} \frac{1}{nq_j} \sum_{i=1}^n \mathbb{1}_{P_j}(X_i) (\mu(X_i) - R_j(X_i)' \beta_j^0)^2 \\ &\leq \max_{1 \leq j \leq J_n^d} \left\| \mathbb{1}_{P_j}(\cdot) (\mu(\cdot) - R_j(\cdot)' \beta_j^0) \right\|_\infty^2 \max_{1 \leq j \leq J_n^d} \frac{1}{nq_j} \sum_{i=1}^n \mathbb{1}_{P_j}(X_i) = O_p \left( J_n^{-2((S+1) \wedge K)} \right). \quad (\text{A.5}) \end{aligned}$$

And so for  $T_{n2}$ , by the above result, Lemmas A.1 and A.4, and  $\sum_{j=1}^{J_n^d} \int_{P_j} f(x)dx = 1$ , we have

$$\begin{aligned} T_{n2} &\leq \left( \max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \|\partial^m R_j(\cdot)\|_\infty^2 \right) \left( \max_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} \left| \hat{\Omega}_j^{-1} \right| \right) \\ &\quad \times \left( \max_{1 \leq j \leq J_n^d} \left| \mathbb{1}_{n,j} \hat{\Omega}_j^{-1/2} R_j' (G - R_j \beta_j^0) / (nq_j) \right|^2 \right) \sum_{j=1}^{J_n^d} \int_{P_j} f(x)dx \\ &\leq O(J_n^{2s}) O_p(1) O_p \left( J_n^{-2((S+1) \wedge K)} \right) = O_p \left( J_n^{-2((S+1) \wedge K - s)} \right). \quad (\text{A.6}) \end{aligned}$$

Finally, Lemma A.2 immediately gives:

$$T_{n3} \leq \max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \left\| \left( (\partial^m R_j(\cdot))' \beta_j^0 - \partial^m \mu_j(\cdot) \right) \right\|_\infty^2 \sum_{j=1}^{J_n^d} \int_{P_j} f(x) dx = O \left( J_n^{-2((S+1) \wedge K - s)} \right). \quad (\text{A.7})$$

Combining the bounds (A.3), (A.6), and (A.7), the result follows from  $\min_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} = 1$  w.p.a. 1 by Lemma A.4.  $\square$

*Proof of Theorem 2.* For  $\beta_j^0$  as in Lemma A.2:

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

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

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  $s \leq S \wedge (K - 1)$ :*

$$\max_{|m| \leq s} \left\| \partial^m \hat{\mu} - \partial^m \mu \right\|_\infty^2 = O_{as} \left( \frac{J_n^{(2-\xi)d+2s} \log(J_n^d)^\xi}{n} + J_n^{-2((S+1) \wedge K - s)} \right).$$

*Proof of Theorem A.1.* First observe that the rate restriction on  $J_n$  given implies that of Theorem 2. The exponential bound of (A.4) and  $n^{-1} J_n^d \log(J_n^d) \rightarrow 0$  gives  $\max_{1 \leq j \leq J_n^d} \sum_{i=1}^n \mathbb{1}_{P_j}(X_i) / (nq_j) = O_{as}(1)$ . Hence Eqn. (A.5) and the steps of Eqn. (A.6) hold almost surely. Coupled with the second conclusion in Lemma A.5, the proof of Theorem 2 can be strengthened to hold with probability one, as  $\min_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} = 1$  w.p.a. 1 using the almost sure Lemma A.4.  $\square$

### A.3 MEAN-SQUARE ERROR

We first give three Lemmas necessary for the proof of Theorem 3. Proofs of these results may be found in the supplemental appendix. Recall that for  $K = 1$ ,  $R'_j R_j = \sum_{i=1}^n \mathbb{1}_{P_j}(X_i)$  is the number of observations in  $P_j$ . Call this  $N_j$ . Further define  $N_{j,-i} = \sum_{l \neq i} \mathbb{1}_{P_j}(X_l)$  and  $N_{j,-i-l} = \sum_{m \neq i,l} \mathbb{1}_{P_j}(X_m)$ . Then  $N_j \sim \text{Bin}(n, q_j)$ ,  $N_{j,-i} \sim \text{Bin}(n-1, q_j)$ , and  $N_{j,-i-l} \sim \text{Bin}(n-2, q_j)$ .

**Lemma A.6.** *Let the conditions of Theorem 3 hold. For remainder terms uniform in  $1 \leq j \leq J_n^d$ ,*

$$\begin{aligned} \mathbb{E} \left[ \mathbb{1}\{N_j > 0\} \frac{1}{N_j} \right] &= \frac{1}{nq_j} + o\left(\frac{J_n^d}{n}\right); & \mathbb{E} \left[ \frac{1}{N_{j,-i} + 1} \right] &= \frac{1 - (1 - q_j)^n}{nq_j}; \\ \mathbb{E} \left[ \frac{1}{(N_{j,-i} + 1)^2} \right] &= \frac{1}{(nq_j)^2} \left( 1 + o\left(\frac{J_n^d}{n}\right) \right); \\ \mathbb{E} \left[ \frac{1}{(N_{j,-i-l} + 2)^2} \right] &= \frac{1}{n(n-1)q_j^2} \left( 1 - \frac{1}{nq_j} \left( 1 + o\left(\frac{J_n^d}{n}\right) \right) \right). \end{aligned}$$

**Lemma A.7.** *Let the conditions of Theorem 3 hold. If  $g(\cdot)$  is continuous on  $\mathcal{X} \subset \mathbb{R}$ , then:  $\int_{P_j} g(w)dw = g(\bar{p}_j) \text{vol}(P_j) + O(J_n^{-d-1})$ , where the order of the remainder is uniform in  $1 \leq j \leq J_n^d$ .*

**Lemma A.8.** *Let the conditions of Theorem 3 hold. If  $g(\cdot)$  is continuous on  $\mathcal{X} \subset \mathbb{R}$ , then:  $\sum_{j=1}^{J_n^d} g(\bar{p}_j) \text{vol}(P_j) = \int_{\mathcal{X}} g(w)dw + O(J_n^{-1})$ .*

*Proof of Theorem 3.* Let  $\mathbf{X} = (X_1, \dots, X_n)$  and expand as follows:

$$\int_{\mathcal{X}} \mathbb{E} \left[ (\hat{\mu}(x) - \mu(x))^2 \right] f(x)dx = \int_{\mathcal{X}} \left\{ \mathbb{E} [\mathbb{V}[\hat{\mu}(x) \mid \mathbf{X}]] + \mathbb{V}[\mathbb{E}[\hat{\mu}(x) \mid \mathbf{X}]] + (\mathbb{E}[\hat{\mu}(x)] - \mu(x))^2 \right\} f(x)dx.$$

We examine each term one at a time. The following two results will be used frequently. First observe that  $q_j = \int_{P_j} f(w)dw = f(\bar{p}_j) \text{vol}(P_j) + O(J^{-d-1})$ , by Lemma A.7. Further, under the conditions placed on the partition and Assumption 1(b),  $\text{vol}(P_j) = \text{vol}(\mathcal{X})/J_n^d$ .

Consider the first term. Using the two results above and Lemmas A.6, A.7, and A.8, we have:

$$\begin{aligned} \int_{\mathcal{X}} \mathbb{E} [\mathbb{V}[\hat{\mu}(x) \mid \mathbf{X}]] f(x)dx &= \sum_{j=1}^{J_n^d} \left( \int_{\mathcal{X}} \mathbb{1}_{P_j}(x) f(x)dx \right) \sum_{i=1}^n \mathbb{E} \left[ \frac{1}{(N_{j,-i} + 1)^2} \right] \mathbb{E} [\mathbb{1}_{P_j}(X_i) \sigma^2(X_i)] \\ &= \sum_{j=1}^{J_n^d} nq_j \left( \frac{1}{(nq_j)^2} + o\left(\left(\frac{J_n^d}{n}\right)^3\right) \right) \left( \int_{P_j} \sigma^2(w) f(w)dw \right) \\ &= \sum_{j=1}^{J_n^d} \frac{\sigma^2(\bar{p}_j) f(\bar{p}_j) \text{vol}(P_j)}{nq_j} \left( 1 + o\left(\frac{J_n^{2d}}{n^2}\right) \right) + O\left(\frac{J_n^{d-1}}{n}\right) \end{aligned}$$

$$= \frac{J_n^d}{|\mathcal{X}|n} \mathbb{E} \left[ \frac{\sigma^2(X)}{f(X)} \right] [1 + o(1)] + o(J_n^d/n). \quad (\text{A.8})$$

For the second variance term, define:  $\bar{\mu}_j \equiv \mathbb{E} [\mathbf{1}_{P_j}(X)\mu(X)]$  and  $\bar{N}_j \equiv \mathbb{E} [(N_{j,-i} + 1)^{-1}] = (1 - (1 - q_j)^n)/(nq_j)$ . Then:

$$\int_{\mathcal{X}} \mathbb{V} [\mathbb{E} [\hat{\mu}(x) | \mathbf{X}]] f(x) dx = \sum_{j=1}^{J_n^d} q_j n \mathbb{E} [\mathbf{1}_{P_j}(X)\mu(X)^2] \mathbb{E} \left[ \frac{1}{(N_{j,-i} + 1)^2} \right] \quad (\text{V}_{n1})$$

$$+ \sum_{j=1}^{J_n^d} q_j n(n-1) \bar{\mu}_j^2 \mathbb{E} \left[ \frac{1}{(N_{j,-i-l} + 2)^2} \right] \quad (\text{V}_{n2})$$

$$- \sum_{j=1}^{J_n^d} q_j (n \bar{\mu}_j \bar{N}_j)^2. \quad (\text{V}_{n3})$$

Now apply the two results above, as well as Lemmas A.6, A.7, and A.8, to get:

$$\begin{aligned} V_{1n} &= \sum_{j=1}^{J_n^d} nq_j \left( \int_{P_j} \mu(x)^2 f(w) dw \right) \frac{1}{(nq_j)^2} \left( 1 + o\left(\frac{J_n^d}{n}\right) \right) \\ &= \frac{1}{n} \sum_{j=1}^{J_n^d} \frac{\mu(\bar{p}_j)^2 f(\bar{p}_j) \text{vol}(P_j)}{f(\bar{p}_j) \text{vol}(P_j)} \left[ 1 + o\left(\frac{J_n^{2d}}{n^2}\right) \right] + O\left(\frac{J_n^{d-1}}{n}\right) \\ &= \frac{J_n^d}{|\mathcal{X}|n} \mathbb{E} \left[ \frac{\mu(X)^2}{f(X)} \right] [1 + o(1)] + o(J_n^d/n). \end{aligned} \quad (\text{A.9})$$

Similarly:

$$\begin{aligned} V_{2n} &= \sum_{j=1}^{J_n^d} q_j n(n-1) \left( \int_{P_j} \mu(w) f(w) dw \right)^2 \frac{1}{n(n-1)q_j^2} \left( 1 - \frac{1}{nq_j} (1 + o(J_n/n)) \right) \\ &= \sum_{j=1}^{J_n^d} \mu(\bar{p}_j)^2 f(\bar{p}_j) \text{vol}(P_j) - \frac{J_n^d}{|\mathcal{X}|n} \mathbb{E} [\mu(X)^2] (1 + o(1)) + o(J_n^d/n). \end{aligned} \quad (\text{A.10})$$

For the final term, similar steps give:

$$V_{3n} = - \sum_{j=1}^{J_n^d} \frac{1}{q_j} \mu(\bar{p}_j)^2 f(\bar{p}_j) \text{vol}(P_j) [1 + O(J_n^{-d-1}) + O((1 - q_j)^n)]. \quad (\text{A.11})$$

Finally, for the bias term, we first compute  $\mathbb{E}[\hat{\mu}(x)]$  using Lemmas A.6 and A.7.

$$\begin{aligned}\mathbb{E}[\hat{\mu}(x)] &= \sum_{j=1}^{J_n^d} \mathbf{1}_{P_j}(x) \mathbb{E} \left[ \frac{1}{N_{j,-i} + 1} \right] n \mathbb{E} [\mathbf{1}_{P_j}(X_i) \mu(X_i)] \\ &= \sum_{j=1}^{J_n^d} \mathbf{1}_{P_j}(x) \frac{1 - (1 - q_j)^n}{q_j} \int_{P_j} \mu(z) f(z) dz = \sum_{j=1}^{J_n^d} \mathbf{1}_{P_j}(x) \mu(\bar{p}_j) \left( 1 + O(J_n^{-d-1}) \right),\end{aligned}$$

where all remainder terms are uniform in  $1 \leq j \leq J_n^d$  and  $x \in P_j$ . Note that  $(p_{\ell,j} - \bar{p}_{\ell,j}) = -(p_{\ell,j-1} - \bar{p}_{\ell,j}) = (p_{\ell,j} - p_{\ell,j-1})/2 = |\mathcal{X}_\ell|/J_n$ . Then we have, applying Assumptions 1(e) and 1(b), the assumed continuity of the density  $f(x)$ , and Lemma A.8:

$$\begin{aligned}\int_{\mathcal{X}} (\mathbb{E}[\hat{\mu}(x)] - \mu(x))^2 f(x) dx &= \int_{\mathcal{X}} \sum_{j=1}^{J_n^d} \mathbf{1}_{P_j}(x) (\mu(\bar{p}_j) - \mu(x))^2 f(x) dx \left( 1 + O(J_n^{-d-1}) \right) \\ &= \sum_{j=1}^{J_n^d} f(\bar{p}_j) \nabla \mu(\bar{p}_j)' \int_{P_j} (x - \bar{p}_j)(x - \bar{p}_j)' dx \nabla \mu(\bar{p}_j) + o(J_n^{-2}) \\ &= J_n^{-2} \frac{1}{12} \sum_{j=1}^{J_n^d} f(\bar{p}_j) \nabla \mu(\bar{p}_j)' D_{\mathcal{X}} \nabla \mu(\bar{p}_j) \text{vol}(P_j) + o(J_n^{-2}) \\ &= J_n^{-2} \frac{1}{12} \mathbb{E} [\nabla \mu(X)' D_{\mathcal{X}} \nabla \mu(X)] [1 + o(1)] + o(J_n^{-2}).\end{aligned}\tag{A.12}$$

Adding Eqns. (A.8), (A.9), (A.10), (A.11), and (A.12) gives the result.  $\square$

#### A.4 BAHADUR REPRESENTATION AND ASYMPTOTIC NORMALITY

We first demonstrate a version of Theorem 4 that holds with probability one. It is convenient to first prove this result as many elements will be used to prove Theorem 4.

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

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

*Proof of Theorem A.2.* Under the linearity condition on  $\theta(\cdot)$  in Assumption 2, we can write the remainder  $\theta(\nu_n)$  from Eqn. (3) as

$$\theta(\nu_n) = \sum_{j=1}^{J_n^d} \Theta_j' \mathbf{1}_{n,j} \Omega_j^{-1} (\Omega_j - \hat{\Omega}_j) \hat{\Omega}_j^{-1} R_j' (Y - G) / (nq_j) \tag{T_{n1}}$$

$$+ \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j} \hat{\Omega}_j^{-1} R_j'(G - R_j \beta_j^0) / (nq_j) \quad (T_{n2})$$

$$+ \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j} (\Theta_j' \beta_j^0 - \theta(\mu_j)) \quad (T_{n3})$$

$$+ \sum_{j=1}^{J_n^d} (\mathbb{1}_{n,j} - 1) \left[ \theta(\mu_j) + \Theta_j' \Omega_j^{-1} R_j'(Y - G) / (nq_j) \right]. \quad (T_{n4})$$

Applying linearity and then continuity of the functional  $\theta(\cdot)$  from Assumption 2, followed by Lemmas A.1, A.3, A.4, and A.5 we have the following bound on  $|T_{n1}|$ :

$$\begin{aligned} |T_{n1}| &= \left| \theta \left( \sum_{j=1}^{J_n^d} (R_j(\cdot))' \mathbb{1}_{n,j} \Omega_j^{-1} (\Omega_j - \hat{\Omega}_j) \hat{\Omega}_j^{-1} \frac{R_j'(Y - G)}{nq_j} \right) \right| \\ &\leq C \max_{|m| \leq s} \left\| \sum_{j=1}^{J_n^d} (\partial^m R_j(\cdot))' \mathbb{1}_{n,j} \Omega_j^{-1} (\Omega_j - \hat{\Omega}_j) \hat{\Omega}_j^{-1} \frac{R_j'(Y - G)}{nq_j} \right\|_{\infty} \\ &\leq C \left( \max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \|\partial^m R_j(\cdot)\|_{\infty} \right) \left( \max_{1 \leq j \leq J_n^d} |\Omega_j - \hat{\Omega}_j| \right) \left( \max_{1 \leq j \leq J_n^d} |\mathbb{1}_{n,j} \hat{\Omega}_j^{-1}| \right) \\ &\quad \times \left( \max_{1 \leq j \leq J_n^d} |\Omega_j^{-1}| \right) \left( \max_{1 \leq j \leq J_n^d} \left| \frac{R_j'(Y - G)}{nq_j} \right| \right) \\ &= O_{as} \left( \frac{J_n^{d(3-\xi)/2+s} \log(J_n^d)^{(1+\xi)/2}}{n} \right). \end{aligned}$$

Following similar logic, by linearity, continuity, Lemma A.1, and the almost sure versions of Lemma A.4 and Eqn. (A.5):

$$\begin{aligned} |T_{n2}| &\leq \left( \max_{1 \leq j \leq J_n^d} \max_{|m| \leq s} \|\partial^m R_j(\cdot)\|_{\infty} \right) \left( \max_{1 \leq j \leq J_n^d} |\mathbb{1}_{n,j} \hat{\Omega}_j^{-1/2}| \right) \left( \max_{1 \leq j \leq J_n^d} \left| \mathbb{1}_{n,j} \hat{\Omega}_j^{-1/2} R_j' \frac{(G - R_j \beta_j^0)}{nq_j} \right| \right) \\ &= O_{as} \left( J_n^{-((S+1) \wedge K - s)} \right). \end{aligned}$$

Identical steps and Lemma A.2 give that  $|T_{n3}| = O_{as}(J_n^{-((S+1) \wedge K - s)})$ . Finally, from  $\min_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} = 1$  w.p.a. 1 it follows that  $T_{n4}$  is smaller order than the other terms. This completes the proof.  $\square$

*Proof of Theorem 4.* Use the same expansion as above. Apply the in probability versions of the same steps as above for  $T_{n2}$ ,  $T_{n3}$ , and  $T_{n4}$ , but for  $T_{n1}$  write:

$$T_{n1} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j} \Omega_j^{-1} (\Omega_j - \hat{\Omega}_j) \Omega_j^{-1} (\Omega_j - \hat{\Omega}_j) \hat{\Omega}_j^{-1} R_j(X_i) \varepsilon_i / q_j \quad (T_{n11})$$

$$-\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{J_n^d} \Theta'_j \mathbb{1}_{n,j} \Omega_j^{-1} (\hat{\Omega}_j - \Omega_j) \Omega_j^{-1} R_j(X_i) \varepsilon_i / q_j. \quad (T_{n12})$$

$T_{n11}$  is handled identically to  $T_{n1}$  above, so we find that

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

For  $T_{n12}$ , begin by defining  $W_j(i, l) = \mathbb{1}_{n,j} \Omega_j^{-1} (R_j(X_i) R_j(X_i)' - \mathbb{E}[R_j(X_i) R_j(X_i)']) \Omega_j^{-1} R_j(X_l) \varepsilon_l$ , so that we can write:  $T_{n12} = \sum_{j=1}^{J_n^d} \frac{1}{(nq_j)^2} \sum_{i=1}^n \sum_{l=1}^n \Theta'_j W_j(i, l)$ . Observe that  $\mathbb{E}[T_{n12}] = 0$  and that unless  $i = h$  and  $l = m$ ,  $\mathbb{E}[W_j(i, l) W_j(h, m)] = 0$ . By Lemmas A.1 and A.3, Assumption 1(c), and  $q_j \asymp J_n^{-d}$ , we have:

$$\begin{aligned} \max_{1 \leq j \leq J_n^d} \mathbb{E}[W_j(i, i) W_j(i, i)'] &\leq O(1) \max_{1 \leq j \leq J_n^d} \mathbb{E}[\mathbb{1}_{P_j}(X_i)] = O(J_n^{-d}), \text{ and} \\ \max_{1 \leq j \leq J_n^d} \mathbb{E}[W_j(i, l) W_j(i, l)'] &= O(J_n^{-2d}) \end{aligned} \quad (A.13)$$

Further note that Assumption 2 and Lemma A.1 give that:

$$\max_{1 \leq j \leq J_n^d} |\Theta_j| \leq C \max_{1 \leq j \leq J_n^d} \left( \max_{|m| \leq s} \|\partial^m R_j(\cdot)\|_\infty \right) = O(J_n^s). \quad (A.14)$$

Therefore the variance of  $T_{n2}$  may be bounded as follows, using  $q_j \asymp J_n^{-d}$ , Eqns. (A.13) and (A.14), linearity and continuity of  $\theta(\cdot)$ , and Lemma A.1:

$$\begin{aligned} \mathbb{E}[T_{n2}^2] &= \sum_{j=1}^{J_n^d} \frac{1}{(nq_j)^4} \sum_{i=1}^n \sum_{l=1}^n \Theta'_j \mathbb{E}[W_j(i, l) W_j(i, l)'] \Theta_j \\ &\leq \frac{C J_n^{4d}}{n^4} \sum_{j=1}^{J_n^d} \Theta'_j \{ n \mathbb{E}[W_j(i, l) W_j(i, l)'] + n(n-1) \mathbb{E}[W_j(i, l) W_j(i, l)'] \} \Theta_j \\ &= \frac{J_n^{4d}}{n^4} \theta \left( \sum_{j=1}^{J_n^d} R_j(\cdot)' \{ n \mathbb{E}[W_j(i, l) W_j(i, l)'] + n(n-1) \mathbb{E}[W_j(i, l) W_j(i, l)'] \} \Theta_j \right) \\ &\leq \frac{C J_n^{4d}}{n^4} \max_{|m| \leq s} \left\| \sum_{j=1}^{J_n^d} (\partial^m R_j(\cdot))' \{ n \mathbb{E}[W_j(i, l) W_j(i, l)'] + n(n-1) \mathbb{E}[W_j(i, l) W_j(i, l)'] \} \Theta_j \right\|_\infty \end{aligned}$$

$$\begin{aligned}
&\leq \frac{CJ_n^{4d}}{n^4} \left( \max_{1 \leq j \leq J_n^d} |\Theta_j| \right) \left( \max_{1 \leq j \leq J_n^d} n\mathbb{E} [W_j(i, l)W_j(i, l)'] + n(n-1)\mathbb{E} [W_j(i, l)W_j(i, l)'] \right) \\
&\quad \times \left( \max_{|m| \leq s} \max_{1 \leq j \leq J_n^d} \sup_{x \in P_j} (\partial^m R_j(\cdot)) \right) \\
&= O_p \left( J_n^{2d+2s}/n^2 \right).
\end{aligned}$$

Hence  $|T_{n2}| = O_p(J_n^{d+s}/n)$ , completing the proof.  $\square$

*Proof of Theorem 5(a).* By Assumption 1(c) and  $\sigma^2(x)$  bounded away from zero on  $\mathcal{X}$ ,  $\Gamma_j \asymp \Omega_j$ . Further, also applying  $q_j \asymp J_n^{-d}$  and Lemma A.3 we have:

$$V_n \asymp \|\Psi_n\|_2^2 \asymp J_n^d \sum_{j=1}^{J_n^d} |\Theta_j|^2. \quad (\text{A.15})$$

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 inequalities, Assumption 1(c),  $V_n \asymp \|\Psi_n\|_2^2$  by Eqn. (A.15), and the conditions of the Theorem,

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

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

For the second conclusion of 5(a), observe that by  $\mathbf{1}_{n,j} = 1$  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

$$\begin{aligned}
T_{n1} &= V_n^{-1} \hat{V}_n - V_n^{-1} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \Theta_j' \hat{\Omega}_j^{-1} \tilde{\Gamma}_j \hat{\Omega}_j^{-1} \Theta_j / q_j, \\
T_{n2} &= V_n^{-1} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \Theta_j' (\hat{\Omega}_j^{-1} + \Omega_j^{-1}) \tilde{\Gamma}_j (\hat{\Omega}_j^{-1} - \Omega_j^{-1}) \Theta_j / q_j,
\end{aligned}$$

$$T_{n3} = V_n^{-1} \sum_{j=1}^{J_n^d} \Theta_j' \Omega_j^{-1} \left( \tilde{\Gamma}_j - \Gamma_j \right) \Omega_j^{-1} \Theta_j / q_j,$$

and  $\tilde{\Gamma}_j = \sum_{i=1}^n R_j(X_i) R_j(X_i)' \varepsilon_i^2 / (nq_j)$ . First, expanding the squared terms,  $T_{n1}$  can be split into two terms, and upon applying Lemmas A.1 and A.4,  $q_j \asymp J_n^{-d}$ , Eqns. (A.4) and (A.15), and the condition of the Theorem, we find that

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

where the final line additionally uses Assumption 1(c) and the final relation of Eqn. (A.15) to give:

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

By Lemma A.1 and otherwise identical steps to the above, we get:

$$\mathbb{E} \left[ \frac{1}{V_n} \sum_{j=1}^{J_n^d} |\Theta_j|^2 |\tilde{\Gamma}_j| / q_j \right] \leq \frac{J_n^d}{V_n} \sum_{j=1}^{J_n^d} |\Theta_j|^2 \frac{1}{nq_j} \sum_{i=1}^n \mathbb{E} [ |R_j(X)|^2 \varepsilon_i^2 ] = O(1).$$

Therefore, applying Lemmas A.3 and A.4:

$$\begin{aligned} |T_{n2}| &= V_n^{-1} \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j} \Theta_j' (\hat{\Omega}_j^{-1} + \Omega_j^{-1}) \tilde{\Gamma}_j \Omega_j^{-1} (\hat{\Omega}_j - \Omega_j) \hat{\Omega}_j^{-1} \Theta_j / q_j \\ &\leq C \left( \max_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} |\hat{\Omega}_j^{-1}|^3 \wedge \max_{1 \leq j \leq J_n^d} |\Omega_j^{-1}|^3 \right) \left( \max_{1 \leq j \leq J_n^d} |\hat{\Omega}_j - \Omega_j| \right) V_n^{-1} \sum_{j=1}^{J_n^d} |\Theta_j|^2 |\tilde{\Gamma}_j| / q_j \\ &= O_p \left( \sqrt{J_n^d \log(J_n^d) / n} \right) = o_p(1). \end{aligned}$$

Finally, referring to the definitions in Eqn. (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  $\delta \in (0, 1)$ ,  $(a + b)^{(1+\delta)/2} \leq a^{(1+\delta)/2} + b^{(1+\delta)/2}$ , the  $c_r$  inequality, Jensen's inequality, Assumption 1(c), and the first relation of Eqn. (A.15):

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

Next, for the case of  $\eta \geq 2$  we utilize only the fourth moment to find that:

$$\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n T_{n3}(i) \right)^2 \right] = \frac{1}{n} \mathbb{E} [T_{n3}(i)^2] \leq \frac{1}{n} V_n^{-2} \mathbb{E} [\Psi_n(X_i)^4 \varepsilon_i^4] = O \left( \left( \frac{\|\Psi_n\|_4}{n^{1/4} \|\Psi_n\|_2} \right)^4 \right) \rightarrow 0,$$

again using Jensen's inequality, Assumption 1(c), and the first relation of Eqn. (A.15). In either case,  $T_{3n} = o_p(1)$  by Markov's inequality.  $\square$

*Proof of Theorem 5(b).* By Assumption 1(c), the Cauchy-Schwarz and triangle inequalities, and the conditions of the Theorem:

$$V_n - V = \mathbb{E}[(\Psi_n(X)^2 - \Psi(X)^2)\sigma^2(X)] \leq C \|\Psi_n - \Psi\|_2 (\|\Psi_n - \Psi\|_2 + 2\|\Psi\|_2) \rightarrow 0, \quad (\text{A.16})$$

whence the second conclusion.

Using the above result, the assumed mean-square convergence of  $\Psi_n(X)$ , and the remainder condition of the Theorem,

$$\begin{aligned} \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{V_n}} &= \sum_{i=1}^n \left[ \frac{\Psi(X_i)\varepsilon_i}{\sqrt{nV}} + \frac{(\Psi_n(X_i) - \Psi(X_i))\varepsilon_i}{\sqrt{nV}} + \frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV}} \left( \frac{\sqrt{V}}{\sqrt{V_n}} - 1 \right) \right] + \frac{\sqrt{n}\theta(\nu_n)}{\sqrt{V_n}} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\Psi(X_i)\varepsilon_i}{\sqrt{V}} + o_p(1). \end{aligned}$$

Convergence in distribution now follows under the assumed moment condition on  $\Psi(X)$  and a standard central limit 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,

$$T_{n3} = \left( \frac{1}{V_n} - \frac{1}{V} \right) \frac{1}{n} \sum_{i=1}^n \Psi_n(X_i)^2 \varepsilon_i^2 + \frac{1}{n} \sum_{i=1}^n \frac{[\Psi_n(X_i)^2 - \Psi(X_i)^2] \varepsilon_i^2}{V} + \frac{1}{nV} \sum_{i=1}^n (\Psi(X_i)^2 \varepsilon_i^2 - V),$$

where the first two terms tend to zero in probability by Eqn. (A.16) (and the steps therein) and Markov's inequality, and the third by the law of large numbers.  $\square$

## 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.
- Cattaneo, M. D., and M. H. Farrell, 2011, Efficient Estimation of the Dose Response Function under Ignorability using Subclassification on the Covariates. *Advances in Econometrics: Missing Data Methods*, 27, forthcoming.
- Chen, X., 2007, Large Sample Sieve Estimation of Semi-Nonparametric Models. In: J. Heckman, and E. Leamer, (Eds.), *Handbook of Econometrics*, vol. 6B of *Handbook of Econometrics*. Elsevier, chap. 76.
- Cochran, W. G., 1968, The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, 295–313.
- 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., and V. N. LaRiccia, 2009, *Maximum Penalized Likelihood Estimation, Volume II: Regression*. Springer.
- Fan, J., and I. Gijbels, 1996, *Local polynomial modelling and its applications*. Chapman and Hall, London.
- Györfi, L., M. Kohler, A. Krzyżak, and H. Walk, 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., and P. E. Todd, 2007, Implementing Nonparametric and Semiparametric Estimators. In: J. Heckman, and E. Leamer, (Eds.), *Handbook of Econometrics*, vol. 6B of *Handbook of Econometrics*. Elsevier, chap. 74.
- Imbens, G., and T. Lemieux, 2008, Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615–635.
- Imbens, G. W., and J. M. Wooldridge, 2009, Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47, 5–86.

- Kohler, M., A. Krzyżak, and H. Walk, 2006, Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis*, 97, 311–323.
- , 2009, Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139, 1286–1296.
- Kong, E., O. Linton, and Y. Xia, 2010, Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory*, 26, 1529–1564.
- Newey, W. K., 1994, Kernel Estimation of Partial Means and a General Variance Estimator. *Econometric Theory*, 10, 233–253.
- , 1997, Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79, 147–168.
- Newey, W. K., and D. L. McFadden, 1994, Large sample estimation and hypothesis testing. In: R. F. Engle, and D. McFadden, (Eds.), *Handbook of Econometrics*, vol. 4 of *Handbook of Econometrics*. Elsevier, chap. 36, pp. 2111–2245.
- Rosenbaum, P. R., and D. B. Rubin, 1983, On the Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, 41–55.
- , 1984, Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79, 516–524.
- 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.

Table 1: Root Mean-Square Error Compared to Alternative Estimators  
DGP 1,  $n=1,000$

	$x_1 = (0, 0)$			$x_2 = (-1, 0)$			$x_3 = (-1.9, 0)$		
	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE
<b>Partitioning Estimator</b>									
$J_n = 3$	-6.226	0.045	6.23	-0.556	0.014	0.568	2.144	0.014	2.148
$J_n = 5$	-4.653	0.128	4.667	0.45	0.035	0.487	1.454	0.035	1.466
$J_n = 7$	-3.306	0.194	3.335	-0.346	0.061	0.425	1.002	0.068	1.035
$J_n = 3$	-6.208	0.04	6.212	-0.004	0.017	0.131	1.199	0.039	1.215
$J_n = 5$	-4.593	0.134	4.608	-0.011	0.05	0.224	0.821	0.098	0.879
$J_n = 7$	-3.181	0.228	3.217	-0.002	0.109	0.331	0.535	0.17	0.676
$J_n = 3$	-4.006	0.112	4.02	-0.218	0.03	0.279	0.983	0.078	1.022
$J_n = 5$	-1.777	0.187	1.829	0.116	0.096	0.331	0.404	0.158	0.566
<b>Local Polynomials</b>									
$h_n = 4/3$	-6.782	0.017	6.784	0.404	0.01	0.416	2.069	0.009	2.071
$h_n = 4/5$	-5.801	0.055	5.806	0.037	0.013	0.119	1.546	0.023	1.554
$h_n = 4/7$	-4.801	0.105	4.812	0.00	0.022	0.149	1.2	0.042	1.218
$h_n = 4/3$	-6.778	0.014	6.779	0.28	0.006	0.29	1.36	0.026	1.37
$h_n = 4/5$	-5.786	0.051	5.79	0.036	0.01	0.106	1.093	0.067	1.123
$h_n = 4/7$	-4.775	0.103	4.786	0.001	0.019	0.137	0.858	0.105	0.917
$h_n = 4/3$	-5.61	0.048	5.614	-0.021	0.014	0.121	1.168	0.051	1.189
$h_n = 4/5$	-3.862	0.116	3.877	-0.033	0.031	0.18	0.738	0.101	0.804
<b>Regression Splines</b>									
$J_n = 3$	-6.683	0.02	6.685	0.424	0.014	0.44	0.962	0.035	0.98
$J_n = 5$	-5.51	0.075	5.517	-0.072	0.028	0.181	1.311	0.094	1.346
$J_n = 7$	-4.258	0.159	4.277	-0.463	0.066	0.529	0.585	0.132	0.689

Table 2: Root Mean-Square Error Compared to Alternative Estimators  
DGP 2,  $n=1,000$

	$x_1 = (0, 0)$			$x_2 = (-1, 0)$			$x_3 = (-1.9, 0)$		
	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE
<b>Partitioning Estimator</b>									
$J_n = 3$	-1.432	0.016	1.438	-0.315	0.01	0.331	0.585	0.01	0.594
$J_n = 5$	-0.974	0.041	0.995	0.212	0.028	0.27	0.317	0.027	0.357
$J_n = 7$	-0.653	0.071	0.705	-0.147	0.053	0.272	0.2	0.054	0.307
$J_n = 3$	-1.426	0.014	1.431	-0.004	0.016	0.128	0.055	0.031	0.184
$J_n = 5$	-0.961	0.039	0.981	-0.007	0.05	0.223	0.033	0.075	0.275
$J_n = 7$	-0.623	0.075	0.68	-0.002	0.109	0.331	0.019	0.141	0.375
$J_n = 3$	-0.728	0.042	0.756	-0.018	0.029	0.171	0.046	0.062	0.253
$J_n = 5$	-0.259	0.111	0.421	-0.004	0.093	0.305	0.013	0.144	0.38
<b>Local Polynomials</b>									
$h_n = 4/3$	-1.578	0.006	1.58	0.259	0.005	0.269	0.495	0.006	0.501
$h_n = 4/5$	-1.303	0.016	1.309	0.018	0.01	0.1	0.288	0.016	0.315
$h_n = 4/7$	-1.017	0.029	1.031	-0.002	0.018	0.135	0.192	0.03	0.259
$h_n = 4/3$	-1.578	0.005	1.579	0.179	0.004	0.19	0.043	0.02	0.147
$h_n = 4/5$	-1.298	0.014	1.304	0.018	0.009	0.098	0.052	0.046	0.221
$h_n = 4/7$	-1.011	0.027	1.024	0.00	0.018	0.135	0.035	0.076	0.277
$h_n = 4/3$	-1.256	0.016	1.263	-0.093	0.012	0.144	0.054	0.038	0.202
$h_n = 4/5$	-0.72	0.039	0.747	-0.025	0.03	0.176	0.026	0.081	0.286
<b>Regression Splines</b>									
$J_n = 3$	-1.722	0.014	1.726	0.231	0.011	0.254	-0.052	0.033	0.189
$J_n = 5$	-1.555	0.035	1.566	-0.033	0.023	0.157	0.331	0.071	0.425
$J_n = 7$	-0.561	0.063	0.614	-0.24	0.049	0.326	-0.177	0.111	0.378

Table 3: Root Mean-Square Error Compared to Alternative Estimators  
DGP 3,  $n=1,000$

	$x_1 = (0, 0)$			$x_2 = (-1, 0)$			$x_3 = (-1.9, 0)$		
	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE
<b>Partitioning Estimator</b>									
$J_n = 3$	-0.114	0.01	0.151	1.621	0.043	1.634	-1.358	0.043	1.374
$J_n = 5$	-0.057	0.029	0.179	-0.203	0.031	0.268	-0.086	0.083	0.3
$J_n = 7$	-0.034	0.054	0.235	0.748	0.092	0.807	0.474	0.136	0.601
$J_n = 3$	-0.113	0.009	0.148	0.388	0.023	0.417	0.752	0.109	0.822
$J_n = 5$	-0.058	0.028	0.178	0.058	0.054	0.24	0.58	0.283	0.787
$J_n = 7$	-0.031	0.06	0.246	0.019	0.115	0.34	0.123	0.409	0.651
$J_n = 3$	-0.024	0.036	0.19	0.618	0.048	0.656	0.408	0.204	0.609
$J_n = 5$	-0.009	0.104	0.323	0.029	0.093	0.306	-0.091	0.237	0.495
<b>Local Polynomials</b>									
$h_n = 4/3$	-0.088	0.004	0.107	0.869	0.014	0.877	-0.805	0.026	0.821
$h_n = 4/5$	-0.099	0.01	0.141	0.607	0.024	0.626	-0.053	0.054	0.239
$h_n = 4/7$	-0.064	0.02	0.155	0.351	0.031	0.393	0.305	0.094	0.432
$h_n = 4/3$	-0.09	0.003	0.107	0.982	0.008	0.986	0.713	0.089	0.773
$h_n = 4/5$	-0.099	0.009	0.139	0.606	0.013	0.616	0.1	0.156	0.407
$h_n = 4/7$	-0.064	0.019	0.153	0.347	0.021	0.376	-0.176	0.197	0.478
$h_n = 4/3$	-0.157	0.011	0.19	0.322	0.016	0.346	-0.116	0.125	0.372
$h_n = 4/5$	-0.028	0.033	0.184	0.095	0.032	0.202	-0.341	0.118	0.484
<b>Regression Splines</b>									
$J_n = 3$	-0.295	0.011	0.314	0.246	0.016	0.277	0.284	0.241	0.568
$J_n = 5$	-0.401	0.023	0.429	0.271	0.053	0.355	-0.163	0.477	0.71
$J_n = 7$	0.258	0.046	0.335	0.226	0.075	0.354	0.807	0.46	1.054