On Binscatter

Supplemental Appendix

Matias D. Cattaneo*   Richard K. Crump†   Max H. Farrell‡   Yingjie Feng§

February 25, 2019

Abstract

This supplement collects all technical proofs, more general theoretical results than those reported in the main paper, and other methodological and numerical results. New theoretical results for partitioning-based series estimation are obtained that may be of independent interest. See also `Stata` and `R` companion software available at

https://sites.google.com/site/nppackages/binsreg/

where replication files are also provided.

---

*Department of Economics and Department of Statistics, University of Michigan.

†Capital Markets Function, Federal Reserve Bank of New York.

‡Booth School of Business, University of Chicago.

§Department of Economics, University of Michigan.

Contents

SA-1   Setup

This section repeats the setup in the main paper, and introduce some notation for the main analysis.

Suppose that $\{(y_i, x_i, \mathbf{w}_i) : 1 \leq i \leq n\}$ is a random sample satisfying the following regression model

$$y_i = \mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma} + \epsilon_i, \qquad \mathbb{E}[\epsilon_i | x_i, \mathbf{w}_i] = 0, \tag{SA-1.1}$$

where y_i is a scalar response variable, x_i is a scalar covariate, $\mathbf{w}_i$ is a vector of additional control variables of dimension d, and the parameter of interest is the nonparametric component $\mu(\cdot)$.

Binscatter estimators are usually constructed based on quantile-spaced partitions. Specifically, the relevant support of x_i is partitioned into J disjoint intervals employing the empirical quantiles, leading to the partitioning scheme $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \widehat{\mathcal{B}}_2, \ldots, \widehat{\mathcal{B}}_J\}$, where

$$\widehat{\mathcal{B}}_j = \begin{cases} \left[x_{(1)}, x_{(\lfloor n/J \rfloor)}\right) & \text{if } j = 1 \\ \left[x_{(\lfloor (j-1)n/J \rfloor)}, x_{(\lfloor jn/J \rfloor)}\right) & \text{if } j = 2, 3, \ldots, J-1 \\ \left[x_{(\lfloor (J-1)n/J \rfloor)}, x_{(n)}\right] & \text{if } j = J \end{cases},$$

$x_{(i)}$ denotes the i-th order statistic of the sample $\{x_1, x_2, \ldots, x_n\}$, and $\lfloor \cdot \rfloor$ is the floor operator. The number of bins J will play the role of tuning parameter for the binscatter method, and is assumed to diverge: $J \to \infty$ as $n \to \infty$ throughout the supplement, unless explicitly stated otherwise.

In the main paper, the p-th order piecewise polynomial basis, for some choice of $p = 0, 1, 2, \ldots,$ is defined as

$$\widehat{\mathbf{b}}(x) = \left[\begin{array}{cccc} \mathbb{1}_{\widehat{\mathcal{B}}_1}(x) & \mathbb{1}_{\widehat{\mathcal{B}}_2}(x) & \cdots & \mathbb{1}_{\widehat{\mathcal{B}}_J}(x) \end{array} \right]' \otimes \left[\begin{array}{cccc} 1 & x & \cdots & x^p \end{array} \right]',$$

where $\mathbb{1}_{\mathcal{A}}(x) = \mathbb{1}(x \in \mathcal{A})$ with $\mathbb{1}(\cdot)$ denoting the indicator function, and $\otimes$ is tensor product operator. Without loss of generality, we redefine $\widehat{\mathbf{b}}(x)$ as a *standardized rotated* basis for convenience of analysis. Specifically, for each $\alpha = 0, \ldots, p$, and $j = 1, \ldots, J$, the polynomial basis of degree α supported on $\widehat{\mathcal{B}}_j$ is rotated and rescaled:

$$\mathbb{1}_{\widehat{\mathcal{B}}_j}(x)x^\alpha \quad \mapsto \quad \sqrt{J} \cdot \mathbb{1}_{\widehat{\mathcal{B}}_j}(x) \left(\frac{x - x_{(\lfloor (j-1)n/J \rfloor)}}{\hat{h}_j} \right)^\alpha,$$

where $\hat{h}_j = x_{(\lfloor jn/J \rfloor)} - x_{(\lfloor (j-1)n/J \rfloor)}$. Thus, each local polynomial is centered at the start of each

bin and scaled by the length of the bin. $\sqrt{J}$ is an additional scaling factor which will help simplify some expressions of our results. We maintain the notation $\widehat{\mathbf{b}}(x)$ for this redefined basis, since it is equivalent to the original one in the sense that they represent the same (linear) function space.

Imposing the restriction that the estimated function is $(s-1)$-times continuously differentiable for $1 \leq s \leq p$, we introduce a new basis

$$\widehat{\mathbf{b}}_s(x) = (\widehat{b}_{s,1}(x), \ldots, \widehat{b}_{s,K_s}(x))' = \widehat{\mathbf{T}}_s \widehat{\mathbf{b}}(x), \qquad K_s = [(p+1)J - s(J-1)],$$

where $\widehat{\mathbf{T}}_s := \widehat{\mathbf{T}}_s(\widehat{\Delta})$ is a $K_s \times (p+1)J$ matrix depending on $\widehat{\Delta}$, which transforms a piecewise polynomial basis to a smoothed binscatter basis. When $s = 0$, we let $\widehat{\mathbf{T}}_0 = \mathbf{I}_{(p+1)J}$, the identity matrix of dimension $(p+1)J$. Thus $\widehat{\mathbf{b}}_0(x) = \widehat{\mathbf{b}}(x)$, the discontinuous basis without any constraints. When $s = p$, $\widehat{\mathbf{b}}_p(x)$ is the well-known B-spline basis of order $p+1$ with simple knots. When $0 < s < p$, they can be defined similarly as B-splines with knots of certain multiplicities. See Definition 4.1 in Section 4 of Schumaker (2007) for more details. Note that we require $s \leq p$, since if $s = p+1$, $\widehat{\mathbf{b}}_s(x)$ reduces to a global polynomial basis of degree p.

A key feature of the transformation matrix $\widehat{\mathbf{T}}_s$ that will be employed in the analysis is that on every row it has *at most* $(p+1)^2$ nonzeros, and on every column it has *at most* $p+1$ nonzeros. The expression of these elements is very cumbersome. The proof of Lemma SA-2.2 describes the structure of $\widehat{\mathbf{T}}_s$ in more detail, and provides an explicit representation for $\widehat{\mathbf{T}}_p$.

Given such a choice of basis, a covariate-adjusted (generalized) binscatter estimator is

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})' = \arg\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \sum_{i=1}^{n}(y_i - \widehat{\mathbf{b}}_s(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma})^2, \qquad s \leq p. \qquad \text{(SA-1.2)}$$

where $\widehat{\mathbf{b}}_s^{(v)}(x) = d^v \widehat{\mathbf{b}}_s(x)/dx^v$ for some $v \in \mathbb{Z}_+$ such that $v \leq p$.

SA-1.1 Assumptions

We impose the following assumption on the data generating process, which is more general than the one presented in the main paper. We use $\lambda_{\min}(\mathbf{A})$ to denote the minimum eigenvalue of a square matrix $\mathbf{A}$.

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

(i) $\{(y_i, x_i, \mathbf{w}_i) : 1 \leq i \leq n\}$ *are i.i.d satisfying* (SA-1), *and* x_i *follows a distribution function* $F(x)$ *with a continuous (Lebesgue) density* $f(x)$ *bounded away from zero;* $\mu(\cdot)$ *is* $(p+1)$-*times continuously differentiable;*

(ii) $\sigma^2(x) := \mathbb{E}[\epsilon_i^2 | x_i = x]$ *is continuous and bounded away from zero, and* $\sup_{x \in \mathcal{X}} \mathbb{E}[|\epsilon_i|^\nu | x_i = x] \lesssim 1$ *for some* $\nu > 2$;

(iii) $\mathbb{E}[\mathbf{w}_i | x_i = x]$ *is* ς-*times continuously differentiable for some* $\varsigma \geq 1$, $\sup_{x \in \mathcal{X}} \mathbb{E}[\|\mathbf{w}_i\|^\nu | x_i = x] \lesssim 1$, $\mathbb{E}[\|\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i | x_i]\|^4 | x_i] \lesssim 1$, $\lambda_{\min}(\mathbb{E}[(\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i | x_i])(\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i | x_i])' | x_i]) \gtrsim 1$, *and* $\mathbb{E}[\epsilon_i^2 | \mathbf{w}_i, x_i] \lesssim 1$.

Part (i) and (ii) are standard conditions employed in nonparametric series literature (Cattaneo, Farrell, and Feng, 2018, and references therein). Part (iii) includes a set of conditions similar to those used in Cattaneo, Jansson, and Newey (2018a,b) to analyze the semiparametric partially linear regression model, which ensures the negligibility of the estimation error of $\widehat{\gamma}$.

The conditions in Part (i) imply that the (relevant) support of x_i, denoted as $\mathcal{X}$, is a compact interval. Without loss of generality, it is normalized to $[0, 1]$ throughout the supplemental appendix.

SA-1.2 Notation

For vectors, $\| \cdot \|$ denotes the Euclidean norm, $\| \cdot \|_\infty$ denotes the sup-norm, and $\| \cdot \|_0$ denotes the number of nonzeros. For matrices, $\| \cdot \|$ is the operator matrix norm induced by the L_2 norm, and $\| \cdot \|_\infty$ is the matrix norm induced by the supremum norm, i.e., the maximum absolute row sum of a matrix. For a square matrix $\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ are the maximum and minimum eigenvalues of $\mathbf{A}$, respectively. We will use S^L to denote the unit circle in $\mathbb{R}^L$, i.e., $\|\mathbf{a}\| = \mathbf{1}$ for any $\mathbf{a} \in S^L$. For a real-valued function $g(\cdot)$ defined on a measure space $\mathcal{Z}$, let $\|g\|_{Q,2} := (\int_{\mathcal{Z}} |g|^2 dQ)^{1/2}$ be its L_2-norm with respect to the measure Q. In addition, let $\|g\|_\infty = \sup_{z \in \mathcal{Z}} |g(z)|$ be L_∞-norm of $g(\cdot)$, and $g^{(v)}(z) = d^v g(z)/dz^v$ be the vth derivative for $v \geq 0$.

For sequences of numbers or random variables, we use $a_n \lesssim b_n$ to denote that $\limsup_n |a_n/b_n|$ is finite, $a_n \lesssim_{\mathbb{P}} b_n$ or $a_n = O_{\mathbb{P}}(b_n)$ to denote $\limsup_{\varepsilon \to \infty} \limsup_n \mathbb{P}[|a_n/b_n| \geq \varepsilon] = 0$, $a_n = o(b_n)$ implies $a_n/b_n \to 0$, and $a_n = o_{\mathbb{P}}(b_n)$ implies that $a_n/b_n \to_{\mathbb{P}} 0$, where $\to_{\mathbb{P}}$ denotes convergence in probability. $a_n \asymp b_n$ implies that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two random variables X and Y, $X =_d Y$ implies that they have the same probability distribution.

3

We employ standard empirical process notation: $\mathbb{E}_n[g(\mathbf{x}_i)] = \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i)$, and $\mathbb{G}_n[g(\mathbf{x}_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n (g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i)])$. In addition, we employ the notion of covering number extensively in the proofs. Specifically, given a measurable space $(S, \mathcal{S})$ and a suitably measurable class of functions $\mathcal{G}$ mapping S to $\mathbb{R}$ equipped with a measurable envelop function $\bar{G}(z) \geq \sup_{g \in \mathcal{G}} |g(z)|$. The *covering number* of $N(\mathcal{G}, L_2(Q), \varepsilon)$ is the minimal number of $L_2(Q)$-balls of radius ε needed to cover $\mathcal{G}$. The covering number of $\mathcal{G}$ relative to the envelop is denoted as $N(\mathcal{G}, L_2(Q), \varepsilon\|\bar{G}\|_{Q,2})$.

Given the random partition $\widehat{\Delta}$, we will use the notation $\mathbb{E}_{\widehat{\Delta}}[\cdot]$ to denote that the expectation is taken with the partition $\widehat{\Delta}$ understood as fixed. To further simplify notation, we let $\{\hat{\tau}_0 \leq \hat{\tau}_1 \leq \cdots \leq \hat{\tau}_J\}$ denote the empirical quantile sequence employed by $\widehat{\Delta}$. Accordingly, let $\{\tau_0 \leq \cdots \leq \tau_J\}$ be the population quantile sequence, i.e., $\tau_j = F^{-1}(j/J)$ for $0 \leq j \leq J$. Then $\Delta = \{\mathcal{B}_1, \ldots, \mathcal{B}_J\}$ denotes the partition based on population quantiles, i.e.,

$$
\mathcal{B}_j = \begin{cases} \left[\tau_0, \tau_1\right) & \text{if } j = 1 \\ \left[\tau_{j-1}, \tau_j\right) & \text{if } j = 2, 3, \ldots, J-1 \\ \left[\tau_{J-1}, \tau_J\right] & \text{if } j = J \end{cases} .
$$

Let $h_j = F^{-1}(j/J) - F^{-1}((j-1)/J)$ be the width of $\mathcal{B}_j$. $\mathbf{b}_s(x)$ denotes the (smooth) binscatter basis based on the *nonrandom* partition Δ. Moreover, x_i's are collected in a matrix $\mathbf{X} = [x_1, \ldots, x_n]'$, all the data are collected in $\mathbf{D} = \{(y_i, x_i, \mathbf{w}_i') : 1 \leq i \leq n\}$, $\lceil z \rceil$ outputs the smallest integer no less than z and $a \wedge b = \min\{a, b\}$.

Finally, we sometimes write $\mathbf{b}_s(x; \bar{\Delta}) = (b_{s,1}(x; \bar{\Delta}), \ldots, b_{s,K_s}(x; \bar{\Delta}))'$ to emphasize a binscatter basis is constructed based on a particular partition $\bar{\Delta}$. Clearly, $\widehat{\mathbf{b}}_s(x) = \mathbf{b}_s(x; \widehat{\Delta})$ and $\mathbf{b}_s(x) = \mathbf{b}_s(x; \Delta)$.

SA-2 Technical Lemmas

This section collects a set of technical lemmas, which are key ingredients of our main theorems. The following expression of the coefficient estimators, also known as "backfitting" in statistics literature, will be quite convenient for theoretical analysis:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\gamma}}), \qquad \widehat{\boldsymbol{\gamma}} = (\mathbf{W}'\mathbf{M_B}\mathbf{W})^{-1}(\mathbf{W}'\mathbf{M_B}\mathbf{Y})$$

where $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\mathbf{B} = (\widehat{\mathbf{b}}_s(x_1), \ldots, \widehat{\mathbf{b}}_s(x_n))'$, $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)'$, $\mathbf{M_B} = \mathbf{I}_n - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. It is well known that the least squares estimator provides a best linear approximation to the target function. For any given partition $\bar{\Delta}$, the population least squares estimator is defined as

$$\boldsymbol{\beta}_\mu(\bar{\Delta}) := \arg\min_{\boldsymbol{\beta}} \mathbb{E}[(\mu(x_i) - \mathbf{b}_s(x_i; \bar{\Delta})'\boldsymbol{\beta})^2].$$

Accordingly, $r_\mu(x; \bar{\Delta}) = \mu(x) - \mathbf{b}_s(x; \bar{\Delta})'\boldsymbol{\beta}_\mu(\bar{\Delta})$ denotes the L_2 approximation error. We let $\widehat{\boldsymbol{\beta}}_\mu :=$ $\boldsymbol{\beta}_\mu(\widehat{\Delta})$, $\boldsymbol{\beta}_\mu := \boldsymbol{\beta}_\mu(\Delta)$, $\widehat{r}_\mu(x) := r_\mu(x; \widehat{\Delta})$ and $r_\mu(x) := r_\mu(x; \Delta)$.

In addition, we introduce the following definitions:

$$\widehat{\mathbf{Q}} := \widehat{\mathbf{Q}}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'], \quad \mathbf{Q} := \mathbf{Q}(\Delta) := \mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'],$$

$$\widehat{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{\Sigma}}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}})^2],$$

$$\bar{\boldsymbol{\Sigma}} := \bar{\boldsymbol{\Sigma}}(\widehat{\Delta}) := \mathbb{E}_n\left[\mathbb{E}\left[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i^2\Big|\mathbf{X}\right]\right], \quad \boldsymbol{\Sigma} := \boldsymbol{\Sigma}(\Delta) := \mathbb{E}\left[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\epsilon_i^2\right],$$

$$\bar{\Omega}(x) := \bar{\Omega}(x; \widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\bar{\boldsymbol{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x), \text{ and}$$

$$\Omega(x) := \Omega(x; \widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}^{-1}\boldsymbol{\Sigma}\mathbf{Q}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x).$$

All quantities with $\widehat{}$ or $\bar{}$ depend on the random partition $\widehat{\Delta}$, and those without any accents are nonrandom with the only exception of $\Omega(x)$, where the basis $\widehat{\mathbf{b}}_s^{(v)}(x)$ still depends on $\widehat{\Delta}$.

The asymptotic properties of partitioning-based estimators rely on a partition that is not be too "irregular" (Cattaneo, Farrell, and Feng, 2018). In the binscatter setting, we let $\bar{f} = \sup_{x \in \mathcal{X}} f(x)$ and $\underline{f} = \inf_{x \in \mathcal{X}} f(x)$, and for any partition $\bar{\Delta}$ with J bins, we let $h_j(\bar{\Delta})$ denote the length of the jth bin in $\bar{\Delta}$. Then, we introduce the family of partitions:

$$\Pi = \left\{\bar{\Delta} : \frac{\max_{1 \leq j \leq J} h_j(\bar{\Delta})}{\min_{1 \leq j \leq J} h_j(\bar{\Delta})} \leq \frac{3\bar{f}}{\underline{f}}\right\}. \tag{SA-2.1}$$

Intuitively, if a partition belongs to Π, then the lengths of its bins do not differ "too" much, a property usually referred to as "quasi-uniformity" in approximation theory. Our first lemma shows that a quantile-spaced partition possesses this property with probability approaching one.

Lemma SA-2.1 (Quasi-Uniformity of Quantile-Spaced Partitions). *Suppose that Assumption SA-1(i) holds. If $\frac{J \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then (i) $\max_{1 \leq j \leq J} |\hat{h}_j - h_j| \lesssim_{\mathbb{P}} J^{-1}\sqrt{J \log J / n}$, and (ii) $\widehat{\Delta} \in \Pi$ with probability approaching one.*

As discussed previously, $\widehat{\mathbf{T}}_s$ links the more complex spline basis with a simple piecewise polynomial basis. Recall that $\widehat{\mathbf{T}}_s = \widehat{\mathbf{T}}_s(\widehat{\Delta})$ depends on the quantile-based partition $\widehat{\Delta}$. The next lemma describes its key features, and gives a precise definition of $\mathbf{T}_s := \mathbf{T}_s(\Delta)$, the transformation matrix corresponding to the nonrandom basis $\mathbf{b}_s(x)$, i.e., $\mathbf{b}_s(x) = \mathbf{T}_s \mathbf{b}_0(x)$.

Lemma SA-2.2 (Transformation Matrix). *Suppose that Assumption SA-1(i) holds. If $\frac{J \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then $\widehat{\mathbf{b}}_s(x) = \widehat{\mathbf{T}}_s \widehat{\mathbf{b}}_0(x)$ with $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$, $\|\widehat{\mathbf{T}}_s\| \lesssim_{\mathbb{P}} 1$, $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{J \log J}{n}}$, and $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \lesssim_{\mathbb{P}} \sqrt{\frac{J \log J}{n}}$.*

The next lemma characterizes the local basis $\widehat{\mathbf{b}}_s(x)$ and the associated Gram matrix.

Lemma SA-2.3 (Local Basis). *Suppose that Assumption SA-1(i) holds. Then*

$$\sup_{x \in \mathcal{X}} \|\widehat{\mathbf{b}}_s^{(v)}(x)\|_0 \leq (p+1)^2, \quad and \quad 1 \lesssim \lambda_{\min}(\mathbf{Q}) \leq \lambda_{\max}(\mathbf{Q}) \lesssim 1.$$

If, in addition, $\frac{J \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then

$$\sup_{x \in \mathcal{X}} \|\widehat{\mathbf{b}}_s^{(v)}(x)\| \lesssim_{\mathbb{P}} J^{\frac{1}{2}+v}, \quad \|\widehat{\mathbf{Q}} - \mathbf{Q}\| \lesssim_{\mathbb{P}} \sqrt{J \log J / n},$$

$$\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1, \quad and \quad \|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}\|_\infty \lesssim_{\mathbb{P}} \sqrt{J \log J / n}.$$

The next lemma shows that the limiting variance is bounded from above and below if properly scaled, which is key to pointwise and uniform inference. Recall that $\bar{\Omega}(x) = \bar{\Omega}(x; \widehat{\Delta})$ and $\Omega(x) = \Omega(x; \widehat{\Delta})$.

Lemma SA-2.4 (Asymptotic Variance). *Suppose that Assumption SA-1(i)-(ii) holds. If $\frac{J \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then*

$$J^{1+2v} \lesssim_{\mathbb{P}} \inf_{x \in \mathcal{X}} \bar{\Omega}(x) \leq \sup_{x \in \mathcal{X}} \bar{\Omega}(x) \lesssim_{\mathbb{P}} J^{1+2v}, \quad and$$

$$J^{1+2v} \lesssim \inf_{x \in \mathcal{X}} \Omega(x) \leq \sup_{x \in \mathcal{X}} \Omega(x) \lesssim J^{1+2v}.$$

As explained before, $\widehat{r}_\mu(x)$ is understood as L_2 approximation error of least squares estimators for $\mu(x)$. The next two lemmas establish bounds on $\widehat{r}_\mu(x)$ and its projection onto the space spanned by $\widehat{\mathbf{b}}_s(x)$ in terms of sup-norm.

Lemma SA-2.5 (L_2 Approximation Error). *Under Assumption SA-1(i), if $\frac{J \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then*

$$\sup_{x \in \mathcal{X}} |\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_\mu - \mu^{(v)}(x)| \lesssim_{\mathbb{P}} J^{-p-1+v}.$$

Lemma SA-2.6 (Projection of L_2 Approximation Error). *Under Assumption SA-1(i), if $\frac{J \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then*

$$\sup_{x \in \mathcal{X}} |\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)]| \lesssim_{\mathbb{P}} J^{-p-1+v}\sqrt{\frac{J \log J}{n}}.$$

The next lemma gives a bound on the variance component of the binscatter estimator, which is the main building block of uniform convergence.

Lemma SA-2.7 (Uniform Convergence: Variance). *Suppose that Assumption SA-1(i)(ii) hold. If $\frac{J^{\frac{\nu}{\nu-2}} \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then*

$$\sup_{x \in \mathcal{X}} |\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\mathbf{b}_s(x_i)\epsilon_i]| \lesssim J^v\sqrt{\frac{J \log J}{n}}.$$

Let $\{a_n : n \geq 1\}$ be a sequence of non-vanishing constants, which will be used later to characterize the strong approximation rate. The next theorem shows that under certain conditions the estimation of $\boldsymbol{\gamma}$ does not impact the asymptotic inference on the nonparametric component.

Lemma SA-2.8 (Covariate Adjustment). *Suppose that Assumption SA-1 holds. If $\frac{J \log J}{n} = o(1)$, $\frac{a_n}{\sqrt{J}} = o(1)$, $a_n\sqrt{n}J^{-p-(\varsigma \wedge (p+1))-\frac{3}{2}} = o(1)$ and , then*

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n}), \quad and \quad \|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i']\|_\infty \lesssim_{\mathbb{P}} J^v \ \text{for each } x \in \mathcal{X}.$$

If, in addition, $\frac{J^{\frac{\nu}{\nu-2}} \log J}{n} \lesssim 1$, then $\sup_{x \in \mathcal{X}} \|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i']\|_\infty \lesssim_{\mathbb{P}} J^v$.

Collecting the previous results, the next lemma constructs the rate of uniform convergence for binscatter estimators.

7

Lemma SA-2.9 (Uniform Convergence)**.** *Suppose that Assumption SA-1 holds. If $\sqrt{n}J^{-p-(\varsigma\wedge(p+1))-\frac{3}{2}} = o(1)$ and $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} \lesssim 1$, then*

$$\sup_{x\in\mathcal{X}} |\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)| \lesssim_{\mathbb{P}} J^v\sqrt{\frac{J\log J}{n}} + J^{-p-1+v}$$

The last lemma shows that the proposed variance estimator is consistent.

Lemma SA-2.10 (Variance Estimate)**.** *Suppose that Assumption SA-1 holds. If $\frac{J^{\frac{\nu}{\nu-2}}(\log J)^{\frac{\nu}{\nu-2}}}{n} = o(1)$ and $\sqrt{n}J^{-p-(\varsigma\wedge(p+1))-\frac{3}{2}} = o(1)$, then*

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\| \lesssim_{\mathbb{P}} J^{-p-1} + \sqrt{\frac{J\log J}{n^{1-\frac{2}{\nu}}}}.$$

As a result,

$$\sup_{x\in\mathcal{X}} |\widehat{\Omega}(x) - \Omega(x)| \lesssim_{\mathbb{P}} J^{1+2v}\left(J^{-p-1} + \sqrt{\frac{J\log J}{n^{1-\frac{2}{\nu}}}}\right).$$

SA-3   Main Results

SA-3.1   Integrated Mean Squared Error

The following theorem proves the result stated in Theorem 1 of the main paper.

Theorem SA-3.1 (IMSE)**.** *Suppose that Assumption SA-1 holds. Let $\omega(x)$ be a continuous weighting function over $\mathcal{X}$ bounded away from zero. If $\frac{J\log J}{n} = o(1)$ and $\sqrt{n}J^{-p-(\varsigma\wedge(p+1))-\frac{3}{2}} = o(1)$, then*

$$\int_{\mathcal{X}} \mathbb{E}\left[\left(\widehat{\mu}^{(v)}(x)-\mu^{(v)}(x)\right)^2\Big|\mathbf{D}\right]\omega(x) = \frac{J^{1+2v}}{n}\mathscr{V}_n(p,s,v)+J^{-2(p+1-v)}\mathscr{B}_n(p,s,v)+o_{\mathbb{P}}\left(\frac{J^{1+2v}}{n}+J^{-2(p+1-v)}\right).$$

where

$$\mathscr{V}_n(p,s,v) := J^{-(1+2v)}\,\mathrm{trace}\left(\mathbf{Q}^{-1}\boldsymbol{\Sigma}\mathbf{Q}^{-1}\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\right) \asymp 1,$$

$$\mathscr{B}_n(p,s,v) := J^{2p+2-2v}\int_{\mathcal{X}}\left(\mathbf{b}_s^{(v)}(x)'\boldsymbol{\beta}_\mu - \mu^{(v)}(x)\right)^2\omega(x)dx \lesssim 1.$$

As a consequence, the IMSE-optimal bin is

$$J_{\mathtt{IMSE}} = \left\lceil\left(\frac{2(p-v+1)\mathscr{B}_n(p,s,v)}{(1+2v)\mathscr{V}_n(p,s,v)}\right)^{\frac{1}{2p+3}}n^{\frac{1}{2p+3}}\right\rceil.$$

Regarding the bias component $\mathscr{B}_n(p, s, v)$, a more explicit but more cumbersome expression is available in the proof, which forms the foundation of our bin selection procedure discussed in Section SA-4. However, for $s = 0$, both variance and bias terms admit concise explicit formulas, as shown in the following corollary. To state the results, we introduce a polynomial function $\mathscr{B}_p(x)$ for $p \in \mathbb{Z}_+$ such that $\binom{2p}{p}\mathscr{B}_p(x)$ is the *shifted* Legendre polynomial of degree p on $[0, 1]$. These polynomials are orthogonal on $[0, 1]$ with respect to the Lebesgue measure. On the other hand, let $\boldsymbol{\psi}(z) = (1, z, \ldots, z^p)'$.

Corollary SA-3.1. *Under the assumptions in Theorem SA-3.1, $\mathscr{V}_n(p, 0, v) = \mathscr{V}(p, 0, v) + o(1)$ and $\mathscr{B}_n(p, 0, v) = \mathscr{B}(p, 0, v) + o(1)$ where*

$$\mathscr{V}(p, 0, v) := \mathrm{trace}\left\{\left(\int_0^1 \boldsymbol{\psi}(z)\boldsymbol{\psi}(z)'dz\right)^{-1}\int_0^1 \boldsymbol{\psi}^{(v)}(z)\boldsymbol{\psi}^{(v)}(z)'dz\right\}\int_{\mathcal{X}} \sigma^2(x)f(x)^{2v}\omega(x)dx,$$

$$\mathscr{B}(p, 0, v) := \frac{\int_0^1 [\mathscr{B}_{p+1-v}(z)]^2 dz}{((p+1-v)!)^2}\int_{\mathcal{X}} \frac{[\mu^{(p+1)}(x)]^2}{f(x)^{2p+2-2v}}\omega(x)dx.$$

Remark SA-3.1. The above corollary implies that the bias constant $\mathscr{B}(p, 0, v)$ is nonzero unless $\mu^{(p+1)}(x)$ is zero almost everywhere on $\mathcal{X}$. For other $s > 0$, notice that $\mathbf{b}_s^{(v)}(x)'\boldsymbol{\beta}_\mu$ can be viewed as an approximation of $\mu^{(v)}(x)$ in the space spanned by piecewise polynomials of order $(p - v)$. The best $L_2(x)$ approximation error in this space, according to the above corollary, is bounded away from zero if rescaled by J^{p+1-v}. $\mathbf{b}_s^{(v)}(x)'\boldsymbol{\beta}_\mu$, as a non-optimal L_2 approximation in such a space, must have a larger L_2 error than the best one (in terms of L_2-norm). Since $\omega(x)$ and $f(x)$ are both bounded and bounded away from zero, the above fact implies that except for the quite special case mentioned previously, $\mathscr{B}(p, s, v) \asymp 1$, a slightly stronger result than that in Theorem SA-3.1. In all analysis that follows, we simply exclude this special case when the leading bias degenerates, and thus $J_{\mathtt{IMSE}} \asymp n^{\frac{1}{2p+3}}$. ⌋

SA-3.2 Pointwise Inference

We consider statistical inference based on the Studentized t-statistic:

$$\widehat{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}.$$

Let $\Phi(\cdot)$ be the cumulative distribution function of a standard normal random variable. The following theorem proves Lemma 1 of the main paper.

Theorem SA-3.2. *Suppose that Assumption SA-1 holds. If $\sup_{x \in \mathcal{X}} \mathbb{E}[|\epsilon_i|^\nu | x_i = x] \lesssim 1$ for some $\nu \geq 3$, $\frac{J^{\frac{\nu}{\nu-2}} (\log J)^{\frac{\nu}{\nu-2}}}{n} = o(1)$ and $nJ^{-2p-3} = o(1)$, then*

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}(\widehat{T}_p(x) \leq u) - \Phi(u) \right| = o(1), \quad \text{for each } x \in \mathcal{X}.$$

Let $\widehat{I}_p(x) = [\widehat{\mu}^{(v)}(x) \pm \mathfrak{c}\sqrt{\widehat{\Omega}(x)/n}]$ for some critical value $\mathfrak{c}$ to be specified. Given the above theorem, we have the following corollary, a result stated in Theorem 2 fo the main paper is valid.

Corollary SA-3.2. *For given p, suppose that the conditions in Theorem SA-3.2 hold, and further assume that $\mu(x)$ and $\mathbb{E}[\mathbf{w}_i | x_i = x]$ are $(p+q+1)$-times continuously differentiable for some $q \geq 1$. If $J = J_{\texttt{IMSE}}$ and $\mathfrak{c} = \Phi^{-1}(1 - \alpha/2)$, then*

$$\mathbb{P}\left[\mu^{(v)}(x) \in \widehat{I}_{p+q}(x) \right] = 1 - \alpha + o(1), \quad \text{for all } x \in \mathcal{X},$$

SA-3.3   Uniform Inference

Recall that $\{a_n : n \geq 1\}$ is a sequence of non-vanishing constants. We will first show that the (feasible) Studentized t-statistic process $\{\widehat{T}_p(x) : x \in \mathcal{X}\}$ can be approximated by a Gaussian process in a proper sense at certain rate.

Theorem SA-3.3 (Strong Approximation). *Suppose that Assumption SA-1 holds. If*

$$\frac{J(\log J)^2}{n^{1-\frac{2}{\nu}}} = o(a_n^{-2}), \quad J^{-1} = o(a_n^{-2}) \quad \text{and} \quad nJ^{-2p-3} = o(a_n^{-2}),$$

then, on a properly enriched probability space, there exists some K_s-dimensional standard normal random vector $\mathbf{N}_{K_s}$ such that for any $\eta > 0$,

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |\widehat{T}_p(x) - Z_p(x)| > \eta a_n^{-1} \right) = o(1), \quad Z_p(x) = \frac{\widehat{\mathbf{b}}_0(x)' \mathbf{T}_s' \mathbf{Q}^{-1} \mathbf{\Sigma}^{1/2}}{\sqrt{\Omega(x)}} \mathbf{N}_{K_s}.$$

The approximating process $\{Z_p(x) : x \in \mathcal{X}\}$ is a Gaussian process conditional on $\mathbf{X}$ by construction. In practice, one can replace all unknowns in $Z_p(x)$ by their sample analogues, and then

construct the following feasible (conditional) Gaussian process:

$$\widehat{Z}_p(x) = \frac{\widehat{\mathbf{b}}_s(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}(x)}}\mathbf{N}_{K_s}.$$

where $\mathbf{N}_{K_s}$ denotes a K_s-dimensional standard normal vector independent of the data $\mathbf{D} = \{(y_i, x_i, \mathbf{w}_i') : 1 \le i \le n\}$.

Theorem SA-3.4 (Plug-in Approximation). *Suppose that the conditions in Theorem SA-3.3 holds. Then, on a properly enrich probability space there exists K_s-dimensional standard normal vector $\mathbf{N}_{K_s}$ independent of $\mathbf{D}$ such that for any $\eta > 0$,*

$$\mathbb{P}\Big[\sup_{x \in \mathcal{X}} |\widehat{Z}_p(x) - Z_p(x)| > \eta a_n^{-1}\Big|\mathbf{D}\Big] = o_{\mathbb{P}}(1).$$

SA-3.4   Applications

Theorem SA-3.3 and SA-3.4 offer a way to approximate the distribution of the *whole t*-statistic process. A direct application of this result is to constructing uniform confidence band, which relies on distributional approximation to the supremum of the *t*-statistic process. The following theorem proves Lemma 2 of the main paper.

Theorem SA-3.5 (Supremum Approximation). *Suppose that the conditions of Theorem SA-3.3 hold with $a_n = \sqrt{\log J}$. Then*

$$\sup_{u \in \mathbb{R}} \Big|\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{T}_p(x)| \le u\Big) - \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_p(x)| \le u\Big|\mathbf{D}\Big)\Big| = o_{\mathbb{P}}(1).$$

Using the above theorem, we have the following corollary, which is a result stated in Theorem 3 of the main paper.

Corollary SA-3.3. *For given p, suppose the conditions in Theorem SA-3.5 hold and $J = J_{\texttt{IMSE}}$. Further, assume that $\mu(x)$ and $\mathbb{E}[\mathbf{w}_i|x_i = x]$ are $(p + q + 1)$-times continuously differentiable for some $q \ge 1$. If $\mathfrak{c} = \inf\Big\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x \in \mathcal{X}} |\widehat{Z}_{p+q}(x)| \le c\,|\mathbf{D}] \ge 1 - \alpha\Big\}$. Then*

$$\mathbb{P}\Big[\mu^{(v)}(x) \in \widehat{I}_{p+q}(x),\ for\ all\ x \in \mathcal{X}\Big] = 1 - \alpha + o(1).$$

11

As another application, the main paper discusses two classes of hypothesis testing problems: testing parametric specifications and certain shape restrictions. To be specific, consider the following two problems:

(i) $\ddot{\mathsf{H}}_0 : \sup_{x \in \mathcal{X}} |\mu^{(v)}(x) - m^{(v)}(x, \boldsymbol{\theta})| = 0$ for some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ v.s.
   $\ddot{\mathsf{H}}_A : \sup_{x \in \mathcal{X}} |\mu^{(v)}(x) - m^{(v)}(x, \boldsymbol{\theta})| > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(ii) $\dot{\mathsf{H}}_0 : \sup_{x \in \mathcal{X}} (\mu^{(v)}(x) - m^{(v)}(x, \bar{\boldsymbol{\theta}})) \leq 0$ for a certain $\bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ v.s.
   $\dot{\mathsf{H}}_A : \sup_{x \in \mathcal{X}} (\mu^{(v)}(x) - m^{(v)}(x, \bar{\boldsymbol{\theta}})) > 0$ for $\bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$.

The testing problem in (i) can be viewed as a two-sided test where the equality between two functions holds *uniformly* over $x \in \mathcal{X}$. In this case, we introduce $\widehat{\boldsymbol{\theta}}$ as a consistent estimator of $\boldsymbol{\theta}$ under $\ddot{\mathsf{H}}_0$. Then we rely on the following test statistic:

$$\ddot{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - m^{(v)}(x, \widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}.$$

The null hypothesis is rejected if $\sup_{x \in \mathcal{X}} |\ddot{T}_p(x)| > \mathfrak{c}$ for some critical value $\mathfrak{c}$.

The testing problem in (ii) can be viewed as a one-sided test where the inequality holds *uniformly* over $x \in \mathcal{X}$. Importantly, it should be noted that under both $\dot{\mathsf{H}}_0$ and $\dot{\mathsf{H}}_A$, we fix $\bar{\boldsymbol{\theta}}$ to be the same value in $\boldsymbol{\Theta}$. In such a case, we introduce $\widehat{\bar{\boldsymbol{\theta}}}$ as a consistent estimator of $\bar{\boldsymbol{\theta}}$ under both $\dot{\mathsf{H}}_0$ and $\dot{\mathsf{H}}_A$. Then we will rely on the following test statistic

$$\dot{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - m^{(v)}(x, \widehat{\bar{\boldsymbol{\theta}}})}{\sqrt{\widehat{\Omega}(x)/n}}.$$

The null hypothesis is rejected if $\sup_{x \in \mathcal{X}} \dot{T}_p(x) > \mathfrak{c}$ for some critical value $\mathfrak{c}$.

The following theorem characterizes the size and power of such tests.

Theorem SA-3.6 (Hypothesis Testing)**.** *Let the conditions in Theorem SA-3.3 holds with* $a_n = \sqrt{\log J}$.

(i) *(Specification) Let* $\mathfrak{c} = \inf \left\{ c \in \mathbb{R}_+ : \mathbb{P}\left[\sup_{x \in \mathcal{X}} |\widehat{Z}_p(x)| \leq c \Big| \mathbf{D} \right] \geq 1 - \alpha \right\}$.

12

Under $\ddot{\mathsf{H}}_0$, if $\sup_{x \in \mathcal{X}} |\mu^{(v)}(x) - m^{(v)}(x; \widehat{\boldsymbol{\theta}})| = o_{\mathbb{P}}\left(\sqrt{\frac{J^{1+2v}}{n \log J}}\right)$, then

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} |\ddot{T}_p(x)| > \mathfrak{c}\right] = \alpha.$$

Under $\ddot{H}_A$, if there exists some $\bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ such that $\sup_{x \in \mathcal{X}} |m^{(v)}(x, \widehat{\boldsymbol{\theta}}) - m^{(v)}(x, \bar{\boldsymbol{\theta}})| = o_{\mathbb{P}}(1)$, and $J^v \sqrt{\frac{J \log J}{n}} = o(1)$, then

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} |\ddot{T}_p(x)| > \mathfrak{c}\right] = 1.$$

(ii) (Shape Restriction) Let $\mathfrak{c} = \inf\left\{c \in \mathbb{R}_+ : \mathbb{P}\left[\sup_{x \in \mathcal{X}} \widehat{Z}_p(x) \le c \big| \mathbf{D}\right] \ge 1 - \alpha\right\}$. Assume that $\sup_{x \in \mathcal{X}} |m^{(v)}(x; \widehat{\overline{\boldsymbol{\theta}}}) - m^{(v)}(x, \bar{\boldsymbol{\theta}})| = o_{\mathbb{P}}\left(\sqrt{\frac{J^{1+2v}}{n \log J}}\right)$.

Under $\dot{H}_0$,

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} \dot{T}_p(x) > \mathfrak{c}\right] \le \alpha.$$

Under $\dot{H}_A$, if $J^v \sqrt{\frac{J \log J}{n}} = o(1)$,

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} \dot{T}_p(x) > \mathfrak{c}\right] = 1.$$

The robust bias-corrected testing procedures given in Theorem 4 and 5 of the main paper are immediate corollaries of Theorem SA-3.6, once the stronger conditions on the smoothness of $\mu(x)$ and $\mathbb{E}[\mathbf{w}_i | x_i = x]$ are assumed. To conserve some space, we do not repeat their statements.

SA-4   Implementation Details

We discuss the implementation details for data-driven selection of the number of bins, based on the integrated mean squared error expansion (see Theorem SA-3.1 and Corollary SA-3.1) presented above. We offer two procedures for estimating the bias and variance constants, and once these estimates ($\widehat{\mathscr{B}}_n(p, s, v)$ and $\widehat{\mathscr{V}}_n(p, s, v)$) are available, the estimated optimal J is

$$\widehat{J}_{\mathtt{IMSE}} = \left\lceil \left(\frac{2(p - v + 1)\widehat{\mathscr{B}}_n(p, s, v)}{(1 + 2v)\widehat{\mathscr{V}}_n(p, s, v)}\right)^{\frac{1}{2p+3}} n^{\frac{1}{2p+3}} \right\rceil.$$

We always let $\omega(x) = f(x)$ as weighting function for concreteness.

SA-4.1 Rule-of-thumb Selector

A rule-of-thumb choice of J is obtained based on Corollary SA-3.1, in which case $s = 0$.

Regarding the variance constants $\mathcal{V}(p, 0, v)$, the unknowns are the density function $f(x)$ and the conditional variance $\sigma^2(x)$. A Gaussian reference model is employed for $f(x)$. For the conditional variance, we note that $\sigma^2(x) = \mathbb{E}[y_i^2 | x_i, \mathbf{w}_i] - (\mathbb{E}[y_i | x_i, \mathbf{w}_i])^2$. The two conditional expectations can be approximated by global polynomial regressions of degree $p + 1$. Then, the variance constant is estimated by

$$\widehat{\mathcal{V}}_{p,0,v} = \text{trace} \left\{ \left(\int_0^1 \boldsymbol{\psi}(z)\boldsymbol{\psi}(z)'dz \right)^{-1} \int_0^1 \boldsymbol{\psi}^{(v)}(z)\boldsymbol{\psi}^{(v)}(z)'dz \right\} \times \frac{1}{n} \sum_{i=1}^n \widehat{\sigma}^2(x_i)\widehat{f}(x_i)^{2v}$$

Regarding the bias constant, the unknowns are $f(x)$, which is estimated using the Gaussian reference model, and $\mu^{(p+1)}(x)$, which can be estimated based on the global regression that approximates $\mathbb{E}[y_i | x_i, \mathbf{w}_i]$. Then the bias constant is estimated by

$$\widehat{\mathcal{B}}(p, 0, v) = \frac{\int_0^1 [\mathcal{B}_{p+1-v}(z)]^2 dz}{((p+1-v)!)^2} \times \frac{1}{n} \sum_{i=1}^n \frac{[\widehat{\mu}^{(p+1)}(x_i)]^2}{\widehat{f}(x_i)^{2p+2-2v}}.$$

The resulting J selector employs the correct rate but an inconsistent constant approximation. Recall that s does not change the rate of J_{IMSE}. Thus, even for other $s > 0$, this selector still gives a correct rate.

SA-4.2 Direct-plug-in Selector

The direct-plug-in selector is implemented based on the binscatter estimators, which apply to any user-specified p, s and v. It requires a preliminary choice of J, for which the rule-of-thumb selector previously described can be used.

More generally, suppose that a preliminary choice J_{pre} is given, and then a binscatter basis $\mathbf{b}_s(x)$ (of order p) can be constructed immediately on the preliminary partition. Implementing a binscatter regression using this basis and partitioning, the variance constant then can be estimated using a standard variance estimator, such as the one in Lemma SA-2.10.

Regarding the bias constant, we employ the uniform approximation (SA-6.6) in the proof of Theorem SA-3.1. The key idea of the bias representation is to "orthogonalize" the leading error

of the uniform approximation based on splines with simple knots (i.e., p smoothness constraints are imposed) with respect to the preliminary binscatter basis $\mathbf{b}_s(x)$. Specifically, the key unknown in the expression of the leading error is $\mu^{(p+1)}(x)$, which can be estimated by implementing a binscatter regression of order $p+1$ (with the preliminary partition unchanged). Plug it in (SA-6.7), and all other quantities in it can be replaced by their sample analogues. Then a bias constant estimate is available.

By this construction, the direct-plug-in selector employs the correct rate and a consistent constant approximation for any p, s and v.

SA-5 Empirical Illustrations

In this section we provide details on the simulation model used in Section 2 of the main paper, and demonstrate the procedures introduced in the main paper with three examples based on publicly available data. All data are obtained from the American Community Survey (ACS) using the 5-year survey estimates. We use the 5-year survey which began in 2013 and finished in 2017. All analyses are performed at the zip code tabulation area level for the United States (excluding Puerto Rico), with the data downloaded from the Census Bureau website: https://factfinder.census.gov/faces/nav/jsf/pages/programs.xhtml?program=acs

Replication files are available at: https://sites.google.com/site/nppackages/replication/

SA-5.1 Simulated Data

In Section 2 of the main paper, we use a simulated dataset to illustrate our main methods. The data generating process is constructed using the real survey dataset on the Gini index and household income described in the next subsection. Specifically, we set the sample size $n = 1,000$, and the independent variable of interest $x_i \sim \mathtt{beta}(2,4)$ where $\mathtt{beta}(2,4)$ is *beta* distribution with parameters 2 and 4. The regression function of interest is

$$\mu(x) = 24x^4 - 98.8x^3 + 112.4x^2 - 44.4x + 3.6.$$

The basic model for the outcome variable y is constructed as

$$y_i = \mu(x_i) + w_i + \epsilon_i, \quad \epsilon_i \sim \mathtt{N}(0, 0.5^2), \quad w_i \sim \mathtt{U}(-1, 1)$$

where $\mathtt{N}(0, 0.5^2)$ is the normal distribution with mean 0 and variance 0.5^2 and $\mathtt{U}(-1, 1)$ is the uniform distribution over $[-1, 1]$. x_i, w_i and ϵ_i are independent of each other.

The basic model is modified in several scenarios. In the disucssion of data variability and heteroskedasticity, we change the conditional variance of ϵ_i. In the discussion of covariate adjustment, we allow the dependence between x_i and w_i. In that case, $w_i = 3(x - 0.5) + \mathtt{U}(-0.5, 0.5)$ where the errors are independent of x_i.

SA-5.2 Example 1: Gini Index versus Household Income

The results are provided in Figure SA-1. The dependent variable is the Gini index and the variable of interest is median household income. The control variables are (i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's degree, (iii) median age of residents, (iv) percentage of residents without health insurance, and (v) the local unemployment rate.

This dataset was used to construct the data generating process used in Section 2 of the main paper, and as an impirical illustration in Section 7 of the main paper.

SA-5.3 Example 2: Internet Access versus Household Income

The results are provided in Figure SA-2. The dependent variable is the percentage of households with internet access and the variable of interest is median household income. The control variables are (i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's degree, (iii) median age of residents, (iv) percentage of residents without health insurance, and (v) the local unemployment rate.
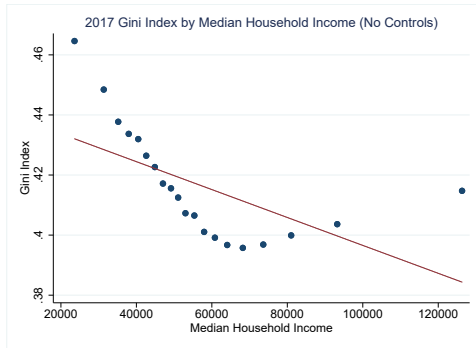
SA-5.4 Example 3: Uninsured Rate versus per Capital Income

The results are provided in Figure SA-3. The dependent variable is the percentage of individuals without health insurance and the variable of interest is per capita income. The control variables are
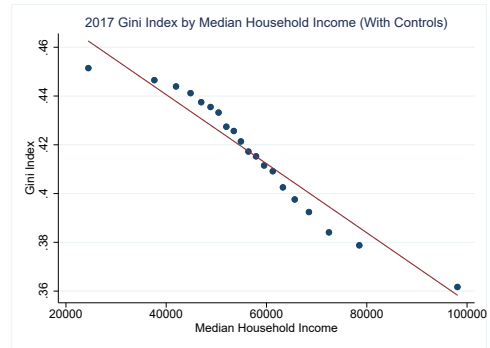
(i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's degree, (iii) median age of residents, and (iv) the local unemployment rate.

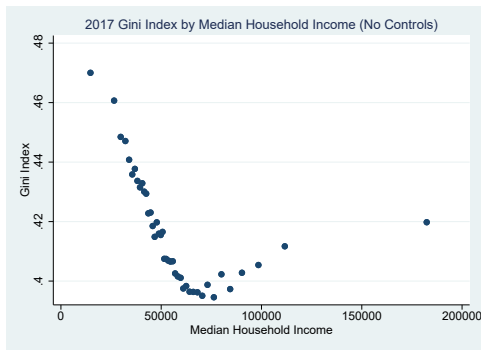# Figure SA-1: **Gini Index versus Household Income**

No Controls
`binscatter`

2017 Gini Index by Median Household Income (No Controls)

With Controls
`binscatter`

2017 Gini Index by Median Household Income (With Controls)

`binsreg`

2017 Gini Index by Median Household Income (No Controls)

`binsreg`

2017 Gini Index by Median Household Income (With Controls)

2017 Gini Index by Median Household Income (No Controls)

2017 Gini Index by Median Household Income (With Controls)

2017 Gini Index by Median Household Income (No Controls)
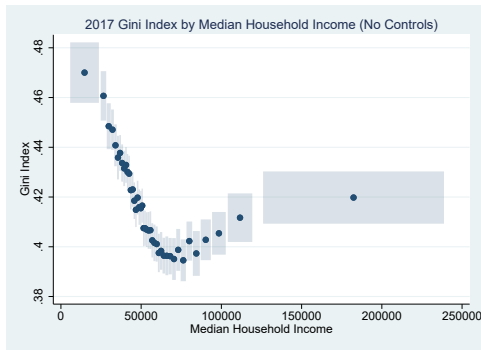

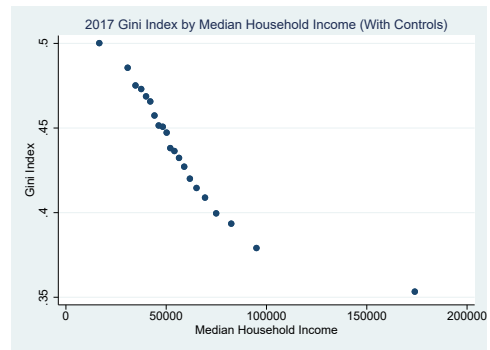2017 Gini Index by Median Household Income (With Controls)

Figure SA-2: **Internet Access versus Household Income**

No Controls
`binscatter`

2017 Percent With Internet Access by Median Household Income (No Controls)

With Controls
`binscatter`

2017 Percent With Internet Access by Median Household Income (With Controls)

`binsreg`

2017 Percent With Internet Access by Median Household Income (No Controls)

`binsreg`

2017 Percent With Internet Access by Median Household Income (With Controls)

2017 Percent With Internet Access by Median Household Income (No Controls)

2017 Percent With Internet Access by Median Household Income (With Controls)

2017 Percent With Internet Access by Median Household Income (No Controls)

2017 Percent With Internet Access by Median Household Income (With Controls)

19

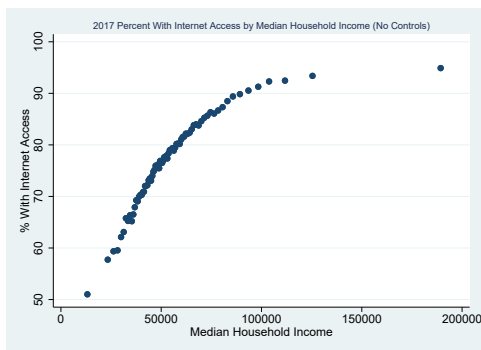Figure SA-3: **Uninsured Rate versus per Capita Income**

No Controls
binscatter

With Controls
binscatter

binsreg

binsreg

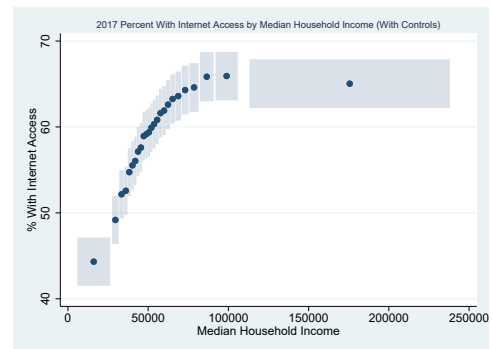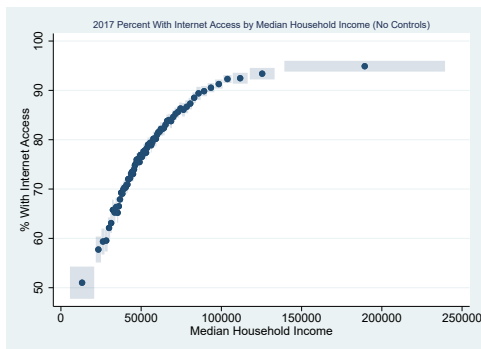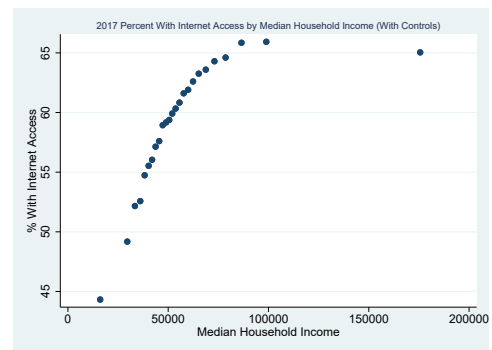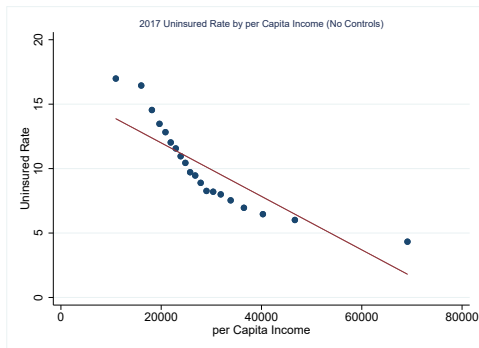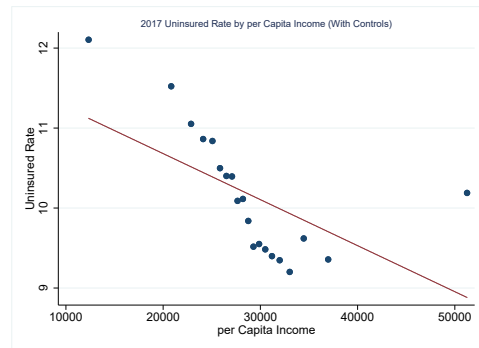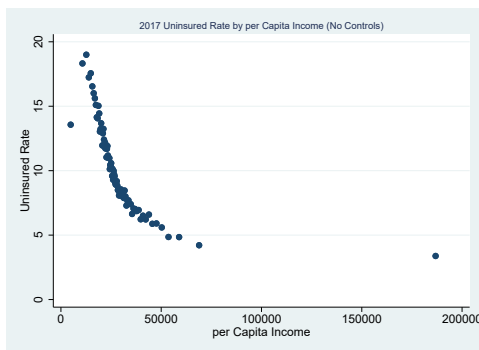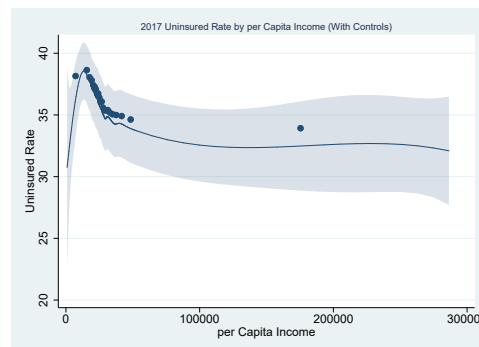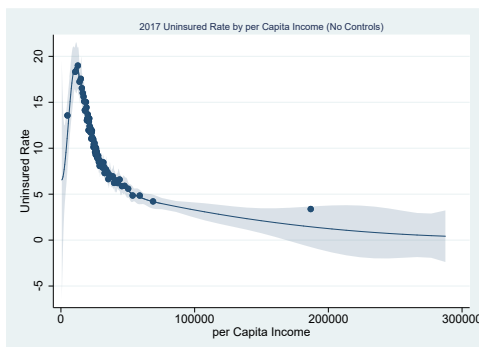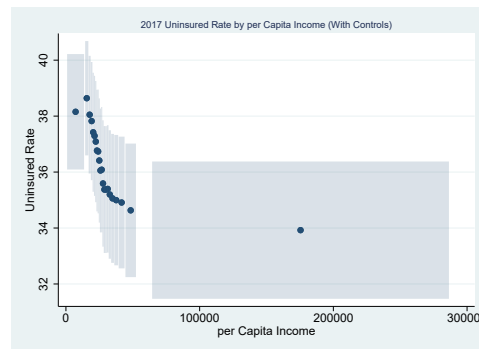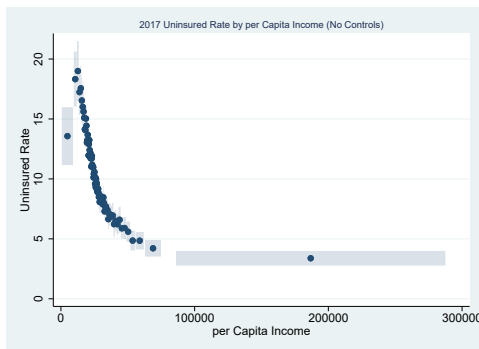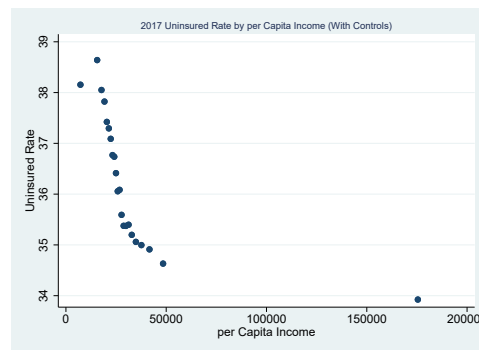## SA-6   Proof

SA-6.1   Proof of Lemma SA-2.1

Proof. The first result follows by Lemma SA2 of Calonico, Cattaneo, and Titiunik (2015). To show the second result, first consider the deterministic partition sequence Δ based on the population quantiles. By mean value theorem,

$$h_j = F^{-1}\Big(\frac{j}{J}\Big) - F^{-1}\Big(\frac{j-1}{J}\Big) = \frac{1}{f(F^{-1}(\xi))} \cdot \frac{1}{J}$$

where ξ is some point between $(j-1)/J$ and j/J. Since f is bounded and bounded away from zero, $\max_{1\leq j\leq J} h_j / \min_{1\leq j\leq J} h_j \leq \bar{f}/\underline{f}$. Using the first result, we have with probability approaching one,

$$\max_{1\leq j\leq J} |\hat{h}_j - h_j| \leq J^{-1}\bar{f}^{-1}/2.$$

Then,

$$\frac{\max_{1\leq j\leq J} \hat{h}_j}{\min_{1\leq j\leq J} \hat{h}_j} = \frac{\max_{1\leq j\leq J} h_j + \max_{1\leq j\leq J} |\hat{h}_j - h_j|}{\min_{1\leq j\leq J} h_j - \max_{1\leq j\leq J} |\hat{h}_j - h_j|} \leq \frac{3\bar{f}}{\underline{f}},$$

and the desired result follows. □

SA-6.2   Proof of Lemma SA-2.2

Proof. For $s = 0$, the result is trivial. For $0 < s \leq p$, $\mathbf{b}_s(x)$ is formally known as B-spline basis of order $p+1$ with knots $\{\hat{\tau}_1, \ldots, \hat{\tau}_{J-1}\}$ of multiplicities $(p-s+1, \ldots, p-s+1)$. See Schumaker (2007, Definition 4.1). Specifically, such a basis is constructed on an extended knot sequence $\{\xi_j\}_{j=1}^{2(p+1)+(p-s+1)(J-1)}$:

$$\xi_1 \leq \cdots \leq \xi_{p+1} \leq 0, \quad 1 \leq \xi_{p+2+(p-s+1)(J-1)} \leq \cdots \leq \xi_{2(p+1)+(p-s+1)(J-1)}.$$

and

$$\xi_{p+2} \leq \cdots \leq \xi_{p+1+(p-s+1)(J-1)} = \underbrace{\hat{\tau}_1, \cdots, \hat{\tau}_1}_{p-s+1}, \cdots, \underbrace{\hat{\tau}_{J-1}, \cdots, \hat{\tau}_{J-1}}_{p-s+1}.$$

By the well-known Recursive Relation of splines, a typical function $\widehat{b}_{s,\ell}(x)$ in $\widehat{\mathbf{b}}_s(x)$ supported on

21

$(\xi_\ell, \xi_{\ell+p+1})$ is expressed as

$$\widehat{b}_{s,\ell}(x) = \sqrt{J} \sum_{j=\ell+1}^{\ell+p+1} C_j(x) \mathbb{1}(x \in [\xi_{j-1}, \xi_j)).$$

where each $C_j(x)$ is a polynomial of degree p as the sum of products of p linear polynomials. See De Boor (1978, Section IX, Equation (19)). Since $s \leq p$, we always have $\xi_\ell < \xi_{\ell+p+1}$. Thus, the support of such a basis function is well defined.

Specifically, all $C_j(x)$'s take the following form:

$$C_j(x) = \sum_{\iota=1}^{M} \prod_{(k,k') \in \mathcal{K}_\iota} \frac{(-1)^{c_{k,k'}}(x - \xi_k)}{\xi_k - \xi_{k'}}.$$

Here, the convention is that "$0/0 = 0$", $M \leq 2^p$ is a constant denoting the number of summands, the cardinality of the index pair set $\mathcal{K}_s$ is exactly p, and $c_{k,k'}$ is a constant used to change the sign of the summand. These indices may depend on j, which is omitted for notation simplicity. As explained previously, such a function is supported on at least one bin.

We want to linearly represent such a function in terms of $\mathbf{b}_0(x)$ with typical element

$$\varphi_{j,\alpha}(x) = \sqrt{J} \cdot \mathbb{1}_{\widehat{\mathcal{B}}_j}(x)\left(\frac{x - \hat{\tau}_{j-1}}{\hat{h}_j}\right)^\alpha, \quad 1 \leq \alpha \leq p, \quad 1 \leq j \leq J. \tag{SA-6.1}$$

Suppose without loss of generality, $\xi_{j-1} < \xi_j$ and (ξ_{j-1}, ξ_j) is a cell within the support of $\widehat{b}_{s,\ell}(x)$. Let $c_{j,\alpha}$ be the coefficient of $\varphi_{j,\alpha}(x)$ in the linear representation of $\widehat{\mathbf{b}}_s(x)$. Using the above results, it takes the following form

$$c_{j,\alpha} = \sum_{\iota=1}^{M} \frac{(\xi_j - \xi_{j-1})^\alpha \sum_{l_\iota=1}^{C_{p,\alpha}} \prod_{k=k_{l_\iota,1}}^{k_{l_\iota,p-\alpha}}(\xi_{j-1} - \xi_k)}{\prod_{(k,k') \in \mathcal{K}_\iota}(-1)^{c_{k,k'}}(\xi_k - \xi_{k'})}.$$

The quantities within the summation only depend on distance between knots, which is no greater than $(p+1)\max_j \hat{h}_j$, since the support covers at most $(p+1)$ bins. Both denominator and numerator are products of p such distances, and hence by Lemma SA-2.1, $\sup_{j,\alpha} |c_{j,\alpha}| \lesssim_{\mathbb{P}} 1$.

Since each row and column of $\widehat{\mathbf{T}}_s$ only contain a finite number of nonzeros, $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\| \lesssim_{\mathbb{P}} 1$. Using the fact $\max_{1 \leq j \leq J} |\hat{h}_j - h_j| \lesssim_{\mathbb{P}} J^{-1}\sqrt{J \log J/n}$, given in the proof of Lemma

22

SA-2.1, and noticing the form $c_{j,\alpha}$, $\max_{k,l} |(\widehat{\mathbf{T}}_s - \mathbf{T}_s)_{k,l}| \lesssim \sqrt{J \log J/n}$ where $(\widehat{\mathbf{T}}_s - \mathbf{T}_s)_{k,l}$ is (k,l)th element of $\widehat{\mathbf{T}}_s - \mathbf{T}_s$. Since $(\widehat{\mathbf{T}}_s - \mathbf{T}_s)$ only has a finite number of nonzeros on every row and column, $\|\widehat{\mathbf{T}} - \mathbf{T}\|_\infty \lesssim_{\mathbb{P}} \sqrt{J \log J/n}$ and $\|\widehat{\mathbf{T}} - \mathbf{T}\| \lesssim_{\mathbb{P}} \sqrt{J \log J/n}$.

Finally, we give an explicit expression of $c_{j,\alpha}$ for the case $s = p$, which may be of independent interest. In this case, $\mathbf{b}_s(x)$ is the usual B-spline basis with simple knots. Let $\widehat{b}_{s,\ell}(x)$ be a typical basis function supported on $[\hat{\tau}_\ell, \hat{\tau}_{\ell+p+1}]$. Then, using recursive formula of B-splines, by induction we have

$$\widehat{b}_{s,\ell}(x) = (\hat{\tau}_{\ell+p+1} - \hat{\tau}_\ell) \sum_{j=\ell}^{\ell+p+1} \frac{(x - \hat{\tau}_j)_+^p}{\prod_{\substack{k=\ell \\ k \neq l}}^{\ell+p+1} (\hat{\tau}_k - \hat{\tau}_j)}, \tag{SA-6.2}$$

where $(z)_+$ equal to z if $z \geq 0$ and 0 otherwise. Since $\widehat{b}_{s,\ell}(x)$ is zero outside of $(\hat{\tau}_\ell, \hat{\tau}_{\ell+p+1})$, $\widehat{b}_{s,\ell}(x)$ can be written as a linear combination of $\varphi_{j,\alpha}(x)$, $j = \ell + 1, \ldots, \ell + p + 1, \alpha = 0, \ldots, m-1$:

$$\widehat{b}_{s,\ell}(x) = \sum_{\alpha=0}^{p} \sum_{j=\ell+1}^{\ell+p+1} c_{j,\alpha} \varphi_{j,\alpha}(x), \quad \text{for some} \quad c_{j,\alpha}, \tag{SA-6.3}$$

For a generic cell $(\hat{\tau}_{j-1}, \hat{\tau}_j) \subset (\hat{\tau}_\ell, \hat{\tau}_{\ell+p+1})$, all truncated polynomials $(x - \hat{\tau}_k)_+^p$ does not contribute to the coefficients of $\varphi_{j,\alpha}(x)$ if $k > j - 1$. For any $\ell \leq k \leq j - 1$, we can expand $(x - \hat{\tau}_k)_+^p$ on $(\hat{\tau}_{j-1}, \hat{\tau}_j)$ as

$$(x - \hat{\tau}_k)^p = (x - \hat{\tau}_{j-1} + \hat{\tau}_{j-1} - \hat{\tau}_k)^p = \sum_{\alpha=0}^{p} \binom{p}{\alpha} (x - \hat{\tau}_{j-1})^\alpha (\hat{\tau}_{j-1} - \hat{\tau}_k)^{p-\alpha}.$$

Thus, the contribution of $(x - \hat{\tau}_k)_+^p$ to the coefficients of $\varphi_{j,\alpha}(x)$ in Equation (SA-6.3), combined with its coefficient in Equation (SA-6.2), is

$$\binom{p}{\alpha} (\hat{\tau}_{j-1} - \hat{\tau}_k)^{p-\alpha} (\hat{\tau}_j - \hat{\tau}_{j-1})^\alpha (\hat{\tau}_{\ell+p+1} - \hat{\tau}_\ell) \Big(\prod_{\substack{k'=\ell \\ k' \neq k}}^{\ell+p+1} (\hat{\tau}_{k'} - \hat{\tau}_k) \Big)^{-1}.$$

Collecting all such coefficients contributed by $(x - \hat{\tau}_k)_+^p$, $k = \ell, \ldots, j$, we obtain

$$c_{j,\alpha} = \sum_{k=\ell}^{j-1} \binom{p}{\alpha} (\hat{\tau}_{j-1} - \hat{\tau}_k)^{p-\alpha} (\hat{\tau}_j - \hat{\tau}_{j-1})^\alpha (\hat{\tau}_{\ell+p+1} - \hat{\tau}_\ell) \Big(\prod_{\substack{k'=\ell \\ k' \neq k}}^{\ell+p+1} (\hat{\tau}_{k'} - \hat{\tau}_k) \Big)^{-1}.$$

$\square$

SA-6.3 Proof of Lemma SA-2.3

Proof. The sparsity of the basis follows by construction. The upper bound on the maximum eigenvalue of $\mathbf{Q}$ follows from Lemma SA-2.2, and the quasi-uniformity property of population quantiles shown in the proof of Lemma SA-2.1. Also, in view of Lemma SA-2.1, the lower bound on the minimum eigenvalue of $\mathbf{Q}$ follows from Theorem 4.41 of Schumaker (2007), by which the minimum eigenvalue of $\mathbf{Q}/J$ (the scaling factor dropped) is bounded by $\min_{1 \le j \le J} h_j$ up to some universal constant.

To show the bound on $\|\widehat{\mathbf{b}}_s^{(v)}(x)\|$, notice that when $s = 0$, for any $x \in \mathcal{X}$ and any $j = 1, \dots, J(p+1)$, $0 \le \widehat{b}_{0,j}(x) \le \sqrt{J}$. Define $\varphi_{j,\alpha}(x)$ as in Equation (SA-6.1). Since

$$\varphi_{j,\alpha}^{(v)} = \sqrt{J}\alpha(\alpha-1)\cdots(\alpha-v+1)\hat{h}_j^{-v}\mathbb{1}_{\widehat{\mathcal{B}}_j}(x)\Big(\frac{x-\hat{\tau}_{j-1}}{\hat{h}_j}\Big)^{\alpha-v} \lesssim \sqrt{J}\hat{h}_j^{-v},$$

the bound on $\|\widehat{\mathbf{b}}_s^{(v)}(x)\|$ simply follows from Lemma SA-2.1 and Lemma SA-2.2.

Now, we prove the convergence of $\widehat{\mathbf{Q}}$. In view of Lemma SA-2.2, it suffices to show the convergence of $\widehat{\mathbf{Q}}$ when $s = 0$, i.e., $\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'] - \mathbb{E}[\mathbf{b}_0(x_i)\mathbf{b}_0(x_i)']\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$. By Lemma SA-2.1, with probability approach 1, $\widehat{\Delta}$ ranges within a family of partitions Π. Let $\mathcal{A}_n$ denote the event on which $\widehat{\Delta} \in \Pi$. Thus, $\mathbb{P}(\mathcal{A}_n^c) = o(1)$. On $\mathcal{A}_n$,

$$\left\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'] - \mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)']\right\| \le \sup_{\bar{\Delta}\in\Pi}\left\|\mathbb{E}_n[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})'] - \mathbb{E}[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})']\right\|.$$

By the relation between matrix norms, the right-hand-side of the above inequality is further bounded by

$$\sup_{\bar{\Delta}\in\Pi}\left\|\mathbb{E}_n[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})'] - \mathbb{E}[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})']\right\|_\infty.$$

Let a_{kl} be a generic (k,l)th entry of the matrix inside the matrix norm, i.e.,

$$|a_{kl}| = \left|\mathbb{E}_n[b_{0,k}(x_i;\bar{\Delta})b_{0,l}(x_i;\bar{\Delta})'] - \mathbb{E}\Big[b_{0,k}(x_i;\bar{\Delta})b_{0,l}(x_i;\bar{\Delta})'\Big]\right|$$

Clearly, if $b_{0,k}(\cdot;\bar{\Delta})$ and $b_{0,l}(\cdot;\bar{\Delta})$ are basis functions with different supports, a_{kl} is zero. Now

24

define the following function class

$$\mathcal{G} = \left\{ x \mapsto b_{0,k}(x; \bar{\Delta}) b_{0,l}(x; \bar{\Delta}) : 1 \leq k, l \leq J(p+1), \bar{\Delta} \in \Pi \right\}.$$

For such a class, $\sup_{g \in \mathcal{G}} |g|_\infty \lesssim J$ and $\sup_{g \in \mathcal{G}} \mathbb{V}[g] \leq \sup_{g \in \mathcal{G}} \mathbb{E}[g^2] \lesssim J$ where the second result follows from the fact that the supports of $b_{0,k}(\cdot; \bar{\Delta})$ and $b_{0,l}(\cdot; \bar{\Delta})$ shrink at the rate of J^{-1}. In addition, each function in $\mathcal{G}$ is simply a dilation and translation of a polynomial function supported on $[0,1]$, plus a zero function, and the number of polynomial degree is finite. Then, by Proposition 3.6.12 of Giné and Nickl (2016), the collection $\mathcal{G}$ of such functions is of VC type, i.e., there exists some constant C_z and $z > 6$ such that

$$N(\mathcal{G}, L_2(Q), \varepsilon \|\bar{G}\|_{L_2(Q)}) \leq \left(\frac{C_z}{\varepsilon} \right)^{2z},$$

for ε small enough where we take $\bar{G} = CJ$ for some constant $C > 0$ large enough. Theorem 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\mathbb{E} \left[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n g(x_i) - \sum_{i=1}^n \mathbb{E}[g(x_i)] \right| \right] \lesssim \sqrt{nJ \log J} + J \log J,$$

implying that

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(x_i) - \mathbb{E}[g(x_i)] \right| \lesssim_{\mathbb{P}} \sqrt{J \log J / n}.$$

Since any row or column of the matrix (a_{kl}) only contains a finite number of nonzero entries, only depending on p, the above result suffices to show that

$$\left\| \mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i) \widehat{\mathbf{b}}_0(x_i)'] - \mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i) \widehat{\mathbf{b}}_0(x_i)'] \right\| \lesssim_{\mathbb{P}} \sqrt{J \log J / n}.$$

Next, Let α_{kl} be a generic (k,l)th entry of $\mathbb{E}_{\widehat{\Delta}} \left[\widehat{\mathbf{b}}_0(x_i) \widehat{\mathbf{b}}_0(x_i)' \right] / J - \mathbb{E} \left[\mathbf{b}_0(x_i) \mathbf{b}_0(x_i)' \right] / J$, where by dividing the matrix by J, we drop the normalizing constant only for notation simplicity. By definition, it is either equal to zero or can be rewritten as

$$\alpha_{kl} = \int_{\widehat{\mathcal{B}}_j} \left(\frac{x - \hat{\tau}_j}{\hat{h}_j} \right)^\ell f(x) dx - \int_{\widehat{\mathcal{B}}_j} \left(\frac{x - \tau_j}{h_j} \right)^\ell f(x) dx$$

$$=\hat{h}_j \int_0^1 z^\ell f(z\hat{h}_j + \hat{\tau}_j)dz - h_j \int_0^1 z^\ell f(zh_j + \tau_j)dz$$

$$=(\hat{h}_j - h_j) \int_0^1 z^\ell f(z\hat{h}_j + \hat{\tau}_j)dz + h_j \int_0^1 z^\ell \Big(f(z\hat{h}_j + \hat{\tau}_j) - f(zh_j + \tau_j) \Big)dz \qquad \text{(SA-6.4)}$$

for some $1 \le j \le J$ and $0 \le \ell \le 2p$. By Assumption SA-1 and Lemma SA2 of Calonico, Cattaneo, and Titiunik (2015), $\max_{1\le j\le J} f(\hat{\tau}_j) \lesssim 1$ and $\max_{1\le j\le J} |\hat{h}_j - h_j| \lesssim_{\mathbb{P}} J^{-1}\sqrt{J \log J/n}$. Also, Lemma SA2 of Calonico, Cattaneo, and Titiunik (2015) implies that

$$\sup_{z\in[0,1]} \max_{1\le j\le J} |\hat{\tau}_j + z\hat{h}_j - (\tau_j + zh_j)| \lesssim_{\mathbb{P}} \sqrt{J \log J/n}.$$

Since $f(\cdot)$ is uniformly continuous on $\mathcal{X}$, the second term in (SA-6.4) is also $O_{\mathbb{P}}(J^{-1}\sqrt{J \log J/n})$. Again, using the sparsity structure of the matrix $[\alpha_{kl}]$, the above result suffices to show that $\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'] - \mathbf{Q}\| \lesssim_{\mathbb{P}} \sqrt{J \log J/n}$.

Given the above fact, it follows that $\|\widehat{\mathbf{Q}}^{-1}\| \lesssim_{\mathbb{P}} 1$. Notice that $\widehat{\mathbf{Q}}$ and $\mathbf{Q}$ are banded matrices with finite band width. Then the bounds on $\|\widehat{\mathbf{Q}}\|_\infty$ and $\|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}\|_\infty$ hold by Theorem 2.2 of Demko (1977). This completes the proof. □

SA-6.4 Proof of Lemma SA-2.4

Proof. Since $\mathbb{E}[\epsilon_i^2|x_i = x]$ is bounded and bounded away from zero uniformly over $x \in \mathcal{X}$, $\widehat{\mathbf{Q}} \lesssim \bar{\mathbf{\Sigma}} \lesssim \widehat{\mathbf{Q}}$. Then, by Lemma SA-2.3, $1 \lesssim_{\mathbb{P}} \lambda_{\min}(\bar{\mathbf{\Sigma}}) \lesssim \lambda_{\max}(\bar{\mathbf{\Sigma}}) \lesssim_{\mathbb{P}} 1$. The upper bound on $\bar{\Omega}(x)$ immediately follows by Lemma SA-2.3.

To establish the lower bound, it suffices to show $\inf_{x\in\mathcal{X}} \|\widehat{\mathbf{b}}_s^{(v)}(x)\| \gtrsim_{\mathbb{P}} J^{1/2+v}$. For $s = 0$, such a bound is trivial by construction. For other s, we only need to consider the case when $\widehat{\Delta} \in \Pi$. Introduce an auxiliary function $\rho(x) = (x - x_0)^v/h_{x_0}^v$ for any arbitrary point $x_0 \in \mathcal{X}$, and h_{x_0} is the length of $\mathcal{B}_{x_0}$, the bin containing x_0 in any given partition $\bar{\Delta} \in \Pi$. Let $\{\psi_j\}_{j=1}^{K_s}$ be the dual basis for B-splines $\breve{\mathbf{b}}(x) := \mathbf{b}_s(x; \bar{\Delta})/\sqrt{J}$, which is constructed as in Theorem 4.41 of Schumaker (2007). The scaling factor $\sqrt{J}$ is dropped temporarily so that the definition of $\breve{\mathbf{b}}(x)$ is consistent with that theorem. Since the B-spline basis reproduce polynomials,

$$J^v \lesssim \rho^{(v)}(x_0) = \sum_{j=1}^{K_s} (\psi_j \rho) \breve{\mathbf{b}}_{s,j}^{(v)}(x_0).$$

For any $x_0 \in \mathcal{X}$, there are only a finite number of basis functions in $\breve{\mathbf{b}}_s(x)$ supported on $\mathcal{B}_{x_0}$. By Theorem 4.41 of Schumaker (2007), for such basis functions $\breve{\mathbf{b}}_{s,j}(x)$, we have $|\psi_j \rho| \lesssim \|\rho\|_{L_\infty[\mathcal{I}_j]}$ where $\mathcal{I}_j$ denotes the support of $\breve{b}_{s,j}(x)$. All points within such $\mathcal{I}_j$ should be no greater than $(p+1)\max_{1 \le j \le J} h_j(\bar{\Delta})$ away from x_0 where $h_j(\bar{\Delta})$ denotes the length of the jth bin in $\bar{\Delta}$. Hence $\|\rho\|_{L_\infty[\mathcal{I}_j]} \lesssim 1$. Then, the desired lower bound follows.

The bound on $\Omega(x)$ can be established similarly. $\qquad\square$

SA-6.5   Proof of Lemma SA-2.5

Proof. By Lemma SA-2.1, it suffices to establish the approximation power of $\mathbf{b}_s(x; \bar{\Delta})$ for all $\bar{\Delta} \in \Pi$. For $v = 0$, by Theorem 6.27 of Schumaker (2007), $\max_{\bar{\Delta} \in \Pi} \min_{\boldsymbol{\beta} \in \mathbb{R}^{K_s}} |\mu(x) - \mathbf{b}_s(x; \bar{\Delta})'\boldsymbol{\beta}| \lesssim J^{-p-1}$. By Huang (2003) and Assumption SA-1, the Lebesgue factor of spline bases is bounded. Then, the bound on uniform approximation error coincides with that for L_2 projection error up to some universal constant.

For other $v > 0$, again, we only need to consider the case when $\bar{\Delta}$ belongs to Π. For any $\bar{\Delta} \in \Pi$, we can take the L_∞ approximation $\|\mu(x) - \mathbf{b}_s(x; \bar{\Delta})'\boldsymbol{\beta}_\infty(\bar{\Delta})\|_\infty \lesssim J^{-p-1}$, $\|\mu^{(v)}(x) - \mathbf{b}_s^{(v)}(x; \bar{\Delta})'\boldsymbol{\beta}_\infty(\bar{\Delta})\|_\infty \lesssim J^{-p-1+v}$ for some $\boldsymbol{\beta}_\infty(\bar{\Delta}) \in \mathbb{R}^{K_s}$. Such a construction exists by Lemma SA-6.1 of Cattaneo, Farrell, and Feng (2018). Then, $\|\mu^{(v)}(x) - \mathbf{b}_s^{(v)}(x; \bar{\Delta})'\boldsymbol{\beta}_\mu(\bar{\Delta})\|_\infty \lesssim \|\mu^{(v)}(x) - \mathbf{b}_s^{(v)}(x; \bar{\Delta})'\boldsymbol{\beta}_\infty(\bar{\Delta})\|_\infty + \|\mathbf{b}_s^{(v)}(x; \bar{\Delta})'(\boldsymbol{\beta}_\infty(\bar{\Delta}) - \boldsymbol{\beta}_\mu(\bar{\Delta}))\|_\infty \lesssim J^{-p-1+v} + \|\mathbf{b}_s^{(v)}(x; \bar{\Delta})'(\boldsymbol{\beta}_\infty(\bar{\Delta}) - \boldsymbol{\beta}_\mu(\bar{\Delta}))\|_\infty$. By definition of $\boldsymbol{\beta}_\mu(\bar{\Delta})$,

$$\boldsymbol{\beta}_\mu(\bar{\Delta}) - \boldsymbol{\beta}_\infty(\bar{\Delta}) = \mathbb{E}[\mathbf{b}_s(x_i; \bar{\Delta})\mathbf{b}_s(x_i; \bar{\Delta})']^{-1}\mathbb{E}[\mathbf{b}_s(x_i; \bar{\Delta})r_\infty(x_i; \bar{\Delta})],$$

where $r_\infty(x_i; \bar{\Delta}) = \mu(x_i) - \mathbf{b}_s(x_i; \bar{\Delta})'\boldsymbol{\beta}_\infty(\bar{\Delta})$. By Lemma SA-2.3, $\|\mathbb{E}[\mathbf{b}_s(x_i; \bar{\Delta})\mathbf{b}_s(x_i; \bar{\Delta})']^{-1}\|_\infty \lesssim 1$ uniformly over $\bar{\Delta} \in \Pi$. Since $\mathbf{b}_s(x_i; \bar{\Delta})$ is supported on a finite number of bins,

$$\|\mathbb{E}[\mathbf{b}_s(x_i; \bar{\Delta})r_\infty(x_i; \bar{\Delta})]\|_\infty \lesssim J^{-p-1-1/2},$$

and then the desired result follows. $\qquad\square$

SA-6.6 Proof of Lemma SA-2.6

Proof. Note that $\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)] = A_1(x) + A_2(x)$, with $A_1(x) := \widehat{\mathbf{b}}_s^{(v)}(x)'(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1})\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)]$ and $A_2(x) := \widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)]$. By definition of $\widehat{r}_\mu(\cdot)$, we have $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)] = 0$.

Define the following function class

$$\mathcal{G} := \left\{ x \mapsto b_{s,l}(x;\bar{\Delta})r_\mu(x;\bar{\Delta}) : 1 \le l \le K_s, \bar{\Delta} \in \Pi \right\}.$$

By Lemma SA-2.5, $\sup_{\bar{\Delta}\in\Pi}|r_\mu(x;\bar{\Delta})|_\infty \lesssim J^{-p-1}$. Then we have $\sup_{g\in\mathcal{G}}|g|_\infty \lesssim J^{-p-1+1/2}$, and $\sup_{g\in\mathcal{G}}\mathbb{V}[g] \lesssim J^{-2(p+1)}$. In addition, any function $g \in \mathcal{G}$ can be rewritten as

$$g(x) = b_{s,l}(x;\bar{\Delta})(\mu(x) - \mathbf{b}_s(x;\bar{\Delta})'\boldsymbol{\beta}_\mu(\bar{\Delta}) = b_{s,l}(x;\bar{\Delta})\mu(x) - \sum_{k=\underline{k}}^{\underline{k}+p} b_{s,l}(x;\bar{\Delta})b_{s,k}(x;\bar{\Delta})\beta_{\mu,k}(\bar{\Delta})$$

for some $1 \le l, \underline{k} \le K_s$ where $\beta_{\mu,k}(\bar{\Delta})$ denotes the kth element in $\boldsymbol{\beta}_\mu(\bar{\Delta})$. Here, we use the sparsity property of the partitioning basis: the summand in the second term is nonzero only if $b_{s,l}(x;\bar{\Delta})$ and $b_{s,k}(x;\bar{\Delta})$ have overlapping supports. For each l, there are only a finite number of such $b_{s,k}(x;\bar{\Delta})$ functions. Then, using the same argument given in the proof of Lemma SA-2.3,

$$N(\mathcal{G}, L_2(Q), \varepsilon\|\bar{G}\|_{L_2(Q)}) \le \left(\frac{J^l}{\varepsilon}\right)^z$$

for some finite l and z and the envelop $\bar{G} = CJ^{-p-1+1/2}$ for C large enough. By Theorem 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}g(x_i)\right| \lesssim J^{-p-1}\sqrt{\frac{\log J}{n}} + \frac{J^{-p-1+1/2}\log J}{n},$$

and, by Lemma SA-2.3, $\|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}\|_\infty \lesssim_\mathbb{P} \sqrt{J\log J/n}$. Then, using the bound on the basis given in Lemma SA-2.3,

$$\sup_{x\in\mathcal{X}}|A_1(x)| \lesssim_\mathbb{P} J^v\sqrt{J}\sqrt{\frac{J\log J}{n}}J^{-p-1}\sqrt{\frac{\log J}{n}} = J^{-p-1+v}\frac{J\log J}{n}, \quad \text{and}$$

$$\sup_{x\in\mathcal{X}}|A_2(x)| \lesssim_\mathbb{P} J^v\sqrt{J}J^{-p-1}\sqrt{\frac{\log J}{n}} = J^{-p-1+v}\sqrt{\frac{J\log J}{n}}.$$

28

These results complete the proof. □

SA-6.7 Proof of Lemma SA-2.7

Proof. By Lemma SA-2.2 and SA-2.3, $\sup_{x \in \mathcal{X}} \|\widehat{\mathbf{b}}^{(v)}(x)'\|_\infty \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$. Define a function class

$$\mathcal{G} = \Big\{ (x_1, \epsilon_1) \mapsto b_{0,l}(x_1; \bar{\Delta})\epsilon_1 : 1 \leq l \leq J(p+1), \bar{\Delta} \in \Pi \Big\}.$$

Then, $\sup_{g \in \mathcal{G}} |g| \lesssim \sqrt{J}|\epsilon_1|$, and hence take an envelop $\bar{G} = C\sqrt{J}|\epsilon_1|$ for some C large enough. Moreover, $\sup_{g \in \mathcal{G}} \mathbb{V}[g] \lesssim 1$ and, as in the proof of Lemma SA-2.3, $\mathcal{G}$ is of VC-type. By Proposition 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\sup_{g \in \mathcal{G}} \Big| \frac{1}{n} \sum_{i=1}^n g(x_i, \epsilon_i) \Big| \lesssim_{\mathbb{P}} \sqrt{\frac{\log J}{n}} + \frac{J^{\frac{\nu}{2(\nu-2)}} \log J}{n} \lesssim \sqrt{\frac{\log J}{n}},$$

and the desired result follows. □

SA-6.8 Proof of Lemma SA-2.8

Proof. We first show the convergence of $\widehat{\boldsymbol{\gamma}}$. We denote the (i,j)th element of $\mathbf{M_B}$ by M_{ij}. Then,

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = \Big(\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n M_{ij} \mathbf{w}_i \mathbf{w}_j' \Big)^{-1} \Big(\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{w}_i M_{ij} (\mu(x_j) + \epsilon_j) \Big)$$

Define $\mathbf{V} = \mathbf{W} - \mathbb{E}[\mathbf{W}|\mathbf{X}]$ and $\mathbf{H} = \mathbb{E}[\mathbf{W}|\mathbf{X}]$. Then,

$$\frac{\mathbf{W'M_B W}}{n} = \frac{\mathbf{V'M_B V}}{n} + \frac{\mathbf{H'M_B H}}{n} + \frac{\mathbf{H'M_B V}}{n} + \frac{\mathbf{V'M_B H}}{n}.$$

We have

$$\frac{\mathbf{V'M_B V}}{n} = \frac{1}{n} \sum_{i=1}^n M_{ii} \mathbf{v}_i \mathbf{v}_i' + \frac{1}{n} \sum_{i=1}^n \sum_{j \neq i} M_{ij} \mathbf{v}_i \mathbf{v}_j' = \frac{1}{n} \sum_{i=1}^n M_{ii} \mathbb{E}[\mathbf{v}_i \mathbf{v}_i'|\mathbf{X}] + O_{\mathbb{P}}\Big(\frac{1}{n}\Big) \gtrsim_{\mathbb{P}} 1,$$

where the penultimate equality holds by Lemma SA-1 of Cattaneo, Jansson, and Newey (2018b) and the last by $\frac{1}{n} \sum_{i=1}^n M_{ii} = \frac{n-K_s}{n} \gtrsim 1$. Moreover, $\frac{\mathbf{H'M_B H}}{n} \geq 0$, and $\frac{\mathbf{H'M_B V}}{n}$ has mean zero

29

conditional on $\mathbf{X}$ and by Lemma SA-1 of Cattaneo, Jansson, and Newey (2018b),

$$\left\|\frac{\mathbf{H}'\mathbf{M_P}\mathbf{V}}{n}\right\|_F \lesssim_\mathbb{P} \frac{1}{\sqrt{n}}\Big(\operatorname{trace}\Big(\frac{\mathbf{H}'\mathbf{H}}{n}\Big)\Big)^{1/2} = o_\mathbb{P}(1),$$

where $\|\cdot\|_F$ denotes Frobenius norm. Therefore, we conclude that $\frac{\mathbf{W}'\mathbf{M_B}\mathbf{W}}{n} \geq 1 + o_\mathbb{P}(1)$.

On the other hand, $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \mathbf{w}_i M_{ij}\epsilon_j$ has mean zero with variance of order $O(1/n)$ by Lemma SA-2 of Cattaneo, Jansson, and Newey (2018b). In addition, as in Lemma 2 of Cattaneo, Jansson, and Newey (2018a), let $\mathbf{G} = (\mu(x_1),\ldots,\mu(x_n))'$ and note that

$$\begin{aligned}
\frac{\mathbf{W}'\mathbf{M_B}\mathbf{G}}{n} &= \frac{\mathbf{H}'\mathbf{M_B}\mathbf{G}}{n} + \frac{\mathbf{V}'\mathbf{M_B}\mathbf{G}}{n}\\
&\lesssim \sqrt{\operatorname{trace}\Big(\frac{\mathbf{H}'\mathbf{M_B}\mathbf{H}}{n}\Big)}\sqrt{\operatorname{trace}\Big(\frac{\mathbf{G}'\mathbf{M_B}\mathbf{G}'}{n}\Big)} + \frac{1}{\sqrt{n}}\Big(\frac{\mathbf{G}'\mathbf{M_B}\mathbf{G}}{n}\Big)^{1/2}\\
&\lesssim_\mathbb{P} J^{-\varsigma\wedge(p+1)}J^{-p-1} + \frac{J^{-p-1}}{\sqrt{n}}.
\end{aligned}$$

Then, the first result follows from the rate restrictions imposed.

To show the second result, note that by Lemma SA-2.2 and SA-2.3, $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_s^{(v)}(x)'\|_\infty \lesssim_\mathbb{P} J^{1/2+v}$, $\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_\mathbb{P} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_\mathbb{P} 1$. $\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\mathbf{w}_i']$ is a $J(p+1)\times d$ matrix, and can be decomposed as follows:

$$\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\mathbf{w}_i'] = \mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\mathbb{E}[\mathbf{w}_i'|x_i]] + \mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)(\mathbf{w}_i' - \mathbb{E}[\mathbf{w}_i'|x_i])].$$

By the argument in the proof of Lemma SA-2.3 and the conditions that $\sup_{x\in\mathcal{X}}|\mathbb{E}[w_{l,i}|x_i = x]| \lesssim 1$ and $\frac{J\log J}{n} = o(1)$, $\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\mathbb{E}[\mathbf{w}_i|x_i]]\|_\infty \lesssim_\mathbb{P} J^{-1/2}$. Regarding the second term, note that it is a mean zero sequence, and for the lth covariate in $\mathbf{w}$, $l = 1,\ldots,d$,

$$\begin{aligned}
&\mathbb{V}\Big[\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)(w_{l,i} - \mathbb{E}[w_{i,l}|x_i])]\Big|\mathbf{X}\Big]\\
&\lesssim \frac{1}{n}\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\mathbb{V}[w_{l,i}|x_i]\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x) \lesssim \frac{J^{1+2v}}{n}.
\end{aligned}$$

Thus the second result follows by Markov's inequality.

Now suppose $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} \lesssim 1$ also holds. Using the argument given in Lemma SA-2.7 and the assumption that $\sup_{x\in\mathcal{X}}\mathbb{E}[|w_{l,i}|^\nu|x_i = x] \lesssim 1$ for all l, we have $\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)(w_{l,i} - \mathbb{E}[w_{i,l}|x_i])]\|_\infty \lesssim_\mathbb{P}$

$\sqrt{\log J/n}$. Thus, the last result follows. $\qquad\square$

SA-6.9 Proof of Lemma SA-2.9

Proof. Noticing that

$$
\begin{aligned}
\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x) =& \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i] + \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)] + \\
& \left(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_\mu - \mu^{(v)}(x)\right) - \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i'](\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}).
\end{aligned}
\tag{SA-6.5}
$$

Then the result follows by Lemma SA-2.5, SA-2.6, SA-2.7 and SA-2.8. $\qquad\square$

SA-6.10 Proof of Lemma SA-2.10

Proof. Since $\widehat{\epsilon}_i := y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}} = \epsilon_i + \mu(x_i) - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) =: \epsilon_i + u_i$, we can write

$$
\begin{aligned}
& \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\epsilon}_i^2] - \mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\sigma^2(x_i)] \\
=& \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'u_i^2] + 2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'u_i\epsilon_i] + \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(\epsilon_i^2 - \sigma^2(x_i))] \\
& + \left(\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\sigma^2(x_i)] - \mathbb{E}\left[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\sigma^2(x_i)\right]\right) \\
=:& \mathbf{V}_1 + \mathbf{V}_2 + \mathbf{V}_3 + \mathbf{V}_4.
\end{aligned}
$$

Now we bound each term in the following.

Step 1: For $\mathbf{V}_1$, we further write $u_i = (\mu(x_i) - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}) - \mathbf{w}_i'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) =: u_{i1} - u_{i2}$. Then

$$
\mathbf{V}_1 = \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(u_{i1}^2 + u_{i2}^2 - 2u_{i1}u_{i2})] =: \mathbf{V}_{11} + \mathbf{V}_{12} - \mathbf{V}_{13}.
$$

Since $\|2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'u_{i1}u_{i2}]\| \leq \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(u_{i1}^2 + u_{i2}^2)]\|$, it suffices to bound $\mathbf{V}_{11}$ and $\mathbf{V}_{12}$. For $\mathbf{V}_{11}$,

$$
\|\mathbf{V}_{11}\| \leq \max_{1 \leq i \leq n}|u_{i1}|^2 \left\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\right\| \lesssim_{\mathbb{P}} \frac{J\log J}{n} + J^{-2(p+1)}
$$

where the last inequality holds by Lemma SA-2.3 and SA-2.9. On the other hand,

$$
\|\mathbf{V}_{12}\| = \left\|\mathbb{E}_n\left[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\left(\sum_\ell^d w_{i\ell}^2(\widehat{\gamma}_\ell - \gamma_\ell)^2 + \sum_{\ell \neq \ell'} w_{i\ell}w_{i\ell'}(\widehat{\gamma}_\ell - \gamma)(\widehat{\gamma}_{\ell'} - \gamma_{\ell'})\right)\right]\right\|
$$

$$\lesssim \left\| \mathbb{E}_n \left[\widehat{\mathbf{b}}_s(x_i) \widehat{\mathbf{b}}_s(x_i)' \left(\sum_\ell^d w_{i\ell}^2 (\widehat{\gamma}_\ell - \gamma_\ell)^2 \right) \right] \right\|$$

by CR-inequality. By Lemma SA-2.8, $\|\widehat{\gamma} - \gamma\|^2 = o_{\mathbb{P}}(J/n)$. Then it suffices to show that for every $\ell = 1, \ldots, d$, $\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)' w_{i\ell}^2]\| \lesssim_{\mathbb{P}} 1$. Under the conditions given in the theorem, this bound can be established using the argument that will be given in Step 3 and 4.

Step 2: For $\mathbf{V}_2$, we have $\mathbf{V}_2 = 2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i(u_{i1} - u_{i2})] =: \mathbf{V}_{21} - \mathbf{V}_{22}$. Then,

$$\|\mathbf{V}_{21}\| \leq \max_{1 \leq i \leq n} |u_{i1}| \left(\left\| \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'] \right\| + \left\| \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i^2] \right\| \right) \lesssim_{\mathbb{P}} \sqrt{\frac{J \log J}{n}} + J^{-p-1}$$

where the last step follows from Lemma SA-2.3 and the result given in the next step. In addition,

$$\|\mathbf{V}_{22}\| = \left\| 2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i \sum_{\ell=1}^d w_{i\ell}(\widehat{\gamma}_\ell - \gamma_\ell)] \right\|.$$

Then, since $\left\| 2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i w_{i\ell}] \right\| \leq \left\| \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(\epsilon_i^2 + w_{i\ell}^2)] \right\|$, the result can be established using the strategy given in the next step.

Step 3: For $\mathbf{V}_3$, in view of Lemma SA-2.1 and SA-2.2, it suffices to show that

$$\sup_{\bar{\Delta} \in \Pi} \left\| \mathbb{E}_n[\mathbf{b}_0(x_i; \bar{\Delta})\mathbf{b}_0(x_i; \bar{\Delta})'(\epsilon_i^2 - \sigma^2(x_i))] \right\| \lesssim_{\mathbb{P}} \sqrt{\frac{J \log J}{n^{\frac{\nu-2}{\nu}}}}.$$

For notational simplicity, we write $\eta_i = \epsilon_i^2 - \sigma^2(x_i)$, $\eta_i^- = \eta_i \mathbb{1}(|\eta_i| \leq M) - \mathbb{E}[\eta_i \mathbb{1}(|\eta_i| \leq M)|x_i]$, $\eta_i^+ = \eta_i \mathbb{1}(|\eta_i| > M) - \mathbb{E}[\eta_i \mathbb{1}(|\eta_i| > M)|x_i]$ for some $M > 0$ to be specified later. Since $\mathbb{E}[\eta_i|x_i] = 0$, $\eta_i = \eta_i^- + \eta_i^+$. Then define a function class

$$\mathcal{G} = \left\{ (x_1, \eta_1) \mapsto b_{0,l}(x_1; \bar{\Delta})b_{0,k}(x_1; \bar{\Delta})\eta_1 : 1 \leq l \leq J(p+1), 1 \leq k \leq J(p+1), \bar{\Delta} \in \Pi \right\}.$$

Then for $g \in \mathcal{G}$, $\sum_{i=1}^n g(x_1, \eta_1) = \sum_{i=1}^n g(x_1, \eta_1^+) + \sum_{i=1}^n g(x_1, \eta_1^-)$.

Now, for the truncated piece, we have $\sup_{g \in \mathcal{G}} |g(x_1, \eta_1^-)| \lesssim JM$, and

$$\sup_{g \in \mathcal{G}} \mathbb{V}[g(x_1, \eta_1^-)] \lesssim \sup_{x \in \mathcal{X}} \mathbb{E}[\eta_1^2|x_1 = x] \sup_{\bar{\Delta} \in \Pi} \sup_{1 \leq l,k \leq J(p+1)} \mathbb{E}[b_{0,l}^2(x_1; \bar{\Delta})b_{0,k}^2(x_1; \bar{\Delta})]$$

$$\lesssim JM \sup_{x \in \mathcal{X}} \mathbb{E}\left[|\eta_1| \Big| x_i = x \right] \lesssim JM.$$

The VC condition holds by the same argument given in the proof of Lemma SA-2.3. Then using Proposition 6.2 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big|\mathbb{E}_n[g(x_i,\eta_i^-)]\Big|\Big] \lesssim \sqrt{\frac{JM\log(JM)}{n}} + \frac{JM\log(JM)}{n}.$$

Regarding the tail, we apply Theorem 2.14.1 of van der vaart and Wellner (1996) and obtain

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big|\mathbb{E}_n[g(x_i,\eta_i^+)]\Big|\Big] \lesssim \frac{1}{\sqrt{n}}J\sqrt{\log J}\mathbb{E}\Big[\sqrt{\mathbb{E}_n[|\eta_i^+|^2]}\Big]$$
$$\leq \frac{1}{\sqrt{n}}J\sqrt{\log J}(\mathbb{E}[\max_{1\leq i\leq n}|\eta_i^+|])^{1/2}(\mathbb{E}[\mathbb{E}_n[|\eta_i^+|]])^{1/2}$$
$$\lesssim \frac{J\sqrt{\log J}}{\sqrt{n}} \cdot \frac{n^{\frac{1}{\nu}}}{M^{(\nu-2)/4}}$$

where the second line follows from Cauchy-Schwarz inequality and the third line uses the fact that

$$\mathbb{E}[\max_{1\leq i\leq n}|\eta_i^+|] \lesssim \mathbb{E}[\max_{1\leq i\leq n}\epsilon_i^2] \lesssim n^{2/\nu}, \quad \text{and}$$
$$\mathbb{E}[\mathbb{E}_n[|\eta_i^+|]] \leq \mathbb{E}[|\eta_1|^+|] \lesssim \frac{\mathbb{E}[|\epsilon|^\nu]}{M^{(\nu-2)/2}}.$$

Then the desired result follows simply by setting $M = J^{\frac{2}{\nu-2}}$ and the sparsity of the basis.

Step 4: For $\mathbf{V}_4$, since by Assumption SA-1, $\sup_{x\in\mathcal{X}}\mathbb{E}[\epsilon_i^2|x_i = x] \lesssim 1$. Then, by the same argument given in the proof of Lemma SA-2.3,

$$\sup_{\bar{\Delta}\in\Pi}\Big\|\mathbb{E}_n[\mathbf{b}_s(x_i;\bar{\Delta})\mathbf{b}_s(x_i;\bar{\Delta})'\sigma^2(x_i)] - \mathbb{E}\Big[\mathbf{b}_s(x_i;\bar{\Delta})\mathbf{b}_s(x_i;\bar{\Delta})'\epsilon_i^2\Big]\Big\| \lesssim_\mathbb{P} \sqrt{J\log J/n}, \quad \text{and}$$
$$\Big\|\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i^2\Big] - \mathbb{E}\Big[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\epsilon_i^2\Big]\Big\| \lesssim_\mathbb{P} \sqrt{J\log J/n}.$$

Then the proof is complete. □

SA-6.11 Proof of Theorem SA-3.1

Proof. The proof is divided into several steps.

Step 1: We rely on the decomposition (SA-6.5). By Lemma SA-2.8, the variance of the last term is of smaller order, and thus it suffices to characterize the conditional variance of $A(x) :=$

$\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s\epsilon_i]$. By Lemma SA-2.3,

$$\int_{\mathcal{X}} \mathbb{V}[A(x)|\mathbf{X}]\omega(x)dx = \frac{1}{n}\operatorname{trace}\left(\mathbf{Q}^{-1}\mathbf{\Sigma}\mathbf{Q}^{-1}\int_{\mathcal{X}}\widehat{\mathbf{b}}_s^{(v)}(x)\widehat{\mathbf{b}}_s^{(v)}(x)'\omega(x)dx\right) + o_{\mathbb{P}}\left(\frac{J^{1+2v}}{n}\right).$$

In fact, using the argument given in the proof of Lemma SA-2.3, we also have

$$\left\|\int_{\mathcal{X}}\widehat{\mathbf{b}}_s^{(v)}(x)\widehat{\mathbf{b}}_s^{(v)}(x)'\omega(x)dx - \int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\right\| = o_{\mathbb{P}}(J^{2v}),$$

and since $\sigma^2(x)$ and $\omega(x)$ are bounded and bounded away from zero,

$$\mathscr{V}_n(p,s,v) = J^{-(1+2v)}\operatorname{trace}\left(\mathbf{Q}^{-1}\mathbf{\Sigma}\mathbf{Q}^{-1}\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\right) \asymp 1.$$

Step 2: By decomposition (SA-6.5),

$$\mathbb{E}[\widehat{\mu}^{(v)}(x)|\mathbf{X},\mathbf{W}] - \mu^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)] + \left(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_\mu - \mu^{(v)}(x)\right)$$

$$- \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i']\mathbb{E}[(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})|\mathbf{X},\mathbf{W}]$$

$$=: \mathfrak{B}_1(x) + \mathfrak{B}_2(x) + \mathfrak{B}_3(x).$$

By Lemma SA-2.6, $\int_{\mathcal{X}}\mathfrak{B}_1(x)^2\omega(x)dx = o_{\mathbb{P}}(J^{-2p-2+2v})$. By Lemma SA-2.8, $\int_{\mathcal{X}}\mathfrak{B}_3(x)^2\omega(x)dx = o_{\mathbb{P}}(J^{-2p-2+2v})$. By Lemma SA-2.5, $\int_{\mathcal{X}}\mathfrak{B}_2(x)^2\omega(x)dx \lesssim_{\mathbb{P}} J^{-2p-2+2v}$. By Cauchy-Schwarz inequality, we can safely ignore the integrals of those cross-product terms in the IMSE expansion, and thus the leading term in the integrated squared bias is

$$J^{2p+2-2v}\int_{\mathcal{X}}\left(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_\mu - \mu^{(v)}(x)\right)^2\omega(x)dx \lesssim_{\mathbb{P}} 1.$$

Then, by Lemma SA-6.1 of Cattaneo, Farrell, and Feng (2018), for $s = p$,

$$\sup_{x \in \mathcal{X}}\left|\mu^{(v)}(x) - \widehat{\mathbf{b}}_p^{(v)}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta}) - \frac{\mu^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{E}_{p+1-v}\left(\frac{x - \hat{\tau}_x^L}{\hat{h}_x}\right)\right| = o_{\mathbb{P}}(J^{-(p+1-v)}) \quad \text{(SA-6.6)}$$

where for each $m \in \mathbb{Z}_+$, $\mathscr{E}_m(\cdot)$ is the mth Bernoulli polynomial, $\hat{\tau}_x^L$ is the start of the (random) interval in $\widehat{\Delta}$ containing x and $\hat{h}_x$ denotes its length. Note that when $s < p$, $\widehat{\mathbf{b}}_p(x)'\boldsymbol{\beta}_\infty$ is still an element in the space spanned by $\widehat{\mathbf{b}}_s(x)$. In other words, it provides a valid approximation of $\mu^{(v)}(x)$

in the larger space in terms of sup-norm. Then it follows that

$$
\begin{aligned}
&\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_\mu - \mu^{(v)}(x)\\
&= \widehat{\mathbf{b}}_s^{(v)}(x)'\Big(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\Big)^{-1}\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\mu(x_i)] - \mu^{(v)}(x)\\
&= \widehat{\mathbf{b}}_s^{(v)}(x)'\Big(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\Big)^{-1}\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_s(x_i)\frac{\mu^{(p+1)}(x_i)}{(p+1)!}\hat{h}_{x_i}^{p+1}\mathscr{E}_{p+1}\Big(\frac{x_i - \hat{\tau}_{x_i}^L}{\hat{h}_{x_i}}\Big)\Big]\\
&\quad - \frac{\mu^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{E}_{p+1-v}\Big(\frac{x - \hat{\tau}_x^L}{\hat{h}_x}\Big) + o_{\mathbb{P}}(J^{-p-1+v})\\
&= J^{-p-1}\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_0(x_i)\frac{\mu^{(p+1)}(x_i)}{(p+1)!f(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i - \hat{\tau}_{x_i}^L}{\hat{h}_{x_i}}\Big)\Big]\\
&\quad - \frac{J^{-p-1+v}\mu^{(p+1)}(x)}{(p+1-v)!f(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x - \hat{\tau}_x^L}{\hat{h}_x}\Big) + o_{\mathbb{P}}(J^{-p-1+v}) \qquad\text{(SA-6.7)}
\end{aligned}
$$

where the last step uses Lemma SA-2.1-SA-2.3, and $o_{\mathbb{P}}(\cdot)$ in the above is understood in terms of sup-norm over $x \in \mathcal{X}$. Taking integral of the squared bias and using Assumption SA-1 and Lemma SA-2.1-SA-2.3 again, we have three leading terms:

$$
\begin{aligned}
M_1(x) :=& \int_{\mathcal{X}}\Big(\frac{J^{-p-1+v}\mu^{(p+1)}(x)}{(p+1-v)!f(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x - \hat{\tau}_x^L}{\hat{h}_x}\Big)\Big)^2\omega(x)dx\\
=& \frac{J^{-2p-2+2v}|\mathscr{E}_{2p+2-2v}|}{(2p+2-2v)!}\int_{\mathcal{X}}\Big[\frac{\mu^{(p+1)}(x)}{f(x)^{p+1-v}}\Big]^2\omega(x)dx + o_{\mathbb{P}}(J^{-2p-2+2v}),\\
M_2(x) :=& J^{-2p-2}\int_{\mathcal{X}}\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_0(x_i)\frac{\mu^{(p+1)}(x_i)}{(p+1)!f(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i - \hat{\tau}_{x_i}^L}{\hat{h}_{x_i}}\Big)\Big]\Big)^2\omega(x)dx\\
=& J^{-2p-2}\boldsymbol{\xi}_{0,f}'\mathbf{T}_s'\mathbf{Q}^{-1}\Big(\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\Big)\mathbf{Q}^{-1}\mathbf{T}_s\boldsymbol{\xi}_{0,f} + o_{\mathbb{P}}(J^{-2p-2+2v}),\\
M_3(x) :=& J^{-2p-2+v}\int_{\mathcal{X}}\Big\{\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_0(x_i)\frac{\mu^{(p+1)}(x_i)}{(p+1)!f(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i - \hat{\tau}_{x_i}^L}{\hat{h}_{x_i}}\Big)\Big]\Big)\\
&\quad\times\frac{\mu^{(p+1)}(x)}{(p+1-v)!f(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x - \hat{\tau}_x^L}{\hat{h}_x}\Big)\Big\}\omega(x)dx\\
=& J^{-2p-2+v}\boldsymbol{\xi}_{0,f}'\mathbf{T}_s'\mathbf{Q}^{-1}\mathbf{T}_s\boldsymbol{\xi}_{v,\omega} + o_{\mathbb{P}}(J^{-2p-2+2v}),
\end{aligned}
$$

where $\mathscr{E}_{2p+2-2v}$ is the $(2p+2-2v)$th Bernoulli number, and for a weighting function $\lambda(\cdot)$ (which can be replaced by $f(\cdot)$ and $\omega(\cdot)$ respectively), we define

$$
\boldsymbol{\xi}_{v,\lambda} = \int_{\mathcal{X}}\mathbf{b}_0^{(v)}(x)\frac{\mu^{(p+1)}(x)}{(p+1-v)!f(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x - \tau_x^L}{h_x}\Big)\lambda(x)dx.
$$

35

τ_x and h_x are defined the same way as $\hat{\tau}_x$ and $\hat{h}_x$, but with respect to Δ, the partition based on population quantiles. Therefore, the leading terms now only rely on the non-random partition Δ as well as other deterministic functions, which are simply equivalent to the leading bias if we repeat the above derivation but set $\widehat{\Delta} = \Delta$. Then the proof is complete. $\qquad\square$

SA-6.12 Proof of Corollary SA-3.1

Proof. The proof is divided into two steps.

Step 1: Consider the special case in which $s = 0$. $\mathscr{V}_n(p, 0, v)$ depends on three matrices: $\mathbf{Q}$, $\boldsymbol{\Sigma}$ and $\int_{\mathcal{X}} \mathbf{b}_0^{(v)}(x) \mathbf{b}_0^{(v)}(x)' \omega(x) dx$. Importantly, they are block diagonal with finite block sizes, and the basis functions that form these matrices have local supports. Then by continuity of $\omega(x)$, $f(x)$ and $\sigma^2(x)$, these matrices can be further approximated:

$$\mathbf{Q} = \check{\mathbf{Q}} \mathfrak{D}_f + o_{\mathbb{P}}(1), \quad \boldsymbol{\Sigma} = \check{\mathbf{Q}} \mathfrak{D}_{\sigma^2 f} + o_{\mathbb{P}}(1), \text{ and } \int_{\mathcal{X}} \mathbf{b}_0^{(v)}(x) \mathbf{b}_0^{(v)}(x)' \omega(x) dx = \check{\mathbf{Q}}_v \mathfrak{D}_\omega + o_{\mathbb{P}}(J^{2v})$$

where

$$\check{\mathbf{Q}} = \int_{\mathcal{X}} \mathbf{b}_0(x) \mathbf{b}_0(x)' dx, \quad \check{\mathbf{Q}}_v = \int_{\mathcal{X}} \mathbf{b}_0^{(v)}(x) \mathbf{b}_0^{(v)}(x)' dx, \quad \mathfrak{D}_f = \operatorname{diag}\{f(\check{x}_1), \ldots, f(\check{x}_{J(p+1)})\},$$

$$\mathfrak{D}_{\sigma^2 f} = \operatorname{diag}\{\sigma^2(\check{x}_1) f(\check{x}_1), \ldots, \sigma^2(\check{x}_{J(p+1)}) f(\check{x}_{J(p+1)})\}, \quad \text{and} \quad \mathfrak{D}_\omega = \operatorname{diag}\{\omega(\check{x}_1), \ldots, \omega(\check{x}_{J(p+1)})\}.$$

"$o_{\mathbb{P}}(\cdot)$" in the above equations means the operator norm of matrix differences is $o_{\mathbb{P}}(\cdot)$, and for $l = 1, \ldots, J(p+1)$, each $\check{x}_l$ is an arbitrary point in the support of $b_{0,l}(x)$. For simplicity, we choose these points such that $x_l = x_{l'}$ if $b_{0,l}(\cdot)$ and $b_{0,l'}(\cdot)$ have the same support. Therefore we have

$$\int_{\mathcal{X}} \mathbb{V}[A(x) | \mathbf{X}] \omega(x) dx = \frac{1}{n} \operatorname{trace}\left(\mathfrak{D}_{\sigma^2 \omega / f} \check{\mathbf{Q}}^{-1} \check{\mathbf{Q}}_v \right) + o_{\mathbb{P}}\left(\frac{J^{1+2v}}{n} \right)$$

where $\mathfrak{D}_{\sigma^2 \omega / f} = \operatorname{diag}\{\sigma^2(\check{x}_1) \omega(\check{x}_1) / f(\check{x}_1), \ldots, \sigma^2(\check{x}_{J(p+1)}) \omega(\check{x}_{J(p+1)}) / f(\check{x}_{J(p+1)})\}$.

Finally, by change of variables, we can write $\check{\mathbf{Q}}^{-1} \check{\mathbf{Q}}_v$ as another block diagonal matrix $\bar{\mathbf{Q}} = \operatorname{diag}\{\bar{\mathbf{Q}}_1, \ldots, \bar{\mathbf{Q}}_J\}$ where the lth block $\bar{\mathbf{Q}}_l$, $l = 1, \ldots, j$, can be written as

$$\bar{\mathbf{Q}}_l = \left(\int_0^1 \boldsymbol{\psi}(z) \boldsymbol{\psi}(z)' dz \right)^{-1} \int_0^1 \boldsymbol{\psi}^{(v)}(z) \boldsymbol{\psi}^{(v)}(z)' dz$$

36

where $\boldsymbol{\psi}(z) = (1, z, \ldots, z^p)$. Employing Lemma SA-2.1 and letting the trace converge to the Riemann integral, we conclude that

$$\int_{\mathcal{X}} \mathbb{V}[A(x)|\mathbf{X}]\omega(x)dx = \frac{J^{1+2v}}{n}\mathscr{V}(p, 0, v) + o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n}\Big).$$

where $\mathscr{V}(p, 0, v) := \mathrm{trace}\Big\{\Big(\int_0^1 \boldsymbol{\psi}(z)\boldsymbol{\psi}(z)'dz\Big)^{-1}\int_0^1 \boldsymbol{\psi}^{(v)}(z)\boldsymbol{\psi}^{(v)}(z)'dz\Big\}\int_{\mathcal{X}} \sigma^2(x)f(x)^{2v}\omega(x)dx.$

Step 2: Now consider the special case in which $s = 0$. By Lemma A.3 of Cattaneo, Farrell, and Feng (2018), we can construct an L_{∞} approximation error

$$r_{\infty}^{(v)}(x; \widehat{\Delta}) := \mu^{(v)}(x) - \widehat{\mathbf{b}}_0^{(v)}(x)'\boldsymbol{\beta}_{\infty}(\widehat{\Delta}) = \frac{\mu^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{B}_{p+1-v}\Big(\frac{x - \hat{\tau}_x^L}{\hat{h}_x}\Big) + o_{\mathbb{P}}(J^{-(p+1-v)})$$

where for each $m \in \mathbb{Z}_+$, $\binom{2m}{m}\mathscr{B}_m(\cdot)$ is the mth shifted Legendre polynomial on $[0, 1]$, $\hat{\tau}_x^L$ is the start of the (random) interval in $\widehat{\Delta}$ containing x and $\hat{h}_x$ denotes its length. In addition,

$$\max_{1 \leq j \leq J(p+1)} |\mathbb{E}_{\widehat{\Delta}}[\widehat{b}_{0,j}(x)r_{\infty}(x; \widehat{\Delta})]|$$

$$= \max_{1 \leq j \leq J(p+1)} \Big|\int_{\mathcal{X}} \widehat{b}_{0,j}(x)r_{\infty}(x; \widehat{\Delta})f(x)dx\Big|$$

$$= \max_{1 \leq j \leq J(p+1)} \Big|\int_{\hat{\tau}_x^L}^{\hat{\tau}_x^L + \hat{h}_x} \widehat{b}_{0,j}(x)r_{\infty}(x; \widehat{\Delta})f(\hat{\tau}_x^L)dx\Big| + o_{\mathbb{P}}(J^{-p-1-1/2})$$

$$= \max_{1 \leq j \leq J(p+1)} \Big|f(\hat{\tau}_x^L)\frac{\mu^{(p+1)}(x)J^{-p-1}}{(p+1)!}\int_{\hat{\tau}_x^L}^{\hat{\tau}_x^L + \hat{h}_x} \widehat{b}_{0,j}(x)\mathscr{B}_{p+1}\Big(\frac{x - \hat{\tau}_x^L}{\hat{h}_x}\Big)dx\Big| + o_{\mathbb{P}}(J^{-p-1-1/2})$$

$$= o_{\mathbb{P}}(J^{-p-1-1/2})$$

where the last line follows by change of variables and the orthogonality of Legendre polynomials. Thus $r_{\infty}(x; \widehat{\Delta})$ is approximately orthogonal to the space spanned by $\widehat{\mathbf{b}}(x)$. Immediately, we have

$$\|\mathbb{E}_{\widehat{\Delta}}[\mathbf{b}(x; \widehat{\Delta})r_{\infty}(x; \widehat{\Delta})]\| = o_{\mathbb{P}}(J^{-p-1}).$$

Since $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x)r_{\mu}(x; \widehat{\Delta})] = 0$,

$$\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}(x)(r_{\mu}(x; \widehat{\Delta}) - r_{\infty}(x; \widehat{\Delta}))]\| = \|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}(x)\widehat{\mathbf{b}}(x)'(\boldsymbol{\beta}_{\infty}(\widehat{\Delta}) - \boldsymbol{\beta}_{\mu}(\widehat{\Delta}))]\| = o_{\mathbb{P}}(J^{-p-1}).$$

By Lemma SA-2.3, $\lambda_{\min}(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)]') \gtrsim_{\mathbb{P}} 1$, and thus $\|\boldsymbol{\beta}_\infty(\widehat{\Delta}) - \boldsymbol{\beta}_\mu(\widehat{\Delta})\| = o_{\mathbb{P}}(J^{-p-1})$. Then,

$$\int_{\mathcal{X}} \left(\widehat{\mathbf{b}}_0^{(v)}(x)'(\boldsymbol{\beta}(\widehat{\Delta}) - \boldsymbol{\beta}_\infty(\widehat{\Delta}))\right)^2 \omega(x)dx$$
$$\leq \lambda_{\max}\left(\int_{\mathcal{X}} \widehat{\mathbf{b}}_0^{(v)}(x)\widehat{\mathbf{b}}_0^{(v)}(x)'\omega(x)dx\right)\|\boldsymbol{\beta}(\widehat{\Delta}) - \boldsymbol{\beta}_\infty(\widehat{\Delta})\|^2 = o_{\mathbb{P}}(J^{-2p-2+2v}).$$

Therefore, we can represent the leading term in the integrated squared bias by L_∞ approximation error: $\int_{\mathcal{X}} \mathfrak{B}_2(x)^2\omega(x)dx = \int_{\mathcal{X}}(\mu^{(v)}(x) - \widehat{\mathbf{b}}^{(v)}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta}))^2\omega(x)dx + o_{\mathbb{P}}(J^{-2p-2+2v})$. Finally, using the results given in Lemma SA-2.1, change of variables and the definition of Riemann integral, we conclude that

$$\int_{\mathcal{X}} \left(\mathbb{E}[\widehat{\mu}^{(v)}(x)|\mathbf{X}, \mathbf{W}] - \mu^{(v)}(x)\right)^2 \omega(x)dx = J^{-2(p+1-v)}\mathscr{B}(p, 0, v) + o_{\mathbb{P}}(J^{-2p-2+2v})$$

where

$$\mathscr{B}(p, 0, v) = \frac{\int_0^1 [\mathscr{B}_{p+1-v}(z)]^2 dz}{((p+1-v)!)^2} \int_{\mathcal{X}} \frac{[\mu^{(p+1)}(x)]^2}{f(x)^{2p+2-2v}}\omega(x)dx.$$

Then the proof is complete. $\qquad\square$

SA-6.13 Proof of Theorem SA-3.2

Proof. By Lemma SA-2.5-SA-2.8, We first show

$$\bar{\Omega}(x)^{-1/2}\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{G}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i] =: \mathbb{G}_n[a_i\epsilon_i]$$

is asymptotically normal. Conditional on $\mathbf{X}$, it is a mean zero independent sequence over i with variance equal to 1. Then by Berry-Esseen inequality,

$$\sup_{u\in\mathbb{R}} \left|\mathbb{P}(\mathbb{G}_n[a_i\epsilon_i] \leq u|\mathbf{X}) - \Phi(u)\right| \leq \min\left(1, \frac{\sum_{i=1}^n \mathbb{E}[|a_i\epsilon_i|^3|\mathbf{X}]}{n^{3/2}}\right).$$

Now, using Lemma SA-2.3 and SA-2.4,

$$\frac{1}{n^{3/2}}\sum_{i=1}^n \mathbb{E}\left[|a_i\epsilon_i|^3\Big|\mathbf{X}\right] \lesssim \bar{\Omega}(x)^{-3/2}\frac{1}{n^{3/2}}\sum_{i=1}^n \mathbb{E}\left[|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)\epsilon_i|^3\Big|\mathbf{X}\right]$$

$$\lesssim \bar{\Omega}(x)^{-3/2} \frac{1}{n^{3/2}} \sum_{i=1}^{n} |\widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{b}}_s(x_i)|^3$$

$$\leq \bar{\Omega}(x)^{-3/2} \frac{\sup_{x \in \mathcal{X}} \sup_{z \in \mathcal{X}} |\widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{b}}_s(z)|}{n^{3/2}} \sum_{i=1}^{n} |\widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{b}}_s(x_i)|^2$$

$$\lesssim_{\mathbb{P}} \frac{1}{J^{3/2+3v}} \cdot \frac{J^{1+v}}{\sqrt{n}} \cdot J^{1+2v} \to 0$$

since $J/n = o(1)$. By Lemma SA-2.10, the above weak convergence still holds if $\bar{\Omega}(x)$ is replaced by $\widehat{\Omega}(x)$. Now, the desired result follows by Lemma SA-2.5, SA-2.6 and SA-2.8. $\qquad\square$

SA-6.14 Proof of Corollary SA-3.2

Proof. Note that for a given p, by Theorem SA-3.1, $J_{\mathtt{IMSE}} \asymp n^{\frac{1}{2p+3}}$. Then, for $(p+q)$th-order binscatter estimator, $nJ_{\mathtt{IMSE}}^{-2p-2q-3} = o(1)$ and $\frac{J_{\mathtt{IMSE}}^2 \log^2 J_{\mathtt{IMSE}}}{n} = o(1)$. Then the conclusion of Theorem SA-3.2 holds for the $(p+q)$th-order binscatter estimator. Then the result immediately follows.

$\qquad\square$

SA-6.15 Proof of Theorem SA-3.3

Proof. The proof is divided into several steps.

Step 1: Note that

$$\sup_{x \in \mathcal{X}} \left| \frac{\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}} - \frac{\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)}{\sqrt{\Omega(x)/n}} \right|$$

$$\leq \sup_{x \in \mathcal{X}} \left| \frac{\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)}{\sqrt{\Omega(x)/n}} \right| \sup_{x \in \mathcal{X}} \left| \frac{\widehat{\Omega}(x)^{1/2} - \Omega(x)^{1/2}}{\widehat{\Omega}(x)^{1/2}} \right|$$

$$\lesssim_{\mathbb{P}} \left(\sqrt{\log J} + \sqrt{n} J^{-p-1-1/2} \right) \left(J^{-p-1} + \sqrt{\frac{J \log J}{n^{1-\frac{2}{\nu}}}} \right)$$

where the last step uses Lemma SA-2.4 and SA-2.9. Then, in view of Lemma SA-2.5, SA-2.6, SA-2.8 and SA-2.10 and the rate restriction given in the lemma, we have

$$\sup_{x \in \mathcal{X}} \left| \frac{\widehat{\mu}^{(v)}(x) - \mu(x)}{\sqrt{\widehat{\Omega}(x)/n}} - \frac{\widehat{\mathbf{b}}_s(x)' \widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}} \mathbb{G}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i] \right| = o_{\mathbb{P}}(a_n^{-1}).$$

Step 2: Let us write $\mathscr{K}(x, x_i) = \Omega(x)^{-1/2} \widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \mathbf{b}_s(x_i)$. Now we rearrange $\{x_i\}_{i=1}^{n}$ as a

sequence of order statistics $\{x_{(i)}\}_{i=1}^n$, i.e., $x_{(1)} \leq \cdots \leq x_{(n)}$. Accordingly, $\{\epsilon_i\}_{i=1}^n$ and $\{\sigma^2(x_i)\}_{i=1}^n$ are ordered as concomitants $\{\epsilon_{[i]}\}_{i=1}^n$ and $\{\sigma_{[i]}^2\}_{i=1}^n$ where $\sigma_{[i]}^2 = \sigma^2(x_{(i)})$. Clearly, conditional on $\mathbf{X}$, $\{\epsilon_{[i]}\}_{i=1}^n$ is still an independent mean zero sequence. Then by Assumption SA-1 and Sakhanenko (1991), there exists a sequence of i.i.d. standard normal random variables $\{\zeta_{[i]}\}_{i=1}^n$ such that

$$\max_{1 \leq \ell \leq n} |S_\ell| := \max_{1 \leq \ell \leq n} \left| \sum_{i=1}^l \epsilon_{[i]} - \sum_{i=1}^l \sigma_{[i]} \zeta_{[i]} \right| \lesssim_{\mathbb{P}} n^{\frac{1}{2+\nu}}.$$

Then, using summation by parts,

$$\sup_{x \in \mathcal{X}} \left| \sum_{i=1}^n \mathscr{K}(x, x_{(i)})(\epsilon_{[i]} - \sigma_{[i]} \zeta_{[i]}) \right|$$

$$= \sup_{x \in \mathcal{X}} \left| \mathscr{K}(x, x_{(n)}) S_n - \sum_{i=1}^{n-1} S_i \left(\mathscr{K}(x, x_{(i+1)}) - \mathscr{K}(x, x_{(i)}) \right) \right|$$

$$\leq \sup_{x \in \mathcal{X}} \max_{1 \leq i \leq n} |\mathscr{K}(x, x_i)| |S_n| + \sup_{x \in \mathcal{X}} \left| \frac{\widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}} \sum_{i=1}^{n-1} S_i \left(\widehat{\mathbf{b}}_s(x_{(i+1)}) - \widehat{\mathbf{b}}_s(x_{(i)}) \right) \right|$$

$$\leq \sup_{x \in \mathcal{X}} \max_{1 \leq i \leq n} |\mathscr{K}(x, x_i)| |S_n| + \sup_{x \in \mathcal{X}} \left\| \frac{\widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}} \right\|_\infty \left\| \sum_{i=1}^{n-1} S_i \left(\widehat{\mathbf{b}}_s(x_{(i+1)}) - \widehat{\mathbf{b}}_s(x_{(i)}) \right) \right\|_\infty$$

By Lemma SA-2.3 and SA-2.4, $\sup_{x \in \mathcal{X}} \sup_{x_i \in \mathcal{X}} |\mathscr{K}(x, x_i)| \lesssim_{\mathbb{P}} \sqrt{J}$, and

$$\sup_{x \in \mathcal{X}} \left\| \frac{\widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}} \right\|_\infty \lesssim_{\mathbb{P}} 1.$$

Then, notice that

$$\max_{1 \leq l \leq K_s} \left| \sum_{i=1}^{n-1} \left(\widehat{b}_{s,l}(x_{(i+1)}) - \widehat{b}_{s,l}(x_{(i)}) \right) S_l \right| \leq \max_{1 \leq l \leq K_s} \sum_{i=1}^{n-1} \left| \widehat{b}_{s,l}(x_{(i+1)}) - \widehat{b}_{s,l}(x_{(i)}) \right| \max_{1 \leq \ell \leq n} |S_\ell|.$$

By construction of the ordering, $\max_{1 \leq l \leq K_s} \sum_{i=1}^{n-1} \left| \widehat{b}_{s,l}(x_{(i+1)}) - \widehat{b}_{s,l}(x_{(i)}) \right| \lesssim \sqrt{J}$. Under the rate restriction in the theorem, this suffices to show that for any $\eta > 0$,

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} \mathbb{G}_n[\mathscr{K}(x, x_i)(\epsilon_i - \sigma_i \zeta_i)] > \eta a_n^{-1} | \mathbf{X} \right) = o_{\mathbb{P}}(1)$$

where we recover the original ordering. Since $\mathbb{G}_n[\widehat{\mathbf{b}}(x_i) \zeta_i \sigma_i] =_{d|\mathbf{X}} \mathbf{N}(0, \bar{\boldsymbol{\Sigma}})$ ($=_{d|\mathbf{X}}$ denotes "equal in

distribution conditional on $\mathbf{X}$"), the above steps construct the following approximating process:

$$\bar{Z}_p(x) := \frac{\widehat{\mathbf{b}}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}}\bar{\boldsymbol{\Sigma}}^{1/2}\mathbf{N}_{K_s}.$$

Then it remains to show $\widehat{\mathbf{Q}}^{-1}$ and $\bar{\boldsymbol{\Sigma}}$ can be replaced by their population analogues without affecting the approximation, which is verified in the next step.

Step 3: Note that

$$\sup_{x\in\mathcal{X}}|\bar{Z}_p(x) - Z_p(x)| \le \sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mathbf{b}}^{(v)}(x)'(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1})}{\sqrt{\Omega(x)}}\bar{\boldsymbol{\Sigma}}^{1/2}\mathcal{N}_{K_s}\right|$$
$$+ \sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mathbf{b}}^{(v)}(x)'\mathbf{Q}^{-1}}{\sqrt{\Omega(x)}}\left(\bar{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2}\right)\mathcal{N}_{K_s}\right|$$
$$+ \sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mathbf{b}}_0^{(v)}(x)'(\widehat{\mathbf{T}}_s - \mathbf{T}_s)\mathbf{Q}^{-1}}{\sqrt{\Omega(x)}}\boldsymbol{\Sigma}^{1/2}\mathcal{N}_{K_s}\right|$$

where each term on the right-hand-side is a mean-zero Gaussian process conditional on $\mathbf{X}$. By Lemma SA-2.2 and SA-2.3, $\|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$ and $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$. Also, using the argument in the proof of Lemma SA-2.3 and Theorem X.3.8 of Bhatia (2013), $\|\bar{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2}\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$. By Gaussian Maximal Inequality (see Chernozhukov, Lee, and Rosen, 2013, Lemma 13),

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}|\bar{Z}_p(x) - Z_p(x)|\Big|\mathbf{X}\right] \lesssim_{\mathbb{P}} \sqrt{\log J}\left(\|\bar{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2}\| + \|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}\| + \|\widehat{\mathbf{T}}_s - \mathbf{T}_s\|\right) = o_{\mathbb{P}}(a_n^{-1})$$

where the last line follows from the imposed rate restriction. Then the proof is complete. $\qquad\square$

SA-6.16 Proof of Theorem SA-3.4

Proof. This conclusion follows from Lemma SA-2.3, SA-2.10 and Gaussian Maximal Inequality as applied in Step 3 in the proof of Theorem SA-3.3. $\qquad\square$

SA-6.17 Proof of Theorem SA-3.5

Proof. We first show that

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{T}_p(x)| \leq u \Big) - \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \leq u \Big) \right| = o(1).$$

By Theorem SA-3.3, there exists a sequence of constants η_n such that $\eta_n = o(1)$ and

$$\mathbb{P}\Big(\Big| \sup_{x \in \mathcal{X}} |\widehat{T}_p(x)| - \sup_{x \in \mathcal{X}} |Z_p(x)| \Big| > \eta_n/a_n \Big) = o(1).$$

Then,

$$\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{T}(x)| \leq u \Big) \leq \mathbb{P}\Big(\Big\{ \sup_{x \in \mathcal{X}} |\widehat{T}(x)| \leq u \Big\} \cap \Big\{ \Big| \sup_{x \in \mathcal{X}} |\widehat{T}_p(x)| - \sup_{x \in \mathcal{X}} |Z_p(x)| \Big| \leq \eta_n/a_n \Big\} \Big) + o(1)$$

$$\leq \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \leq u + \eta_n/a_n \Big) + o(1)$$

$$\leq \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \leq u \Big) + \sup_{u \in \mathbb{R}} \mathbb{E}\Big[\mathbb{P}\Big(\Big| \sup_{x \in \mathcal{X}} |Z_p(x)| - u \Big| \leq \eta_n/a_n \Big| \mathbf{X} \Big) \Big]$$

$$\leq \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \leq u \Big) + \mathbb{E}\Big[\sup_{u \in \mathbb{R}} \mathbb{P}\Big(\Big| \sup_{x \in \mathcal{X}} |Z_p(x)| - u \Big| \leq \eta_n/a_n \Big| \mathbf{X} \Big) \Big] + o(1).$$

Now, apply Anti-Concentration Inequality (conditional on $\mathbf{X}$) to the second term:

$$\sup_{u \in \mathbb{R}} \mathbb{P}\Big(\Big| \sup_{x \in \mathcal{X}} |Z_p(x)| - u \Big| \leq \eta_n/a_n \Big| \mathbf{X} \Big) \leq 4\eta_n a_n^{-1} \mathbb{E}\Big[\sup_{x \in \mathcal{X}} |Z_p(x)| \Big| \mathbf{X} \Big] + o(1)$$

$$\lesssim_{\mathbb{P}} \eta_n a_n^{-1} \sqrt{\log J} + o(1) \to 0$$

where the last step uses Gaussian Maximal Inequality (see Chernozhukov, Lee, and Rosen, 2013, Lemma 13). By Dominated Convergence Theorem,

$$\mathbb{E}\Big[\sup_{u \in \mathbb{R}} \mathbb{P}\Big(\Big| \sup_{x \in \mathcal{X}} |Z_p(x)| - u \Big| \leq \eta_n/a_n \Big| \mathbf{X} \Big) \Big] = o(1).$$

The other side of the inequality follows similarly.

By similar argument, using Theorem SA-3.4 we have

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_p(x)| \leq u \Big| \mathbf{X} \Big) - \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \leq u \Big| \mathbf{X} \Big) \right| = o_{\mathbb{P}}(1).$$

Then it remains to show that

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \le u \Big) - \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \le u | \mathbf{X} \Big) \right| = o_{\mathbb{P}}(1). \tag{SA-6.8}$$

Now, we note that we can write

$$Z_p(x) = \frac{\widehat{\mathbf{b}}_0^{(v)}(x)'}{\sqrt{\widehat{\mathbf{b}}_0^{(v)}(x)' \mathbf{V} \widehat{\mathbf{b}}_0^{(v)}(x)}} \breve{N}_{K_s}$$

where $\mathbf{V} = \mathbf{T}_s' \mathbf{Q}^{-1} \mathbf{\Sigma} \mathbf{Q}^{-1} \mathbf{T}_s$ and $\breve{\mathbf{N}}_{K_s} := \mathbf{T}_s' \mathbf{Q}^{-1} \mathbf{\Sigma}^{1/2} \mathbf{N}_{K_s}$ is a K_s-dimensional normal random vector. Importantly, by this construction, $\breve{\mathbf{N}}_{K_s}$ and $\mathbf{V}$ do not depend on $\widehat{\Delta}$ and x, and they are only determined by the deterministic partition Δ.

Now, first consider $v = 0$. For any two partitions $\Delta_1, \Delta_2 \in \Pi$, for any $x \in \mathcal{X}$, there exists $\breve{x} \in \mathcal{X}$ such that

$$\mathbf{b}_0^{(v)}(x; \Delta_1) = \mathbf{b}_0^{(v)}(\breve{x}; \Delta_2),$$

and vice versa. Therefore, the following two events are equivalent: $\{\omega : \sup_{x \in \mathcal{X}} |Z_p(x; \Delta_1)| \le u\} = \{\omega : \sup_{x \in \mathcal{X}} |Z_p(x; \Delta_2)| \le u\}$ for any u. Thus,

$$\mathbb{E}\Big[\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \le u \Big| \mathbf{X} \Big) \Big] = \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |Z_p(x)| \le u \Big| \mathbf{X} \Big).$$

Then for $v = 0$, the desired result follows.

For $v > 0$, simply notice that $\widehat{\mathbf{b}}_0^{(v)}(x) = \widehat{\mathfrak{T}}_v \widehat{\mathbf{b}}_0(x)$ for some transformation matrix $\widehat{\mathfrak{T}}_v$. Clearly, $\widehat{\mathfrak{T}}_v$ takes a similar structure as $\widehat{\mathbf{T}}_s$: each row and each column only have a finite number of nonzeros. Each nonzero element is simply $\hat{h}_j^{-v}$ up to some constants. By the similar argument given in the proof of Lemma SA-2.2, it can be shown that $\|\widehat{\mathfrak{T}}_v - \mathfrak{T}_v\| \lesssim \sqrt{J \log J/n}$ where $\mathfrak{T}_v$ is the population analogue ($\hat{h}_j$ replaced by h_j). Repeating the argument given in the proof of Theorem SA-3.3 and SA-3.4, we can replace $\widehat{\mathfrak{T}}_v$ in $Z_p(x)$ by $\mathfrak{T}_v$ without affecting the approximation rate. Then the desired result follows by repeating the argument given for $v = 0$ above. $\qquad \square$

SA-6.18   Proof of Corollary SA-3.3

Proof. Given $J = J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$, the rate restrictions required in Theorem SA-3.5 are satisfied. Let $\eta_{1,n} = o(1)$, $\eta_{2,n} = o(1)$ and $\eta_{3,n} = o(1)$. Then,

$$
\begin{aligned}
\mathbb{P}\left[\sup_{x \in \mathcal{X}} |\widehat{T}_{p+q}(x)| \leq \mathfrak{c}\right] &\leq \mathbb{P}\left[\sup_{x \in \mathcal{X}} |Z_{p+q}(x)| \leq \mathfrak{c} + \eta_{1,n}/a_n\right] + o(1) \\
&\leq \mathbb{P}\left[\sup_{x \in \mathcal{X}} |Z_{p+q}(x)| \leq c^0(1 - \alpha + \eta_{3,n}) + (\eta_{1,n} + \eta_{2,n})/a_n\right] + o(1) \\
&\leq \mathbb{P}\left[\sup_{x \in \mathcal{X}} |Z_{p+q}(x)| \leq c^0(1 - \alpha + \eta_{3,n})\right] + o(1) \to 1 - \alpha,
\end{aligned}
$$

where $c^0(1 - \alpha + \eta_{3,n})$ denotes the $(1 - \alpha + \eta_{3,n})$-quantile of $\sup_{x \in \mathcal{X}} |Z_{p+q}(x)|$, the first inequality holds by Theorem SA-3.3, the second by Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), and the third by Anti-Concentration Inequality in Chernozhukov, Chetverikov, and Kato (2014). The other side of the bound follows similarly.   □

SA-6.19   Proof of Theorem SA-3.6

Proof. Throughout this proof, we let $\eta_{1,n} = o(1)$, $\eta_{2,n} = o(1)$ and $\eta_{3,n} = o(1)$ be sequences of vanishing constants. Moreover, let A_n be a sequence of diverging constants such that $\sqrt{\log J} A_n \leq \sqrt{\frac{n}{J^{1+2v}}}$

(i): For Test (i), note that under $\ddot{\mathsf{H}}_0$,

$$
\sup_{x \in \mathcal{X}} |\ddot{T}_p(x)| \leq \sup_{x \in \mathcal{X}} \left|\frac{\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}\right| + \sup_{x \in \mathcal{X}} \left|\frac{\mu^{(v)}(x) - m^{(v)}(x; \widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|.
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\left[\sup_{x \in \mathcal{X}} |\ddot{T}_p(x)| > \mathfrak{c}\right] &\leq \mathbb{P}\left[\sup_{x \in \mathcal{X}} |\widehat{T}_p(x)| > \mathfrak{c} - \sup_{x \in \mathcal{X}} \left|\frac{\mu^{(v)}(x) - m^{(v)}(x; \widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|\right] \\
&\leq \mathbb{P}\left[\sup_{x \in \mathcal{X}} |Z_p(x)| > \mathfrak{c} - \eta_{1,n}/a_n - \sup_{x \in \mathcal{X}} \left|\frac{\mu^{(v)}(x) - m^{(v)}(x; \widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|\right] + o(1) \\
&\leq \mathbb{P}\left[\sup_{x \in \mathcal{X}} |Z_p(x)| > c^0(1 - \alpha - \eta_{3,n}) - (\eta_{1,n} + \eta_{2,n})/a_n - \right.
\end{aligned}
$$

44

$$\sup_{x\in\mathcal{X}}\left|\frac{\mu^{(v)}(x)-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|\right]+o(1)$$

$$\leq\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)|>c^0(1-\alpha-\eta_{3,n})\Big]+o(1)$$

$$=\alpha+o(1)$$

where $c^0(1-\alpha-\eta_{3,n})$ denotes the $(1-\alpha-\eta_{3,n})$-quantile of $\sup_{x\in\mathcal{X}}|Z_p(x)|$, the second inequality holds by Theorem SA-3.3, the third by Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), the fourth by the condition that $\sup_{x\in\mathcal{X}}\left|\frac{\mu^{(v)}(x)-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|=o_\mathbb{P}(\frac{1}{\sqrt{\log J}})$ and Anti-Concentration Inequality in Chernozhukov, Chetverikov, and Kato (2014). The other side of the bound follows similarly.

On the other hand, under $\ddot{\mathsf{H}}_A$,

$$\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\ddot{T}_p(x)|>\mathfrak{c}\Big]$$

$$=\mathbb{P}\Big[\sup_{x\in\mathcal{X}}\Big|\widehat{T}_p(x)+\frac{\mu^{(v)}(x)-m^{(v)}(x,\bar{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}+\frac{m^{(v)}(x,\bar{\boldsymbol{\theta}})-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big|>\mathfrak{c}\Big]$$

$$\geq\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)|\leq\sup_{x\in\mathcal{X}}\Big|\frac{\mu^{(v)}(x)-m^{(v)}(x,\bar{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}+\frac{m^{(v)}(x,\bar{\boldsymbol{\theta}})-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big|-\mathfrak{c}\Big]-o(1)$$

$$\geq\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)|\leq\sqrt{\log J}A_n-\eta_{1,n}/a_n\Big]-o(1)$$

$$\geq1-o(1).$$

where the third line holds by Lemma SA-2.4, Lemma SA-2.10, Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015) and $J^v\sqrt{J\log J/n}=o(1)$, the fourth by definition of A_n and Theorem SA-3.3, and the last by Concentration Inequality given in Lemma 12 of Chernozhukov, Lee, and Rosen (2013).

(ii): For Test (ii), note that under $\dot{\mathsf{H}}_0$,

$$\sup_{x\in\mathcal{X}}\dot{T}_p(x)\leq\sup_{x\in\mathcal{X}}\widehat{T}_p(x)+\sup_{x\in\mathcal{X}}\frac{|m^{(v)}(x,\bar{\boldsymbol{\theta}})-m^{(v)}(x,\widehat{\bar{\boldsymbol{\theta}}})|}{\sqrt{\widehat{\Omega}(x)/n}}.$$

Then,

$$\mathbb{P}\Big[\sup_{x\in\mathcal{X}}\dot{T}_p(x) > \mathfrak{c}\Big] \leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}\widehat{T}_p(x) > \mathfrak{c} - \sup_{x\in\mathcal{X}}\frac{|m^{(v)}(x,\bar{\boldsymbol{\theta}}) - m^{(v)}(x,\widehat{\bar{\boldsymbol{\theta}}})|}{\sqrt{\widehat{\Omega}(x)/n}}\Big]$$

$$\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} Z_p(x) > \mathfrak{c} - \eta_{1,n}/a_n\Big] + o(1)$$

$$\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} Z_p(x) > c^0(1 - \alpha - \eta_{3,n}) - (\eta_{1,n} + \eta_{2,n})/a_n\Big] + o(1)$$

$$\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} Z_p(x) > c^0(1 - \alpha - \eta_{3,n})\Big] + o(1)$$

$$= \alpha + o(1)$$

where $c^0(1 - \alpha - \eta_{3,n})$ denotes the $(1 - \alpha - \eta_{3,n})$-quantile of $\sup_{x\in\mathcal{X}} Z_p(x)$, the second line holds by Lemma SA-3.3, the third by Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), the fourth by Anti-Concentration Inequality in Chernozhukov, Chetverikov, and Kato (2014).

On the other hand, under $\dot{\mathsf{H}}_A$,

$$\mathbb{P}\Big[\sup_{x\in\mathcal{X}}\dot{T}_p(x) > \mathfrak{c}\Big] = \mathbb{P}\Big[\sup_{x\in\mathcal{X}}\Big(\widehat{T}_p(x) + \frac{\mu^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}\Big) > 0\Big]$$

$$\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)| < \sup_{x\in\mathcal{X}}\frac{\mu^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}, \; \sup_{x\in\mathcal{X}}\frac{\mu^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} > \mathfrak{c}\Big]$$

$$\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)| < \sup_{x\in\mathcal{X}}\frac{\mu^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}\Big] - o(1)$$

$$\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)| < \sqrt{\log J}A_n\Big] - o(1)$$

$$\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)| < \sqrt{\log J}A_n - \eta_{1,n}/a_n\Big] - o(1)$$

$$\geq 1 - o(1)$$

where the third line holds by Theorem SA-2.4, Lemma SA-2.10, Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), the assumption that $\sup_{x\in\mathcal{X}}|m^{(v)}(x,\widehat{\bar{\boldsymbol{\theta}}}) - m^{(v)}(x,\bar{\boldsymbol{\theta}})| = o_{\mathbb{P}}(1)$ and $J^v\sqrt{J\log J/n} = o(1)$, the fourth by definition of A_n, and the fifth by Lemma SA-3.3, and the last by Concentration Inequality given in Lemma 12 of Chernozhukov, Lee, and Rosen (2013). □

46

References

BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND K. KATO (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186(2), 345–366.

BHATIA, R. (2013): *Matrix Analysis*. Springer.

CALONICO, S., M. D. CATTANEO, AND R. TITIUNIK (2015): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110(512), 1753–1769.

CATTANEO, M. D., M. H. FARRELL, AND Y. FENG (2018): "Large Sample Properties of Partitioning-Based Estimators," arXiv:1804.04916.

CATTANEO, M. D., M. JANSSON, AND W. K. NEWEY (2018a): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 34(2), 277–301.

———— (2018b): "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113(523), 1350–1361.

CHERNOZHUKOV, V., D. CHETVERIKOV, AND K. KATO (2014): "Anti-Concentration and Honest Adaptive Confidence Bands," *Annals of Statistics*, 42(5), 1787–1818.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

DE BOOR, C. (1978): *A Practical Guide to Splines*. Springer-Verlag New York.

DEMKO, S. (1977): "Inverses of Band Matrices and Local Convergence of Spline Projections," *SIAM Journal on Numerical Analysis*, 14(4), 616–619.

GINÉ, E., AND R. NICKL (2016): *Mathematical Foundations of Infinite-Dimensional Statistical Models*, vol. 40. Cambridge University Press.

HUANG, J. Z. (2003): "Local Asymptotics for Polynomial Spline Regression," *Annals of Statistics*, 31(5), 1600–1635.

SAKHANENKO, A. (1991): "On the Accuracy of Normal Approximation in the Invariance Principle,"
Siberian Advances in Mathematics, 1, 58–91.

SCHUMAKER, L. (2007): *Spline Functions: Basic Theory.* Cambridge University Press.

VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes: With
Application to Statistics.* Springer.