On Binscatter
Supplemental Appendix*

Matias D. Cattaneo[†] Richard K. Crump[‡] Max H. Farrell[§] Yingjie Feng[¶]

August 3, 2021

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 linear and non-linear partitioning-based series estimation are obtained that may be of independent interest. Companion general-purpose software and replication files are available at https://nppackages.github.io/binsreg/.

*Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805 and SES-2019432. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

[†]Department of Operations Research and Financial Engineering, Princeton University.

[‡]Capital Markets Function, Federal Reserve Bank of New York.

[§]Booth School of Business, University of Chicago.

[¶]School of Economics and Management, Tsinghua University.

Contents

SA-1 Setup

Suppose that $(y_i, x_i, \mathbf{w}_i')$, $1 \le i \le n$, is a random sample where $y_i \in \mathcal{Y}$ is a scalar response variable, $x_i \in \mathcal{X}$ is a scalar covariate, and $\mathbf{w}_i \in \mathcal{W}$ is a vector of additional control variables of dimension d. For a general loss function $\rho(\cdot;\cdot)$ and a strictly monotonic transformation function $\eta(\cdot)$, define

$$(\mu_0(\cdot), \boldsymbol{\gamma}_0) = \operatorname*{arg\,min}_{\mu \in \mathcal{M}, \boldsymbol{\gamma} \in \mathbb{R}^d} \mathbb{E}[\rho(y_i; \eta(\mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}))], \tag{SA-1.1}$$

where $\mathcal{M}$ is a space of functions satisfying certain smoothness conditions to be specified later. The parameter of interest is the nonparametric component $\mu_0(\cdot)$ and transformations thereof.

This setup is general. For example, consider $\boldsymbol{\gamma}_0 = \mathbf{0}$. If $\rho(\cdot;\cdot)$ is a squared loss and $\eta(\cdot)$ is an identity function, $\mu_0(x)$ is the conditional expectation of y_i given $x_i = x$. Let $\mathbb{1}(\cdot)$ denote the indicator function. If $\rho(y;\eta) = (q - \mathbb{1}(y < \eta))(y - \eta)$ for some $0 < q < 1$ and $\eta(\cdot)$ is an identity function, then $\mu_0(x)$ is the qth conditional quantile of y_i given $x_i = x$. Introducing a transformation function $\eta(\cdot)$ is useful. For instance, it may accommodate the logistic regression for binary responses. When $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, the parametric and the nonparametric components are additively separable, and thus (SA-1.1) becomes a generalized partially linear model.

Binscatter estimators are 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

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

$x_{(i)}$ 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 plays 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.

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

$$\widehat{\mathbf{b}}(x) = \Big[\, \mathbb{1}_{\widehat{\mathcal{B}}_1}(x) \quad \mathbb{1}_{\widehat{\mathcal{B}}_2}(x) \quad \cdots \quad \mathbb{1}_{\widehat{\mathcal{B}}_J}(x) \,\Big]' \otimes \Big[\, 1 \quad x \quad \cdots \quad x^p \,\Big]',$$

where $\mathbb{1}_{\mathcal{A}}(x) = \mathbb{1}(x \in \mathcal{A})$ and $\otimes$ is the 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:

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

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.

To impose 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. 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 $\widehat{\mathbf{T}}_s$ 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 cumbersome. The proof of Lemma SA-1.2 describes the structure of $\widehat{\mathbf{T}}_s$ in more detail and provides an explicit representation for $\widehat{\mathbf{T}}_s$.

Given a choice of basis, we consider the following generalized binscatter estimator:

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta\big(\widehat{\mathbf{b}}_s(x_i)'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\gamma}\big)\Big), \qquad \text{(SA-1.2)}$$

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

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \quad \widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) := \underset{\boldsymbol{\beta}\in\mathbb{R}^{K_s}}{\arg\min} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta(\widehat{\mathbf{b}}_s(x_i)'\boldsymbol{\beta} + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}})\Big), \tag{SA-1.3}$$

$s \leq p$. The representation (SA-1.3) allows us to be more general and agnostic about the estimation of $\boldsymbol{\gamma}_0$, and also simplifies some of the proofs. More specifically, our theory requires only a sufficiently fast convergence rate of $\widehat{\boldsymbol{\gamma}}$ (see Assumption SA-GL(iii) below), which in general non-linear/non-differentiable cases can be justified in different ways: e.g., joint estimation, backfitting, profiling, split-sampling, etc. For the special case of semi-linear least squares regression (i.e., squared loss and identify link functions), the estimation procedure (SA-1.3) is simply a partial-out representation of (SA-1.2), and thus we directly verify the required convergence rate of $\widehat{\boldsymbol{\gamma}}$ to $\boldsymbol{\gamma}$ in Section SA-2. In the general case, the required rate of convergence can be verified on a case-by-case basis (e.g., logistic regression, quantile regression, etc.) depending on the specific structure of $\widehat{\boldsymbol{\gamma}}$.

We first impose some basic conditions on the data generating process.

Assumption SA-DGP (Data Generating Process)**.** *$\{(y_i, x_i, \mathbf{w}_i') : 1 \leq i \leq n\}$ is i.i.d satisfying (SA-1.1); x_i has absolutely continuous distribution function $F_X(\cdot)$ with a continuous (Lebesgue) density $f_X(\cdot)$ bounded away from zero on a compact support $\mathcal{X}$; $\mu_0(\cdot)$ is $(p+1)$-times continuously differentiable.*

SA-1.1 Summary of Technical Contributions

This supplemental appendix collects a comprehensive collection of new theoretical results for generalized non-linear partitioning-based estimators with semi-linear covariate-adjustment and random partitioning based on empirical quantiles. Canonical binscatter and all other binscatter methods discussed in the main paper are special cases of the generic setup considered herein, but several of the results below may be of independent interest in other contexts.

Section SA-1.3 presents new technical lemmas for random partitions based on empirical quantiles. Those results include general characterizations of the "regularity" of the random partitioning scheme (Lemmas SA-1.1 and SA-1.2) and of the associated random basis functions (Lemmas SA-1.3 and SA-1.4). These results give sharp control on the underlying random binning scheme of binscatter methods.

Section SA-2 studies large sample point estimation and distributional properties of the least squares estimator of $\mu_0(x)$ in (SA-1.2); that is, when $\rho(y;u) = (y-u)^2$ and $\eta(u) = u$. We study this case separately because its closed form solution allows for sharper results under weaker regularity conditions, and because least squares binscatter is arguably the most popular approach in empirical work. New results include (i) technical lemmas for Gram matrix (Lemma SA-2.1), asymptotic variance (Lemmas SA-2.2 and SA-2.3), approximation error (Lemma SA-2.4) and covariate adjustments (Lemma SA-2.5); (ii) stochastic linearization and uniform convergence rate (Theorem SA-2.1) and variance estimation (Theorem SA-2.2); (iii) pointwise distributional approximation (Theorem SA-2.3); (iv) conditional strong approximation (Theorem SA-2.4) and feasible implementation thereof (Theorem SA-2.5); (v) integrated mean squared error (IMSE) expansions (Theorem SA-2.6) and IMSE-optimal tuning parameter selection. All these results explicitly account for the random binning scheme. The most noteworthy novel result in this section is the conditional strong approximation, which circumvents a fundamental lack of uniformity of the random binning basis $\widehat{\mathbf{b}}_s^{(v)}(x)$, while still delivering a sufficiently fast uniform coupling requiring only $J^2/n \to 0$ (up to $\log(n)$ terms). This rate condition not only improves on previous results in the literature, but also allows for canonical binscatter (i.e., our results show that there exists a sequence $J \to \infty$ such that bias and variance are simultaneously controlled even when $p = s = 0$).

Section SA-3 studies large sample point estimation and distributional properties of the generalized non-linear estimator of $\mu_0(x)$ in (SA-1.2). New results include (i) technical lemmas for Gram matrix (Lemma SA-3.1), asymptotic variance (Lemmas SA-3.2 and SA-3.3), approximation error (Lemma SA-3.4) and uniform consistency (Lemma SA-3.5); (ii) stochastic linearization and uniform convergence rate (Theorem SA-3.1) and variance estimation (Theorem SA-3.2); (iii) pointwise distributional approximation (Theorem SA-3.3); (iv) conditional strong approximation (Theorem SA-3.4) and feasible implementation thereof (Theorem SA-3.5); (v) IMSE expansions (Theorem SA-3.6) and IMSE-optimal tuning parameter selection. All these results explicitly account for the random binning scheme. This section includes two most noteworthy novel results. First, a sharp Bahadur expansion for general non-linear semiparametric partitioning-based estimators with much faster rate of convergences than previously available in the literature is established: the result requires only $J^2/n \to 0$ (up to $\log(n)$ terms), while previous results required $J^4/n \to 0$ (up to $\log(n)$ terms) or worse. Therefore, our results allow for Canonical Binscatter and generalizations thereof, which would have been excluded by prior results (i.e., for previous technical results there was no sequence $J \to \infty$ such that bias and variance are simultaneously controlled). Furthermore, our new Bahadur expansion allows us to employ the same novel conditional strong approximation approach mentioned above, albeit with some important technical differences, to establish uniform inference results for generalized non-linear binscatter methods under essentially the same tuning parameter conditions imposed for least square binscatter methods. Second, a new Nagar-type IMSE expansion for generalized non-linear partitioning-based estimators with semi-linear covariate-adjustment and random partitioning based on empirical quantiles is established, which has no antecedent in the literature to the best of our knowledge.

Section SA-4 employs the technical results in Sections SA-2 and SA-3 to study estimation and inference for $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ and $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$, respectively. New results include valid confidence band estimators, consistent hypothesis tests about parametric specification and shape restrictions, and a detailed discussion on other parameters of interest, among other results. All these results explicitly account for the random binning scheme and semi-linear covariate-adjustment with random evaluation point. The most noteworthy novel result in this section is the proof technique to transform our strong approximation results (Theorems SA-2.4 and SA-3.4), and there feasible versions, into statements about the Kolmogorov distance for the suprema and related functionals of the t-statistics processes of interest. Our technical approach circumvents a fundamental lack of uniformity of the random binning basis $\widehat{\mathbf{b}}^{(v)}_{s}(x)$, while still delivering a sufficiently fast uniform coupling requiring only $J^2/n \to 0$ (up to $\log(n)$ terms). This proof technique can also be used to analyze other functionals such as the L_p distance, Kullback–Leibler divergence, and $\arg\max$ statistic.

Last but not least, at the end of each technical subsection, we include a remark labelled "Improvements over literature" that discusses more details of the technical improvements presented in that subsection and gives related references.

SA-1.2 Notation

For background definitions, see van der vaart and Wellner (1996), Bhatia (2013), Giné and Nickl (2016), and references therein.

Matrices and Norms. 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. $[\mathbf{A}]_{ij}$ denotes the (i,j)th entry of a generic matrix $\mathbf{A}$. We will use S^L to denote the unit circle in $\mathbb{R}^L$, i.e., $\|\mathbf{a}\| = 1$ for any $\mathbf{a} \in S^L$. For a real-valued function $g(\cdot)$ defined on a measure space $\mathcal{Z}$, let $\|g\|_{\mathbb{Q},2} := (\int_{\mathcal{Z}} |g|^2 d\mathbb{Q})^{1/2}$ be its L_2-norm with respect to the measure $\mathbb{Q}$. 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 $v \geq 0$.

Asymptotics. 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 $b_n \lesssim a_n$.

Empirical Process. We employ standard empirical process notation: $\mathbb{E}_n[g(\mathbf{v}_i)] = \frac{1}{n}\sum_{i=1}^n g(\mathbf{v}_i)$, and $\mathbb{G}_n[g(\mathbf{v}_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n (g(\mathbf{v}_i) - \mathbb{E}[g(\mathbf{v}_i)])$ for a sequence of random variables $\{\mathbf{v}_i\}_{i=1}^n$. 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 $N(\mathcal{G}, L_2(\mathbb{Q}), \varepsilon)$ is the minimal number of $L_2(\mathbb{Q})$-balls of radius ε needed to cover $\mathcal{G}$ for a measure $\mathbb{Q}$. The covering number of $\mathcal{G}$ relative to the envelope is denoted as $N(\mathcal{G}, L_2(\mathbb{Q}), \varepsilon\|\bar{G}\|_{\mathbb{Q},2})$.

Partitions. Given the random partition $\widehat{\Delta}$, we 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 $\Delta_0 = \{\mathcal{B}_1, \ldots, \mathcal{B}_J\}$ denotes the partition based on population quantiles, i.e.,

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

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_0. We sometimes write $\mathbf{b}_s(x;\Delta) = (b_{s,1}(x;\Delta), \ldots, b_{s,K_s}(x;\Delta))'$

to emphasize a binscatter basis is constructed based on a particular partition Δ. Clearly, $\widehat{\mathbf{b}}_s(x) = \mathbf{b}_s(x;\widehat{\Delta})$ and $\mathbf{b}_s(x) = \mathbf{b}_s(x;\Delta_0)$. For any given partition Δ, the *population* least squares projection of $\mu_0(\cdot)$ is given by $\mathbf{b}_s(\cdot;\Delta)'\boldsymbol{\beta}_0(\Delta)$ with

$$\boldsymbol{\beta}_0(\Delta) := \underset{\boldsymbol{\beta}\in\mathbb{R}^{K_s}}{\arg\min}\ \mathbb{E}[(\mu_0(x_i) - \mathbf{b}_s(x_i;\Delta)'\boldsymbol{\beta})^2]. \tag{SA-1.4}$$

Accordingly, given the random partition $\widehat{\Delta}$ and the nonrandom partition Δ_0, we have

$$\widehat{\boldsymbol{\beta}}_0 := \boldsymbol{\beta}_0(\widehat{\Delta}) := \underset{\boldsymbol{\beta}\in\mathbb{R}^{K_s}}{\arg\min}\ \mathbb{E}_{\widehat{\Delta}}[(\mu_0(x_i) - \mathbf{b}_s(x_i;\widehat{\Delta})'\boldsymbol{\beta})^2], \quad \text{and}$$

$$\boldsymbol{\beta}_0 := \boldsymbol{\beta}_0(\Delta_0) := \underset{\boldsymbol{\beta}\in\mathbb{R}^{K_s}}{\arg\min}\ \mathbb{E}[(\mu_0(x_i) - \mathbf{b}_s(x_i;\Delta_0)'\boldsymbol{\beta})^2].$$

The corresponding L_2 projection error is $r_{0,v}(x;\Delta) = \mu_0^{(v)}(x) - \mathbf{b}_s^{(v)}(x;\Delta)'\boldsymbol{\beta}_0(\Delta)$. We therefore define the approximation errors

$$\widehat{r}_{0,v}(x) := r_{0,v}(x;\widehat{\Delta}), \qquad \text{and} \qquad r_{0,v}(x) := r_{0,v}(x;\Delta_0).$$

For $v = 0$, we write $\widehat{r}_0(x) := \widehat{r}_{0,0}(x)$ and $r_0(x) := r_{0,0}(x)$

Other. Let $\mathbf{X} = [x_1,\ldots,x_n]'$, $\mathbf{W} = [\mathbf{w}_1,\cdots,\mathbf{w}_n]'$, and $\mathbf{D} = [(y_i, x_i, \mathbf{w}_i')' : i = 1,2,\ldots,n]$. For two random variables X and Y, $X =_d Y$ implies that they have the same probability distribution. $\lceil z\rceil$ outputs the smallest integer no less than z and $a \wedge b = \min\{a,b\}$. "w.p.a. 1" means "with probability approaching one".

SA-1.3 Preliminary Lemmas

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

$$\Pi = \Big\{\Delta : \frac{\max_{1\le j\le J} h_j(\Delta)}{\min_{1\le j\le J} h_j(\Delta)} \le \frac{3\bar{f}_X}{\underline{f}_X}\Big\}. \tag{SA-1.5}$$

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-1.1 (Quasi-Uniformity of Quantile-Spaced Partitions)**.** *Suppose that Assumption SA-DGP holds. If* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then (i)* $\max_{1\le j\le J}|\hat{h}_j - h_j| \lesssim_{\mathbb{P}} J^{-1}\Big(\frac{J\log J}{n}\Big)^{1/2}$*, and (ii)* $\widehat{\Delta}\in\Pi$ *w.p.a.* 1*.*

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 empirical-quantile-based partition $\widehat{\Delta}$. The next lemma describes its key features. We let $\mathbf{T}_s := \mathbf{T}_s(\Delta_0)$ be 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-1.2 (Transformation Matrix)**.** *Suppose that Assumption SA-DGP 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* $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$*,* $\|\widehat{\mathbf{T}}_s\| \lesssim_{\mathbb{P}} 1$*,* $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\|_\infty \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n}\Big)^{1/2}$*, and* $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n}\Big)^{1/2}$*.*

The following lemma provides some simple bounds on the basis.

Lemma SA-1.3 (Local Basis)**.** *Suppose that Assumption SA-DGP holds. Then,* $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_s^{(v)}(x)\|_0 \le (p+1)^2$*. If, in addition,* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then* $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_s^{(v)}(x)\| \lesssim_{\mathbb{P}} J^{\frac{1}{2}+v}$*.*

The following lemma characterizes the approximation error $\widehat{r}_{0,v}(x)$ in terms of the sup norm.

Lemma SA-1.4 (Approximation Error)**.** *Suppose that Assumption SA-DGP holds. If* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\sup_{x\in\mathcal{X}}|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0 - \mu_0^{(v)}(x)| \lesssim_{\mathbb{P}} J^{-p-1+v}.$$

Remark SA-1.1 (Improvements over literature)**.** Lemmas SA-1.1-SA-1.4 show some basic characteristics of the binscatter basis, which are used in the subsequent main analysis. Compared with other studies of splines (see, e.g., Shen, Wolfe, and Zhou, 1998; Huang, 2003; Schumaker, 2007), we formally take into account the randomness of the partition formed by empirical quantiles. ⌟

SA-2 Least Squares Binscatter

In this section, we consider a squared loss function combined with an identity link: $\rho(y;\eta) = (y-\eta)^2$ and $\eta(\theta) = \theta$. Our setup corresponds to the partially linear regression model in the semiparametrics literature:

$$y_i = \mu_0(x_i) + \mathbf{w}_i'\gamma_0 + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|x_i, \mathbf{w}_i] = 0. \tag{SA-2.1}$$

In the main paper, we define the following parameter of interest:

$$\Upsilon_{\mathbf{w}}^{(v)}(x) = \frac{\partial^v}{\partial x^v}\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$$

for some evaluation point $\mathbf{w}$. Given the assumption $\mathbb{E}[\epsilon_i|x_i, \mathbf{w}_i] = 0$, $\Upsilon_{\mathbf{w}}^{(0)}(x) = \mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0$ and $\Upsilon_{\mathbf{w}}^{(v)}(x) = \mu_0^{(v)}(x)$ for $v > 0$.

It is well known that the estimator given in (SA-1.2) admits the following "backfitting" expression, which will be convenient for our theoretical analysis:

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

where $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\mathbf{B} = (\widehat{\mathbf{b}}_s(x_1), \ldots, \widehat{\mathbf{b}}_s(x_n))'$, $\mathbf{M}_{\mathbf{B}} = \mathbf{I}_n - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$.

Given an estimator $\widehat{\mathbf{w}}$ of the evaluation point $\mathbf{w}$, we have the following estimator of $\Upsilon_{\mathbf{w}}^{(v)}(x)$:

$$\widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x) = \begin{cases} \widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\gamma} & \text{if } v = 0 \\ \widehat{\mu}^{(v)}(x) & \text{if } v \geq 1 \end{cases}, \qquad \widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}.$$

In this section, we will focus on the nonparametric component $\mu_0(\cdot)$, i.e., $\Upsilon_{\mathbf{0}}(x)$ and the corresponding estimator $\widehat{\mu}^{(v)}(x)$. The properties of $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x)$ can be derived similarly, given the fast rate of convergence of the parametric component given in Lemma SA-2.5 below. We will revisit this issue later in Section SA-4.4.

Now, we introduce the following quantities:

$$\begin{aligned}
&\widehat{\mathbf{Q}} := \widehat{\mathbf{Q}}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'], \quad \mathbf{Q}_0 := \mathbf{Q}(\Delta_0) := \mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'],\\
&\widehat{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{\Sigma}}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\epsilon}_i^2], \quad \bar{\boldsymbol{\Sigma}} := \bar{\boldsymbol{\Sigma}}(\widehat{\Delta}) := \mathbb{E}_n[\mathbb{E}[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i^2|\mathbf{X}]],\\
&\boldsymbol{\Sigma}_0 := \boldsymbol{\Sigma}(\Delta_0) := \mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\epsilon_i^2],\\
&\widehat{\Omega}(x) := \widehat{\Omega}(x; \widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x),\\
&\bar{\Omega}(x) := \bar{\Omega}(x; \widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\bar{\boldsymbol{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x), \quad \text{and}
\end{aligned}$$

$$\Omega(x) := \Omega(x;\widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0\mathbf{Q}_0^{-1}\widehat{\mathbf{b}}_s^{(v)}(x),$$

where $\widehat{\epsilon}_i = y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}$. 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 dependence on v of $\bar{\Omega}(x)$ and $\Omega(x)$ is omitted for simplicity.

We impose an additional condition for the least squares case.

Assumption SA-LS (Squared Loss)**.**

(i) $\mathbb{E}[\epsilon_i|x_i,\mathbf{w}_i] = 0$, $\sigma^2(x) := \mathbb{E}[\epsilon_i^2|x_i = x]$ *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* $\nu > 2$.

(ii) $\mathbb{E}[\mathbf{w}_i|x_i = x]$ *is* ς*-times continuously differentiable for some* $\varsigma \geq 1$*;* $\sup_{x\in\mathcal{X}}\mathbb{E}[\|\mathbf{w}_i\|^\nu|x_i = x] \lesssim 1$*;* $\max_{1\leq i\leq n}\mathbb{E}[\|\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i|x_i]\|^4|x_i] \lesssim_{\mathbb{P}} 1$, $\min_{1\leq i\leq n}\lambda_{\min}(\mathbb{E}[(\mathbf{w}_i-\mathbb{E}[\mathbf{w}_i|x_i])(\mathbf{w}_i-\mathbb{E}[\mathbf{w}_i|x_i])'|x_i]) \gtrsim_{\mathbb{P}} 1$*, and* $\max_{1\leq i\leq n}\mathbb{E}[\epsilon_i^2|\mathbf{w}_i,x_i] \lesssim_{\mathbb{P}} 1$.

Part (i) imposes some mild moment conditions on the error term which are commonly used in the nonparametric series estimation literature. Part (ii) includes a set of conditions similar to those used in Cattaneo, Jansson, and Newey (2018a,b) to analyze the semiparametric partially linear regression model. They ensure the negligibility of the estimation error of $\widehat{\boldsymbol{\gamma}}$.

SA-2.1 Technical Lemmas

This section collects a set of technical lemmas, which are key ingredients of our main theorems. The first lemma characterizes the local basis $\widehat{\mathbf{b}}_s(x)$ and the associated Gram matrix.

Lemma SA-2.1 (Gram)**.** *Suppose that Assumption SA-DGP holds. Then,* $1 \lesssim \lambda_{\min}(\mathbf{Q}_0) \leq \lambda_{\max}(\mathbf{Q}_0) \lesssim 1$*. If, in addition,* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\|\widehat{\mathbf{Q}} - \mathbf{Q}_0\| \lesssim_{\mathbb{P}} \left(\frac{J\log J}{n}\right)^{1/2}, \quad \|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1, \quad \textit{and} \quad \|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\|_\infty \lesssim_{\mathbb{P}} \left(\frac{J\log J}{n}\right)^{1/2}.$$

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

Lemma SA-2.2 (Asymptotic Variance)**.** *Suppose that Assumptions SA-DGP and SA-LS(i) hold. If* $\frac{J \log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

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

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-2.3 (Uniform Convergence: Variance)**.** *Suppose that Assumptions SA-DGP and SA-LS(i) hold. If* $\frac{J^{\frac{\nu}{\nu-2}} \log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\sup_{x \in \mathcal{X}} \left| \widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \mathbb{E}_n[\mathbf{b}_s(x_i)\epsilon_i] \right| \lesssim_{\mathbb{P}} J^v \Big(\frac{J \log J}{n} \Big)^{1/2}.$$

As explained before, $\widehat{r}_0(x)$ is understood as L_2 approximation error of least squares estimators for $\mu_0(x)$. The next two lemmas establish bounds on $\widehat{r}_0(x)$ and its projection onto the space spanned by $\widehat{\mathbf{b}}_s(x)$ in terms of sup-norm.

Lemma SA-2.4 (Projection of Approximation Error)**.** *Under Assumption SA-DGP, if* $\frac{J \log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\sup_{x \in \mathcal{X}} \left| \widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_0(x_i)] \right| \lesssim_{\mathbb{P}} J^{-p-1+v} \Big(\frac{J \log J}{n} \Big)^{1/2}.$$

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 gives conditions under which the estimation of $\boldsymbol{\gamma}_0$ does not impact the asymptotic inference on the nonparametric component.

Lemma SA-2.5 (Covariate Adjustment)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. If* $\frac{J \log J}{n} = o(1)$, $\frac{a_n}{\sqrt{J}} = o(1)$*, and* $a_n \sqrt{n} J^{-p-(\varsigma \wedge (p+1))-\frac{3}{2}} = o(1)$*, then*

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n}), \quad \textit{and} \quad \|\widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i']\|_\infty \lesssim_{\mathbb{P}} J^v \ \textit{for each } x \in \mathcal{X}.$$

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

Remark SA-2.1 (Improvements over literature)**.** The results in this subsection give novel rates of approximations for semi-linear partitioning-based estimators with random partitions. Compared to standard semi-linear regression results, our results provide sharper approximation rates due to the specific binscatter basis, and also formally take into account the randomness of the partition formed by empirical quantiles. See Cattaneo, Jansson, and Newey (2018a,b), and reference therein, for related literature. ⌟

SA-2.2 Bahadur Representation

Theorem SA-2.1 (Bahadur Representation). *Suppose that Assumptions SA-DGP and SA-LS hold. If* $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} \lesssim 1$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\sup_{x\in\mathcal{X}}\Big|\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x) - \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i]\Big| \lesssim_{\mathbb{P}} J^v\Big(\frac{1}{\sqrt{n}} + J^{-p-1-(\varsigma\wedge(p+1))} + J^{-p-1}\Big).$$

An immediate corollary of Theorem SA-2.1 is the uniform convergence of $\widehat{\mu}(\cdot)$.

Corollary SA-2.1 (Uniform Convergence). *Suppose that Assumptions SA-DGP and SA-LS hold. 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*

$$\sup_{x\in\mathcal{X}}|\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)| \lesssim_{\mathbb{P}} J^v\Big(\frac{J\log J}{n}\Big)^{1/2} + J^{-p-1+v}.$$

Based on the above facts, we can also show that the proposed variance estimator is consistent.

Theorem SA-2.2 (Variance Estimate). *Suppose that Assumptions SA-DGP and SA-LS hold. 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*

$$\Big\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\Big\| \lesssim_{\mathbb{P}} J^{-p-1} + \Big(\frac{J\log J}{n^{1-\frac{2}{\nu}}}\Big)^{1/2}, \quad \text{and} \quad \sup_{x\in\mathcal{X}}\Big|\widehat{\Omega}(x) - \Omega(x)\Big| \lesssim_{\mathbb{P}} J^{1+2v}\Big(J^{-p-1} + \Big(\frac{J\log J}{n^{1-\frac{2}{\nu}}}\Big)^{1/2}\Big).$$

Remark SA-2.2 (Improvements over literature). The results in this subsection improve on the linear series estimation literature (Belloni, Chernozhukov, Chetverikov, and Kato, 2015; Cattaneo, Farrell, and Feng, 2020) by formally taking into account the randomness of the partition formed by empirical quantiles, and by accounting for the semi-linear regression estimation structure. The final approximation rate in the Bahadur-type (linear) representation is sharp for the binscatter basis (with or without random binning). ⌟

SA-2.3 Pointwise Inference

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

$$\widehat{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}.$$

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-2.3 (Pointwise Asymptotic Distribution)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. 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*

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

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 of the main paper.

Corollary SA-2.2 (Confidence Intervals)**.** *For given* p*, suppose that the conditions in Theorem SA-2.3 hold for* $\nu = 4$*, and further assume that* $\mu_0(x)$ *and* $\mathbb{E}[\mathbf{w}_i|x_i = x]$ *are* $(p + q + 1)$*-times continuously differentiable for some* $q \geq 1$*. If* $J = J_{\texttt{IMSE}}$ *for* $J_{\texttt{IMSE}}$ *defined in Section SA-2.5 and* $\mathfrak{c} = \Phi^{-1}(1 - \alpha/2)$*, then*

$$\mathbb{P}\Big[\mu_0^{(v)}(x) \in \widehat{I}_{p+q}(x)\Big] = 1 - \alpha + o(1), \quad \textit{for all } x \in \mathcal{X}.$$

Remark SA-2.3 (Improvements over literature)**.** The results in this subsection improve upon Cattaneo, Farrell, and Feng (2020, Section 5), the best results available for partitioning-based estimation, by formally taking into account the randomness of the partition formed by empirical quantiles, and by accounting for the semi-linear regression estimation structure. ⌟

SA-2.4 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-2.4 (Strong Approximation). *Suppose that Assumptions SA-DGP and SA-LS hold. If*

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

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 $\xi > 0$,

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

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 $\mathbf{D}$.

Theorem SA-2.5 (Plug-in Approximation). *Suppose that the conditions in Theorem SA-2.4 hold. Then, on a properly enrich probability space there exists a K_s-dimensional standard normal random vector $\mathbf{N}_{K_s}$ independent of $\mathbf{D}$ such that for any $\xi > 0$,*

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

Remark SA-2.4 (Improvements over literature). Theorems SA-2.4 and SA-2.5 offer a new easy-to-implement approach to conduct binscatter-based uniform inference. We formally take into account the randomness of the empirical-quantile-based partition and approximate the *whole* t-statistic process by a (conditional) Gaussian process under seemingly minimal rate conditions. In fact, it can be shown that when $a_n = \sqrt{\log n}$ and a subexponential moment restriction holds for the error term, it suffices that $J/n = o(1)$, up to $\log n$ terms. In contrast, a strong approximation of the t-statistic process for general series estimators was obtained based on Yurinskii coupling in Belloni, Chernozhukov, Chetverikov, and Kato (2015), which requires $J^5/n = o(1)$, up to $\log n$ terms.

Alternatively, a strong approximation of the *supremum* of the t-statistic process can be obtained under weaker rate restrictions. For instance, Chernozhukov, Chetverikov, and Kato (2014a) requires $J/n^{1-2/\nu} = o(1)$, up to $\log n$ terms, a result that applies exclusively to the suprema of the stochastic process. $\lrcorner$

SA-2.5 Integrated Mean Squared Error

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

Theorem SA-2.6 (IMSE)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. Let $\omega(x)$ be a continuous weighting function over $\mathcal{X}$ bounded away from zero. If $\sqrt{n}J^{-p-(\varsigma\wedge(p+1))-\frac{3}{2}} = o(1)$ and $\frac{J\log J}{n} = o(1)$, then*

$$\begin{aligned}&\int_{\mathcal{X}} \mathbb{E}\Big[\Big(\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)\Big)^2\Big|\mathbf{X},\mathbf{W}\Big]\omega(x)dx\\&= \frac{J^{1+2v}}{n}\mathscr{V}_n(p,s,v) + J^{-2(p+1-v)}\mathscr{B}_n(p,s,v) + o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n} + J^{-2(p+1-v)}\Big),\end{aligned}$$

where

$$\begin{aligned}\mathscr{V}_n(p,s,v) &:= J^{-(1+2v)}\operatorname{trace}\Big(\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\Big) \asymp 1,\\ \mathscr{B}_n(p,s,v) &:= J^{2p+2-2v}\int_{\mathcal{X}}\Big(\mathbf{b}_s^{(v)}(x)'\boldsymbol{\beta}_0 - \mu_0^{(v)}(x)\Big)^2\omega(x)dx \lesssim 1.\end{aligned}$$

As a consequence, the IMSE-optimal choice of J is $J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$ whenever $\mathscr{B}_n(p,s,v) \gtrsim 1$. See Remark SA-2.5 below for discussion of the lower bound on $\mathscr{B}_n(p,s,v)$. More precisely, if $\mathscr{B}_n = \mathscr{B}(p,s,v) + o(1)$ and $\mathscr{V}_n(p,s,v) = \mathscr{V}(p,s,v) + o(1)$, then we can take

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

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-5. 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 $\boldsymbol{\varphi}(z) = (1, z, \dots, z^p)'$.

Corollary SA-2.3. *Under the assumptions in Theorem SA-2.6, $\mathscr{V}_n(p,0,v) = \mathscr{V}(p,0,v) + o(1)$ and $\mathscr{B}_n(p,0,v) = \mathscr{B}(p,0,v) + o(1)$ where*

$$\mathscr{V}(p,0,v) := \operatorname{trace}\Big\{\Big(\int_0^1 \boldsymbol{\varphi}(z)\boldsymbol{\varphi}(z)'dz\Big)^{-1}\int_0^1 \boldsymbol{\varphi}^{(v)}(z)\boldsymbol{\varphi}^{(v)}(z)'dz\Big\}\int_{\mathcal{X}} \sigma^2(x) f_X(x)^{2v}\omega(x)dx,$$
$$\mathscr{B}(p,0,v) := \frac{\int_0^1[\mathscr{B}_{p+1-v}(z)]^2dz}{((p+1-v)!)^2}\int_{\mathcal{X}}\frac{[\mu_0^{(p+1)}(x)]^2}{f_X(x)^{2p+2-2v}}\omega(x)dx.$$

Remark SA-2.5. The above corollary implies that the bias constant $\mathscr{B}(p,0,v)$ is nonzero unless $\mu_0^{(p+1)}(x)$ is zero almost everywhere on $\mathcal{X}$. For other $s>0$, notice that $\mathbf{b}_s^{(v)}(x)'\boldsymbol{\beta}_0$ can be viewed as an approximation of $\mu_0^{(v)}(x)$ 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}. $\mathbf{b}_s^{(v)}(x)'\boldsymbol{\beta}_0$, 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 $f_X(x)$ 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-2.6. In all analysis in this paper, we simply exclude this special case by assuming the leading bias is non-degenerate, and thus $J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$. ⌟

Finally, using Lemma SA-2.5, we have the following corollary about the IMSE expansion of the estimator $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ of the general parameter $\Upsilon^{(v)}_{\mathbf{w}}(x)$, which corresponds to Theorem 1 in the main paper.

Corollary SA-2.4. *Suppose that the assumptions in Theorem SA-2.6 hold and $\widehat{\mathbf{w}}$ is either non-random or generated based on $\mathbf{W}$ such that $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$. Then,*

$$\int_{\mathcal{X}} \mathbb{E}\Big[\Big(\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - \Upsilon^{(v)}_{\mathbf{w}}(x)\Big)^2\Big|\mathbf{X},\mathbf{W}\Big]\omega(x)dx$$
$$= \frac{J^{1+2v}}{n}\mathscr{V}_n(p,s,v) + J^{-2(p+1-v)}\mathscr{B}_n(p,s,v) + o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n} + J^{-2(p+1-v)}\Big),$$

where $\mathscr{V}_n(p,s,v)$ and $\mathscr{B}_n(p,s,v)$ are defined as in Theorem SA-2.6.

Remark SA-2.6 (Improvements over literature)**.** The results in this subsection improve upon

Cattaneo, Farrell, and Feng (2020, Section 4), the best results available for partitioning-based estimation, by formally taking into account the randomness of the partition formed by empirical quantiles, and by accounting for the semi-linear regression estimation structure. ⌟

SA-3 Generalized Non-Linear Binscatter

In this section, we consider a general loss function $\rho(\cdot;\cdot)$ associated with a general (inverse) link function $\eta(\cdot)$. We also assume that a preliminary estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}_0$ exists and impose high-level conditions on $\widehat{\boldsymbol{\gamma}}$ directly. Such estimators and their properties can be usually found in the semi-parametrics literature. To simplify notation, we write

$$\begin{aligned}
&\eta_i = \eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0), && \widehat{\eta}_i = \eta(\widehat{\mu}(x_i) + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}),\\
&\eta_{i,1} = \eta^{(1)}(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0), && \widehat{\eta}_{i,1} = \eta^{(1)}(\widehat{\mu}(x_i) + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}),\\
&\widehat{\mu}(x_i) = \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}, \quad \epsilon_i = y_i - \eta_i, && \text{and} \;\; \widehat{\epsilon}_i = y_i - \widehat{\eta}_i.
\end{aligned}$$

In the main paper, we define the following parameter of interest

$$\vartheta^{(v)}_{\mathbf{w}}(x) = \frac{\partial^v}{\partial x^v}\eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)$$

for some evaluation point $\mathbf{w}$. Accordingly, we can estimate it by

$$\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x) = \frac{\partial^v}{\partial x^v}\eta(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}})$$

for some estimate $\widehat{\mathbf{w}}$ of the evaluation point $\mathbf{w}$. In this section, we focus on the estimator of the nonparametric component $\widehat{\mu}(x)$. Its properties are the building blocks of the analysis of the estimator $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$. We will revisit the estimation and inference of the more general parameter $\vartheta^{(v)}_{\mathbf{w}}(x)$ later in Section SA-4.4.

Now, we impose the following conditions for this general case:

Assumption SA-GL (General Loss)**.**

(i) *$\rho(y;\eta)$ is absolutely continuous with respect to $\eta \in \mathbb{R}$, which admits a piecewise Lipschitz derivative $\psi(y;\eta) \equiv \psi(y-\eta)$ that has at most m discontinuity points for some finite $m \in \mathbb{Z}_+$;*

$\eta(\cdot)$ is strictly monotonic and two-times continuously differentiable; $\rho(y;\eta(\theta))$ is convex with respect to θ.

(ii) *$\mathbb{E}[\psi(\epsilon_i)|x_i,\mathbf{w}_i]=0$, $\sigma^2(x,\mathbf{w}):=\mathbb{E}[\psi(\epsilon_i)^2|x_i=x,\mathbf{w}_i=\mathbf{w}]$ is continuous and bounded away from zero uniformly over $x\in\mathcal{X}$ and $\mathbf{w}\in\mathcal{W}$, and for some $\nu>2$, $\sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}}\mathbb{E}[|\psi(\epsilon_i)|^\nu|x_i=x,\mathbf{w}_i=\mathbf{w}]\lesssim 1$.*

(iii) *The preliminary estimator $\widehat{\boldsymbol{\gamma}}$ satisfies that $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|\lesssim_{\mathbb{P}}\mathfrak{r}_\gamma$ for $\mathfrak{r}_\gamma=o(\sqrt{J/n}+J^{-p-1})$.*

(iv) *The conditional density of y_i given x_i and $\mathbf{w}_i$, denoted by $f_{Y|XW}(y|x,\mathbf{w})$, satisfies that $\sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}}\sup_{y\in\mathcal{Y}_{x\mathbf{w}}}f_{Y|XW}(y|x,\mathbf{w})\lesssim 1$ where $\mathcal{Y}_{x\mathbf{w}}$ is the support of the conditional density of y_i given $x_i=x$ and $\mathbf{w}_i=\mathbf{w}$; the support $\mathcal{W}$ of $\mathbf{w}_i$ is bounded; $\sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}}|\eta^{(1)}(\mu_0(x)+\mathbf{w}'\boldsymbol{\gamma}_0)|\lesssim 1$.*

(v) *$\Psi(x,\mathbf{w};\eta):=\mathbb{E}[\psi(y_i;\eta)|x_i=x,\mathbf{w}_i=\mathbf{w}]$ is twice continuously differentiable with respect to η; $\inf_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}}\varkappa(x,\mathbf{w})\geq C$ for some constant $C>0$ and $\mathbb{E}[\varkappa(x_i,\mathbf{w}_i)|x_i=x]$ is continuous on $\mathcal{X}$ where $\varkappa(x,\mathbf{w}):=\Psi_1(x,\mathbf{w};\eta(\mu_0(x)+\mathbf{w}'\boldsymbol{\gamma}_0))(\eta^{(1)}(\mu_0(x)+\mathbf{w}'\boldsymbol{\gamma}_0))^2$ and $\Psi_1(x,\mathbf{w};\eta):=\frac{\partial}{\partial\eta}\Psi(x,\mathbf{w};\eta)$.*

(vi) *For some estimator $\widehat{\Psi}_1$ of Ψ_1, $\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(\widehat{\varkappa}(x_i,\mathbf{w}_i)-\varkappa(x_i,\mathbf{w}_i))\|\lesssim_{\mathbb{P}}J^{-p-1}+\Big(\frac{J\log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2}$ where $\widehat{\varkappa}(x_i,\mathbf{w}_i)=\widehat{\Psi}_1(x_i,\mathbf{w}_i;\widehat{\eta}_i)\widehat{\eta}_{i,1}^2$.*

Part (vi) is a high-level condition that ensures we have a valid feasible estimator of the population Gram matrix, i.e., $\mathbf{Q}_0$ defined below. The rate of convergence of $\widehat{\eta}_{i,1}$ can be deduced from Corollary SA-3.1 below. Thus, part (vi) can be viewed as a requirement on $\widehat{\Psi}_1$ only. Note that $\widehat{\Psi}_1$ does not have to be consistent for Ψ_1 in a pointwise or uniform sense. It suffices that the estimator $\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\varkappa}(x_i,\mathbf{w}_i)]$ based on $\widehat{\Psi}_1$ as a whole is consistent. See Section SA-5 for several examples of the estimator $\widehat{\Psi}_1$.

We re-define several quantities introduced before which accommodates the more general loss:

$$\begin{aligned}
\widehat{\mathbf{Q}}&:=\widehat{\mathbf{Q}}(\widehat{\Delta}):=\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\Psi}_1(x_i,\mathbf{w}_i;\widehat{\eta}_i)\widehat{\eta}_{i,1}^2],\\
\bar{\mathbf{Q}}&:=\bar{\mathbf{Q}}(\widehat{\Delta}):=\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Psi_1(x_i,\mathbf{w}_i;\eta_i)\eta_{i,1}^2],\\
\mathbf{Q}_0&:=\mathbf{Q}(\Delta_0):=\mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\Psi_1(x_i,\mathbf{w}_i;\eta_i)\eta_{i,1}^2],
\end{aligned}$$

$$
\begin{aligned}
&\widehat{\mathbf{\Sigma}} := \widehat{\mathbf{\Sigma}}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\psi(\widehat{\epsilon}_i)^2\widehat{\eta}_{i,1}^2],\\
&\bar{\mathbf{\Sigma}} := \bar{\mathbf{\Sigma}}(\widehat{\Delta}) := \mathbb{E}_n\Big[\mathbb{E}\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\psi(\epsilon_i)^2\eta_{i,1}^2\Big|\mathbf{X},\mathbf{W}\Big]\Big],\\
&\mathbf{\Sigma}_0 := \mathbf{\Sigma}(\Delta_0) := \mathbb{E}\Big[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\psi(\epsilon_i)^2\eta_{i,1}^2\Big],\\
&\widehat{\Omega}(x) := \widehat{\Omega}(x;\widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x),\\
&\bar{\Omega}(x) := \bar{\Omega}(x;\widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\bar{\mathbf{\Sigma}}\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x), \quad \text{and}\\
&\Omega(x) := \Omega(x;\widehat{\Delta}) := \widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0\mathbf{Q}_0^{-1}\widehat{\mathbf{b}}_s^{(v)}(x).
\end{aligned}
$$

SA-3.1 Technical Lemmas

Lemma SA-3.1 (Gram)**.** *Suppose that Assumptions SA-DGP and SA-GL hold. Then,* $1 \lesssim \lambda_{\min}(\mathbf{Q}_0) \leq \lambda_{\max}(\mathbf{Q}_0) \lesssim 1$*. If, in addition,* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$
\begin{aligned}
&\|\bar{\mathbf{Q}} - \mathbf{Q}_0\| \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n}\Big)^{1/2}, \quad 1 \lesssim \lambda_{\min}(\bar{\mathbf{Q}}) \leq \lambda_{\max}(\bar{\mathbf{Q}}) \lesssim 1, \quad [\bar{\mathbf{Q}}^{-1}]_{ij} \lesssim \varrho^{|i-j|} \quad \textit{w.p.a. } 1,\\
&\|\bar{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1, \quad \textit{and} \quad \|\bar{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\|_\infty \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n}\Big)^{1/2},
\end{aligned}
$$

where $\varrho \in (0,1)$ *is some absolute constant.*

The next lemma shows that the limiting variance is bounded from above and below.

Lemma SA-3.2 (Asymptotic Variance)**.** *Suppose that Assumptions SA-DGP and SA-GL hold. If* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

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

The next lemma gives a bound on the variance component of the general binscatter estimator.

Lemma SA-3.3 (Uniform Convergence: Variance)**.** *Suppose that Assumptions SA-DGP and SA-GL hold. If* $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$
\sup_{x\in\mathcal{X}} \Big|\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)]\Big| \lesssim_{\mathbb{P}} J^v\Big(\frac{J\log J}{n}\Big)^{1/2}.
$$

Lemma SA-3.4 (Projection of Approximation Error)**.** *Under Assumptions SA-DGP and SA-GL,* *if* $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\sup_{x\in\mathcal{X}}\left|\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\mathbb{E}_n\left[\widehat{\mathbf{b}}_s(x_i)\left(\eta_{i,1}\psi(\epsilon_i)-\eta^{(1)}(\widehat{\mathbf{b}}(x_i)'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)\psi(y_i;\eta(\widehat{\mathbf{b}}(x_i)'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\right)\right]\right|$$
$$\lesssim_{\mathbb{P}} J^{-p-1+v}+J^{\frac{2v-p-1}{2}}\Big(\frac{J\log J}{n}\Big)^{1/2}+\frac{J^{1+v}\log J}{n}.$$

Lemma SA-3.5 (Uniform Consistency)**.** *Under Assumptions SA-DGP and SA-GL, if* $\frac{J^2\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_0\|_\infty = o_{\mathbb{P}}(J^{-1/2}).$$

Remark SA-3.1. We conjecture that the rate restriction $\frac{J^2\log J}{n} = o(1)$ is stronger than needed. In fact, for piecewise polynomials (i.e., $s=0$), we can show that $\frac{J\log J}{n} = o(1)$ suffices to establish the uniform consistency of $\widehat{\boldsymbol{\beta}}$. In that special case ($s=0$), the condition $\frac{J^2\log n}{n} = o(1)$ in all theorems below can be dropped.

Our result holds without imposing any smoothness restrictions on the estimation space. Specifically, the estimation procedure (SA-1.3) searches for solutions in $\mathbb{R}^{K_s}$, leading to an estimation space $\{\widehat{\mathbf{b}}_s(x)'\boldsymbol{\beta}:\boldsymbol{\beta}\in\mathbb{R}^{K_s}\}$. In contrast, many studies of series (or sieve) methods restrict the functions in the estimation space to satisfy certain smoothness conditions, e.g., Lipschitz continuity, to derive the uniform consistency. See, for example, Chernozhukov, Imbens, and Newey (2007). ⌟

Finally, before we present our main results, the following maximal inequality is useful.

Lemma SA-3.6 (Maximal Inequality)**.** *Let* $Z_1,\cdots,Z_n$ *be independent but not necessarily identically distributed random variables taking values in a measurable space* $(\mathcal{S};\mathscr{S})$*. Denote the distribution of* Z_i *by* $\mathbb{P}_i$*, and let* $\bar{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n\mathbb{P}_i$*. Let* $\mathcal{F}$ *be a class of Borel measurable functions from* $\mathcal{S}$ *to* $\mathbb{R}$ *which is pointwise measurable. Let* $\bar{F}$ *be a measurable envelope function for* $\mathcal{F}$*. Suppose that* $\|\bar{F}\|_{L_2(\bar{\mathbb{P}})}<\infty$*. Let* $\bar{\sigma}>0$ *satisfy* $\sup_{f\in\mathcal{F}}\|f\|_{L_2(\bar{\mathbb{P}})}\leq\bar{\sigma}\leq\|\bar{F}\|_{L_2(\bar{\mathbb{P}})}$ *and define* $\bar{\bar{F}} = \max_{1\leq i\leq n}\bar{F}(Z_i)$*. Then, with* $\delta = \bar{\sigma}/\|\bar{F}\|_{L_2(\bar{\mathbb{P}})}$*,*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(f(Z_i)-\mathbb{E}[f(Z_i)]\Big)\Big|\Big]\lesssim\|\bar{F}\|_{L_2(\bar{\mathbb{P}})}J(\delta,\mathcal{F},\bar{F})+\frac{\|\bar{\bar{F}}\|_{L_2(\mathbb{P})}J(\delta,\mathcal{F},\bar{F})^2}{\delta^2\sqrt{n}},$$

where

$$J(\delta,\mathcal{F},\bar{F})=\int_0^\delta\sqrt{1+\sup_{\mathbb{Q}}N(\mathcal{F},L_2(\mathbb{Q}),\varepsilon\|\bar{F}\|_{L_2(\mathbb{Q})})}d\varepsilon.$$

Remark SA-3.2 (Improvements over literature)**.** Most of the results in this subsection are new to the literature, even in the case of non-random partitioning and without covariate-adjustments, because they take advantage of the specific binscatter structure (i.e., locally bounded series basis). The closest antecedent in the literature is Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019). Furthermore, relative to prior work, our results formally take into account the randomness of the partition formed by empirical quantiles, and account for the semi-linear regression estimation structure. ⌟

SA-3.2 Bahadur Representation

Theorem SA-3.1 (Bahadur Representation)**.** *Under Assumptions SA-DGP and SA-GL, if* $\frac{J^{\frac{\nu}{\nu-2}}\log n}{n}+\frac{J^2\log n}{n}+\frac{\log n}{J}=o(1)$, *then*

$$\sup_{x\in\mathcal{X}}\Big|\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)+\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)]\Big|$$
$$\lesssim_{\mathbb{P}} J^v\Big\{\Big(\frac{J\log n}{n}\Big)^{3/4}\sqrt{\log n}+J^{-\frac{p+1}{2}}\Big(\frac{J\log^2 n}{n}\Big)^{1/2}+J^{-p-1}+\mathfrak{r}_\gamma\Big\}.$$

The following corollary is an immediate result of Lemma SA-3.3, Lemma SA-3.4 and Theorem SA-3.1. The proof is omitted.

Corollary SA-3.1 (Uniform Convergence)**.** *Suppose that the conditions of Theorem SA-3.1 hold. If, in addition,* $\frac{(\log n)^2}{J}=o(1)$, *then*

$$\sup_{x\in\mathcal{X}}|\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)|\lesssim_{\mathbb{P}} J^v\Big(\Big(\frac{J\log n}{n}\Big)^{1/2}+J^{-p-1}\Big).$$

The next theorem shows that the proposed variance estimator is consistent.

Theorem SA-3.2 (Variance Estimate)**.** *Suppose that Assumptions SA-DGP and SA-GL hold. If* $\frac{J^{\frac{\nu}{\nu-2}}(\log n)^{\frac{\nu}{\nu-2}}}{n}+\frac{J^2\log n}{n}+\frac{(\log n)^2}{J}=o(1)$, *then*

$$\Big\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}_0\Big\|\lesssim_{\mathbb{P}} J^{-p-1}+\Big(\frac{J\log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2}\quad\textit{and}\quad\sup_{x\in\mathcal{X}}\Big|\widehat{\Omega}(x)-\Omega(x)\Big|\lesssim_{\mathbb{P}} J^{1+2v}\Big(J^{-p-1}+\Big(\frac{J\log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2}\Big).$$

Remark SA-3.3 (Improvements over literature). Theorem SA-3.1 and Corollary SA-3.1 construct the Bahadur expansion and uniform convergence of general binscatter-based M-estimators under mild rate restrictions. Specifically, we require $J^2/n = o(1)$ up to $\log n$ terms when $\nu \geq 4$. In fact, for piecewise polynomials ($s = 0$), we can show that the Bahadur expansion still holds under $J/n = o(1)$ up to $\log n$ terms when a subexponential moment restriction holds for the error term ϵ_i, which is analogous to the result for kernel-based estimators in the literature (see, e.g., Kong, Linton, and Xia, 2010). For series estimators, similar results were established for particular choices of loss functions under more stringent conditions in the literature. For example, Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019) considers series-based quantile regression, and Theorem 2 and Corollary 2 therein can be used to establish a Bahadur expansion and uniform convergence of the resulting estimators under $J^4/n^{1-\varepsilon} = o(1)$ for some $\varepsilon > 0$.

The results in Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019) are slightly stronger than that in our Theorem SA-3.1 in the sense that the expansion holds uniformly over both the evaluation point $x \in \mathcal{X}$ and the desired quantiles $u \in \mathcal{U}$ for a compact set of quantile indices $\mathcal{U} \subset (0, 1)$. Our results regarding Bahadur expansion can be extended to achieve the same level of uniformity. In general, the parameter of interest (SA-1.1) and the estimator (SA-1.2) are defined for each particular choice of the loss function within a function class $\mathcal{F}$. For the class of check functions used in quantile regression or other function classes with low complexity, it can be shown that the Bahadur expansion still holds uniformly over the evaluation point $x \in \mathcal{X}$ and the loss function $\rho \in \mathcal{F}$ under rate restrictions similar to those in Theorem SA-3.1, thereby providing an improvement over the literature. $\lrcorner$

SA-3.3 Pointwise Inference

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

$$\widehat{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}.$$

The following theorem proves the pointwise asymptotic normality of the binscatter estimator.

Theorem SA-3.3 (Pointwise Asymptotic Distribution). *Suppose that Assumptions SA-DGP and*

SA-GL *hold. If* $\sup_{x\in\mathcal{X}}\mathbb{E}[|\psi(\epsilon_i)|^\nu|x_i=x]\lesssim 1$ *for some* $\nu\geq 3$, $\frac{J^{\frac{\nu}{\nu-2}}(\log n)^{\frac{\nu}{\nu-2}}}{n}+\frac{J^2\log n}{n}+nJ^{-2p-3}=o(1)$, *then*

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

Remark SA-3.4 (Improvements over literature)**.** The result in this subsection is new to the literature, even in the case of non-random partitioning and without covariate-adjustments, because it takes advantage of the specific binscatter structure (i.e., locally bounded series basis). The closest antecedent in the literature is Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019). Furthermore, relative to prior work, our results formally take into account the randomness of the partition formed by empirical quantiles, and account for the semi-linear regression estimation structure. ⌟

SA-3.4 Uniform Inference

Let $\{a_n : n\geq 1\}$ be 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-3.4 (Strong Approximation)**.** *Under Assumptions SA-DGP and SA-GL, if*

$$\frac{J(\log n)^2}{n^{1-\frac{2}{\nu}}}+\Big(\frac{J(\log n)^5}{n}\Big)^{1/2}+nJ^{-2p-3}+J^{-p-1}(\log n)^2+nJ^{-1}\mathfrak{r}_\gamma^2=o(a_n^{-2})\quad \textit{and}\quad \frac{J^2\log n}{n}=o(1),$$

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 $\xi>0$,

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

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-3.5 (Plug-in Approximation)**.** *Suppose that the conditions in Theorem SA-3.4 hold. Then, on a properly enriched probability space there exists a K_s-dimensional standard normal random vector $\mathbf{N}_{K_s}$ independent of $\mathbf{D}$ such that for any $\xi > 0$,*

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

Remark SA-3.5 (Improvements over literature)**.** Theorems SA-3.4 and SA-3.5 provide empirical researchers with powerful tools for uniform inference based on binscatter methods. Importantly, we take into account the randomness of the empirical-quantile-based partition and construct a novel strong approximation of general binscatter-based M-estimators under mild rate restrictions. For $a_n = \sqrt{\log n}$ and $\nu \ge 4$, we require $J^2/n = o(1)$, up to $\log n$ terms. In the literature, similar results were only available in some special cases under stringent rate restrictions. For instance, Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019) considers strong approximations of general series-based quantile regression estimators. For the binscatter basis considered in this paper, their Theorem 11 can be applied to construct strong approximation of the t-statistic process based on pivotal coupling that achieves the approximation rate $a_n = n^{-\varepsilon'}$ under $J^4/n^{1-\varepsilon} = o(1)$ for some constants $\varepsilon, \varepsilon' > 0$, whereas their Theorem 12 can be used to construct strong approximation based on Gaussian processes under $J^5/n^{1-\varepsilon} = o(1)$. It should be noted that their notion of strong approximation is stronger than ours in the sense that it holds uniformly over both the evaluation point $x \in \mathcal{X}$ and the desired quantile $u \in \mathcal{U}$ for a compact set of quantile indices $\mathcal{U} \subset (0, 1)$. $\lrcorner$

SA-3.5 Integrated Mean Squared Error

Theorem SA-3.6 (IMSE)**.** *Suppose that Assumptions SA-DGP and SA-GL hold. Let $\omega(x)$ be a continuous weighting function over $\mathcal{X}$ bounded away from zero. If $\frac{J^{(\frac{\nu}{\nu-2}\vee 2)}\log n}{n} + \frac{(\log n)^5}{J} = o(1)$, then*

$$\int_{\mathcal{X}} \Big(\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)\Big)^2 \omega(x)dx = \mathtt{AISE} + o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n} + J^{-2(p+1-v)}\Big)$$

where

$$
\begin{aligned}
\mathbb{E}[\texttt{AISE}|\mathbf{X},\mathbf{W}] &= \frac{J^{1+2v}}{n}\mathscr{V}_n(p,s,v) + J^{-2(p+1-v)}\mathscr{B}_n(p,s,v) + o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n} + J^{-2(p+1-v)}\Big),\\
\mathscr{V}_n(p,s,v) &:= J^{-(1+2v)}\,\mathrm{trace}\Big(\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\Big) \asymp 1,\\
\mathscr{B}_n(p,s,v) &:= J^{2p+2-2v}\int_{\mathcal{X}}\Big(r_{0,v}(x) - \mathbf{b}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_s(x_i)\varkappa(x_i,\mathbf{w}_i)r_0(x_i)]\Big)^2\omega(x)dx \lesssim 1.
\end{aligned}
$$

As in the least squares case, as long as $\mathscr{B}_n(p,s,v) \gtrsim 1$, the above theorem implies that the (approximate) IMSE-optimal number of bins satisfies that $J_{\texttt{AIMSE}} \asymp n^{\frac{1}{2p+3}}$. Relying on the IMSE expansion in Theorem SA-3.6, one may design a data-driven procedure to select the IMSE-optimal number of bins for general binscatter-based M-estimators.

Remark SA-3.6 (Improvements over literature)**.** The results in this subsection are new to the literature, even in the case of non-random partitioning and without covariate-adjustments, for both general non-linear series estimators and binscatter (piecewise polynomials and splines) non-linear series estimators in particular. Furthermore, our results formally take into account the randomness of the partition formed by empirical quantiles, and account for the semi-linear regression estimation structure. ⌟

SA-4 Applications

Our results can be used to draw inference on many parameters of interest, for example,

$$
\mu_0^{(v)}(x), \quad \vartheta(x;\mathbf{w}) := \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0),
$$

and transformations thereof. For simplicity, we will focus on $\mu_0^{(v)}(x)$ first, and then discuss the inference on $\vartheta(x;\mathbf{w})$ and $\frac{\partial}{\partial x}\vartheta(x;\mathbf{w})$ in Section SA-4.4.

Remark SA-4.1 (Improvements over literature)**.** The upcoming results in this section are new to the literature, even in the case of non-random partitioning and without covariate-adjustments, because they take advantage of the specific binscatter structure (i.e., locally bounded series basis). Furthermore, relative to prior work, our results formally take into account the randomness of the partition formed by empirical quantiles, account for the semi-linear regression estimation structure, and consider an array of linear and non-linear estimation and inference problems. In particular, the approach taken in Theorems SA-4.1 and SA-4.4 to establish strong approximation and related distributional approximations for linear and non-linear binscatter statistics may be of independent interest. ⌟

SA-4.1 Confidence Bands

Theorems SA-2.4, SA-2.5, SA-3.4 and SA-3.5 offer a way to approximate the distribution of the *whole* t-statistic process. A direct application of these results is to constructing uniform confidence bands, 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-4.1 (Supremum Approximation)**.** *Let $a_n = \sqrt{\log J}$. Suppose that the conditions of Theorem SA-2.4 hold for the squared loss, or the conditions of Theorem SA-3.4 hold for a general loss. Then,*

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

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

Corollary SA-4.1. *For a given $p = p_0$, suppose the conditions in Theorem SA-2.6 hold with $\mathscr{B}_n \gtrsim 1$ and let $J = J_{\texttt{IMSE}}$. In addition, for some $p = p_0+q$ ($q \geq 1$), Assumptions SA-DGP and SA-LS hold for the squared loss, or Assumptions SA-DGP and SA-GL hold and $\sqrt{\log J}\mathfrak{r}_\gamma = o(\sqrt{J/n})$ for a general loss. If $\mathfrak{c} = \inf\Big\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}} |\widehat{Z}_{p+q}(x)| \leq c\,|\mathbf{D}] \geq 1-\alpha\Big\}$, then*

$$\mathbb{P}\Big[\mu_0^{(v)}(x) \in \widehat{I}_{p+q}(x),\ \textit{for all}\ x \in \mathcal{X}\Big] = 1-\alpha+o(1).$$

Remark SA-4.2. The above results construct valid uniform confidence bands for general binscatter-based M-estimators under mild rate restrictions. Specifically, when $\nu \geq 4$, we require $J^2/n = o(1)$, up to $\log n$ terms. In contrast, Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019) considers general series-based quantile regression estimators, and Theorem 15 therein can be used to construct confidence bands for binscatter estimators via various resampling methods under $J^4/n^{1-\varepsilon} = o(1)$ for some $\varepsilon > 0$. ⌟

SA-4.2 Parametric Specification Tests

As another application, we can test parametric specifications of the unknown function $\mu_0(x)$ and other transformations thereof. To be specific, consider the following problem:

$$\ddot{\mathsf{H}}_0: \sup_{x\in\mathcal{X}} |\mu_0^{(v)}(x) - m^{(v)}(x,\boldsymbol{\theta}_0)| = 0 \text{ for some } \boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \quad \text{v.s.}$$
$$\ddot{\mathsf{H}}_A: \sup_{x\in\mathcal{X}} |\mu_0^{(v)}(x) - m^{(v)}(x,\boldsymbol{\theta}_0)| > 0 \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

This testing problem 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 $\boldsymbol{\theta}_0$ 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 $\mathfrak{c}$.

Theorem SA-4.2 (Specification Tests)**.** *Let $a_n = \sqrt{\log J}$. Suppose that the conditions in Theorem SA-2.4 hold for the squared loss, or the conditions in Theorem SA-3.4 hold for a general loss. Let $\mathfrak{c} = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}} |\widehat{Z}_p(x)| \le c|\mathbf{D}] \ge 1-\alpha\}$.*

Under $\ddot{\mathsf{H}}_0$, if $\sup_{x\in\mathcal{X}} |\mu_0^{(v)}(x) - m^{(v)}(x; \widehat{\boldsymbol{\theta}})| = o_{\mathbb{P}}\Big(\sqrt{\frac{J^{1+2v}}{n\log J}}\Big)$, then

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

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 $J^v\Big(\frac{J\log J}{n}\Big)^{1/2} = o(1)$, then

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

The robust bias-corrected testing procedure given in Theorem 4 of the main paper is an immediate corollary of Theorem SA-4.2, given the choice of J, the rate of convergence of $\widehat{\boldsymbol{\gamma}}$ in Theorem SA-2.5, and the condition on $\widehat{\mathbf{w}}$. To conserve some space, we do not repeat their statements.

SA-4.3 Shape Restriction Tests

The third application of our results is to test certain shape restrictions on the unknown functions. To be specific, consider the following problem:

$$\dot{\mathsf{H}}_0: \sup_{x\in\mathcal{X}} (\mu_0^{(v)}(x) - m^{(v)}(x,\bar{\boldsymbol{\theta}})) \leq 0 \text{ for a certain } \bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta} \quad \text{v.s.}$$

$$\dot{\mathsf{H}}_A: \sup_{x\in\mathcal{X}} (\mu_0^{(v)}(x) - m^{(v)}(x,\bar{\boldsymbol{\theta}})) > 0 \text{ for } \bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta}.$$

This testing problem 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-4.3 (Shape Restriction Tests). *Let $a_n = \sqrt{\log J}$. Suppose that the conditions in Theorem SA-2.4 hold for the squared loss, or the conditions in Theorem SA-3.4 hold for a general loss. Let $\mathfrak{c} = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}} \widehat{Z}_p(x) \leq c|\mathbf{D}] \geq 1-\alpha\}$. Assume that $\sup_{x\in\mathcal{X}} |m^{(v)}(x; \widehat{\bar{\boldsymbol{\theta}}}) - m^{(v)}(x,\bar{\boldsymbol{\theta}})| = o_{\mathbb{P}}\Big(\sqrt{\frac{J^{1+2v}}{n\log J}}\Big)$.*

Under $\dot{H}_0$,

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

Under $\dot{H}_A$, if $J^v\Big(\frac{J\log J}{n}\Big)^{1/2} = o(1)$,

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

The robust bias-corrected testing procedure given in Theorem 5 of the main paper is an immediate corollary of Theorem SA-4.3, given the choice of J, the rate of convergence of $\widehat{\boldsymbol{\gamma}}$ in Theorem SA-2.5, and the condition on $\widehat{\mathbf{w}}$. To conserve some space, we do not repeat their statements.

SA-4.4 Binscatter Parameters

The results above can be extended to inference on other parameters of interest, once an expansion similar to Theorem SA-GL is established. In this subsection, we focus on

$$\vartheta(x; \mathbf{w}) := \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0) \quad \text{and} \quad \vartheta^{(1)}(x; \mathbf{w}) := \frac{\partial}{\partial x}\vartheta(x; \mathbf{w}),$$

where $\mathbf{w}$ is a particular user-specified evaluation point. Some simple choices are $\mathbf{w} = \mathbf{0}$, $\mathbb{E}[\mathbf{w}_i]$, and $\mathsf{Median}(\mathbf{w}_i)$. The corresponding strong approximation results are constructed based on the following estimators:

$$\widehat{\vartheta}(x; \widehat{\mathbf{w}}) = \eta(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}) \qquad \widehat{\vartheta}^{(1)}(x; \widehat{\mathbf{w}}) = \eta^{(1)}(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}})\widehat{\mu}^{(1)}(x),$$

where $\widehat{\mathbf{w}}$ is an estimator of $\mathbf{w}$.

Theorem SA-4.4. *Assume* $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n})$.

(i) Let $\sigma_\vartheta(x) = |\eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)|(\widehat{\mathbf{b}}_s(x)'\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\widehat{\mathbf{b}}_s(x))^{1/2}$. *Suppose that the conditions of Theorem SA-3.4 hold. Then, there exists some* K_s*-dimensional standard normal random vector* $\mathbf{N}_{K_s}$ *such that for any* $\xi > 0$,

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|\frac{\sqrt{n}(\widehat{\vartheta}(x,\widehat{\mathbf{w}}) - \vartheta(x,\mathbf{w}))}{\sigma_\vartheta(x)} - \frac{\widehat{\mathbf{b}}_s(x)'\eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\mathbf{Q}_0^{-1}}{\sigma_\vartheta(x)}\boldsymbol{\Sigma}_0^{1/2}\mathbf{N}_{K_s}\Big| > \xi a_n^{-1}\Big) = o(1).$$

(ii) Let $\sigma_{\vartheta,1}(x) = |\eta^{(1)}(\mu_0(x)+\mathbf{w}'\boldsymbol{\gamma}_0)|(\widehat{\mathbf{b}}_s^{(1)}(x)'\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\widehat{\mathbf{b}}_s^{(1)}(x))^{1/2}$. *Suppose that the conditions of Theorem SA-3.4 hold for some* $p \geq 1$. *Then, there exists some* K_s*-dimensional standard normal random vector* $\mathbf{N}_{K_s}$ *such that for any* $\xi > 0$,

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|\frac{\sqrt{n}(\widehat{\vartheta}^{(1)}(x,\widehat{\mathbf{w}}) - \vartheta^{(1)}(x,\mathbf{w}))}{\sigma_{\vartheta,1}(x)} - \frac{\widehat{\mathbf{b}}_s^{(1)}(x)'\eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\mathbf{Q}_0^{-1}}{\sigma_{\vartheta,1}(x)}\boldsymbol{\Sigma}_0^{1/2}\mathbf{N}_{K_s}\Big| > \xi a_n^{-1}\Big) = o(1).$$

As in Theorem SA-3.5, feasible approximation processes can be constructed by replacing all the unknown quantities, i.e., $\mathbf{Q}_0$, $\boldsymbol{\Sigma}_0$, μ_0, $\mathbf{w}$ and $\boldsymbol{\gamma}_0$, by corresponding estimators $\widehat{\mathbf{Q}}$, $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\mu}$, $\widehat{\mathbf{w}}$ and $\widehat{\boldsymbol{\gamma}}$.

SA-5 Implementation Details

SA-5.1 Standard Error Computation

With the variance estimator $\widehat{\Omega}(x)$ given in Section SA-3, we have obtained the standard error of $\widehat{\mu}^{(v)}(x)$. Regarding $\widehat{\vartheta}(x;\widehat{\mathbf{w}})$ or $\widehat{\vartheta}^{(1)}(x;\widehat{\mathbf{w}})$, the following construction can be used to to compute their standard errors:

$$\widehat{\sigma}_{\vartheta} := |\eta^{(1)}(\widehat{\mu}(x)+\widehat{\mathbf{w}}'\widehat{\gamma})|\Big(\widehat{\mathbf{b}}_s(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x)\Big)^{1/2},$$
$$\widehat{\sigma}_{\vartheta,1} := |\eta^{(1)}(\widehat{\mu}(x)+\widehat{\mathbf{w}}'\widehat{\gamma})|\Big(\widehat{\mathbf{b}}_s^{(1)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(1)}(x)\Big)^{1/2}.$$

Recall the formula for the estimator $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}_0$:

$$\widehat{\mathbf{\Sigma}} = \mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\psi(\widehat{\epsilon}_i)^2\eta^{(1)}(\widehat{\mu}(x_i)+\mathbf{w}_i'\widehat{\gamma})^2\Big].$$

Note that it only relies on known or estimable quantities such as the derivative of the loss function $\psi(\cdot)$, the derivative of the inverse link function $\eta^{(1)}(\cdot)$, the residual $\widehat{\epsilon}_i$ and the estimate $\widehat{\mu}(\cdot)$. Thus, $\widehat{\mathbf{\Sigma}}$ and other types of heteroskedasticity-robust "meat" matrix estimators can be easily constructed using the data. Then, it remains to obtain an estimator $\widehat{\mathbf{Q}}$ of $\mathbf{Q}_0$, which in general relies on another estimator $\widehat{\Psi}_1(\cdot)$ and can be constructed in a case-by-case basis. In the following we discuss several examples.

Example 1 (Least Squares Regression). For least squares regression, the loss function $\rho(y,\eta) = \frac{1}{2}(y-\eta)^2$ and the (inverse) link function $\eta(\theta)=\theta$. Therefore, $\psi(\epsilon_i) = -\epsilon_i$ and $\eta_{i,1}=1$. Thus, the formula for $\widehat{\mathbf{Q}}$ given in Section SA-3 reduces to that given in Section SA-2, which is immediately feasible in practice.

Example 2 (Logistic Regression). For logistic regression, the loss function is given by the corresponding likelihood function, i.e., $-\rho(y,\eta) = y\log\eta + (1-y)\log(1-\eta)$, and the inverse link is given by the logistic function $\eta(\theta) = \frac{e^\theta}{1+e^\theta}$. Accordingly, an estimator of $\mathbf{Q}_0$ is given by

$$\widehat{\mathbf{Q}} = \mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\eta}_i(1-\widehat{\eta}_i)\Big], \quad \widehat{\eta}_i = \eta(\widehat{\mu}(x_i)+\mathbf{w}_i'\widehat{\gamma}).$$

Example 3 (Quantile Regression). For quantile regression, $\rho(y,\eta) = (q-\mathbb{1}(y<\eta))(y-\eta)$ for some $q \in (0,1)$ and $\eta(\theta) = \theta$. Accordingly, $\psi(\epsilon_i) = \mathbb{1}(\epsilon_i < 0) - q$, and one needs to estimate

$$\mathbf{Q}_0 = \mathbb{E}\Big[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)' f_{Y|XW}(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0 | x_i, \mathbf{w}_i)\Big].$$

The key is to estimate the conditional density $f_{Y|XW}(\cdot|x_i, \mathbf{w}_i)$ evaluated at the conditional quantile of interest $(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0)$, whose reciprocal is termed "sparsity function" in the literature. Many different methods have been proposed. For example, the sparsity function is simply the derivative of the conditional quantile function with respect to the quantile, which can be estimated by using the difference quotient of the estimated conditional quantile function. Alternatively, $\mathbf{Q}_0$ can be viewed as a matrix-weighted density function, and one can construct a corresponding estimator based on kernel density estimation ideas. In addition, one can use bootstrapping methods to estimate the variance, avoiding the technical difficulty of estimating the sparsity function. See Section 3.4 and Section 3.9 of Koenker (2005) for more discussion of variance estimation for quantile regression.

SA-5.2 Number of Bins Selector

We discuss the implementation details for data-driven selection of the number of bins, based on the integrated mean squared error expansion for the squared loss (see Theorem SA-2.6, Corollary SA-2.3 and Corollary SA-2.4). Note that for general loss functions, Theorem SA-3.6 implies that the (approximate) IMSE-optimal number of bins has the same order as that for the squared loss. Therefore, the selectors given below can provide a choice of J with the "correct" rate that balances the leading bias and variance in the IMSE expansion in general, and it achieves optimality in the special case of least squares regression. For other loss functions, one may design data-driven procedures to select the IMSE-optimal J based on Theorem SA-3.6.

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

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

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

SA-5.2.1 Rule-of-thumb Selector

A rule-of-thumb choice of J is obtained based on Corollary SA-2.3, in which case $s = 0$.

Regarding the variance constants $\mathscr{V}(p, 0, v)$, the unknowns are the density function $f_X(x)$ and the conditional variance $\sigma^2(x)$. A Gaussian reference model is employed for $f_X(x)$. 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

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

Regarding the bias constant, the unknowns are $f_X(x)$, which is estimated using the Gaussian reference model, and $\mu_0^{(p+1)}(x)$, which can be estimated based on the global polynomial regression that approximates $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$. Then, the bias constant is estimated by

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

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

SA-5.2.2 Direct-plug-in Selector

The direct-plug-in selector is implemented based on binscatter estimators, which applies 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 $J_{\texttt{pre}}$ 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 Theorem SA-2.2.

Regarding the bias constant, we employ the uniform approximation (SA-7.6) in the proof of Theorem SA-2.6. 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 $\mu_0^{(p+1)}(x)$, which can be estimated by implementing a binscatter regression of order $p+1$ (with the preliminary partition unchanged). Plug it in (SA-7.7), and all other quantities in that equation 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-6 Empirical Illustrations

In this section we provide details on the simulation model used in Sections 2, 3 and 4 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://github.com/nppackages-replication/CCFF_2021_wp`.

SA-6.1 Simulated Data

In 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 $x_i \sim \texttt{beta}(2, 4)$ where $\texttt{beta}(2, 4)$ is *beta* distribution with parameters 2 and 4. The regression function of interest is

$$\mu_0(x) = 24x^4 - 98.8x^3 + 112.4x^2 - 44.4x + 21.6.$$

The basic model for the outcome variable y_i is constructed as

$$y_i = \mu_0(x_i) + w_i + \epsilon_i.$$

Regarding the control variable w_i, we allow for its dependence on x_i: $w_i = 3(x_i - 0.5) + \mathtt{U}(-0.5, 0.5)$, where $\mathtt{U}(-0.5, 0.5)$ is a random variable that is uniformly distributed over $[-0.5, 0.5]$ and independent of x_i. In the discussion of covariate adjustment, we also consider a case in which w_i is independent of x_i and uniformly distributed: $w_i \sim \mathtt{U}(-1, 1)$.

Regarding the error term ϵ_i, we allow for heteroskedasticity (over x_i):

$$\epsilon_i = 0.2 * (|\mathsf{S}_i - 3.5| - 4.5)^2 * \upsilon_{i,1} + \upsilon_{i,2},$$

where

$$\mathsf{S}_i = \begin{cases} 1 & \text{if } x_i \in \left[x_{(1)}, x_{(\lfloor n/10 \rfloor)}\right) \\ j & \text{if } x_i \in \left[x_{(\lfloor (j-1)n/10 \rfloor)}, x_{(\lfloor jn/10 \rfloor)}\right) \text{ and } 2 \leq j \leq 9 \\ 10 & \text{if } x_i \in \left[x_{(\lfloor 9n/10 \rfloor)}, x_{(n)}\right] \end{cases},$$

and $\upsilon_{i,1}, \upsilon_{i,2} \sim \mathtt{N}(0, 0.5^2)$, that is, they are normally distributed with mean 0 and variance 0.5^2. $\upsilon_{i,1}$ and $\upsilon_{i,2}$ are independent of each other and are independent of (x_i, w_i).

For group comparison purpose, we generate another dataset of the same sample size ($n = 1,000$). The data generating process is the same as above except that a different regression function is specified:

$$\mu_1(x) = \mu_0(x) - 25(x - 0.5)^2 + 2.$$

Finally, in Section 4 of the main paper, we introduce a binary response variable to illustrate the binscatter-based logistic regression: $t_i = \mathbb{1}(y_i \leq y_{(\lfloor n/3 \rfloor)})$ where $y_{(\lfloor n/3 \rfloor)}$ is the $\frac{1}{3}$-empirical-quantile of y_i in the first dataset (generated based on regression function $\mu_0(\cdot)$). In our simulated dataset, $y_{(\lfloor n/3 \rfloor)} = 15.77$.

SA-6.2 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.

The top row of Figure SA-1 shows the output from the command `binscatter` with and without control variables. The second row shows the output from `binsreg` using a data-driven (optimal) choice of number of bins. The third row displays confidence intervals and confidence bands for the choice $p = 1$ and $s = 1$. Finally, the bottom row displays point estimates (line) and confidence bands for the choice $p = 3$ and $s = 3$.

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

SA-6.3 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.

The top row of Figure SA-2 shows the output from the command `binscatter` with and without control variables. The second row shows the output from `binsreg` using a data-driven (optimal) choice of number of bins. The third row utilizes a binary variable for whether the median age in the zip code is below 40 years.[1]. The fourth row displays the pairwise comparison along with the point estimate (line) and confidence bands for the choice $p = 3$ and $s = 3$.

[1] For the pairwise comparisons with control variables we remove the median age in the set of controls.

SA-6.4 Example 3: Uninsured Rate versus per Capital Income

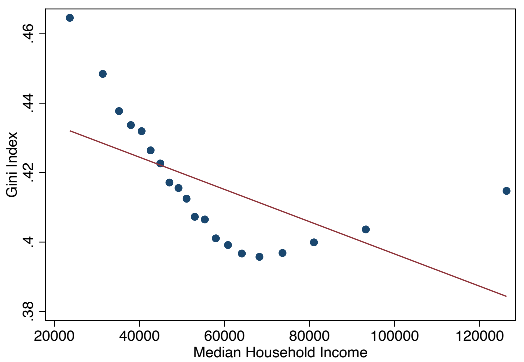
The results are provided in Figure SA-3. The dependent variable is the percentage of individuals without health insurance and the variable of interest is per capita income. The control variables are (i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's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

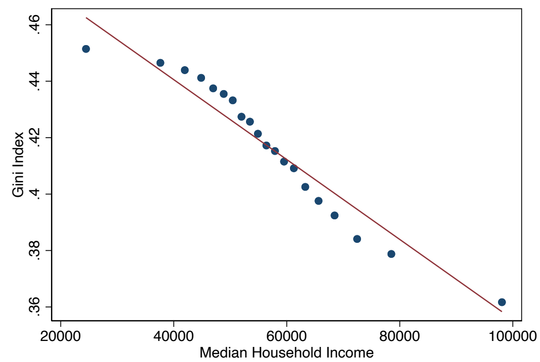
The top row of Figure SA-3 shows the output from the command `binscatter` with and without control variables. The second row shows the output from `binsreg` using a data-driven (optimal) choice of number of bins. The third row displays pointwise confidence intervals for the choice of $p = 3$ and $s = 3$ along with a global cubic fit (orange line) to the data. Finally, the bottom row displays point estimates (line) and confidence bands for the choice $p = 3$ and $s = 3$.

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

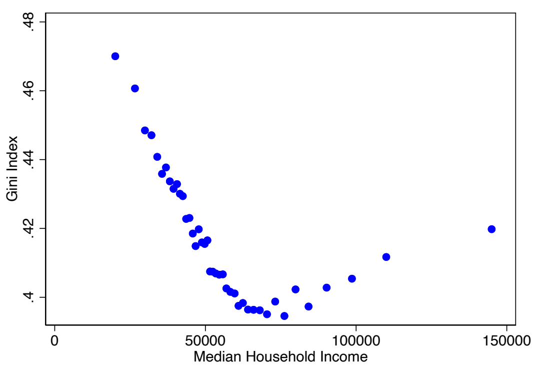
No Controls | **With Controls**

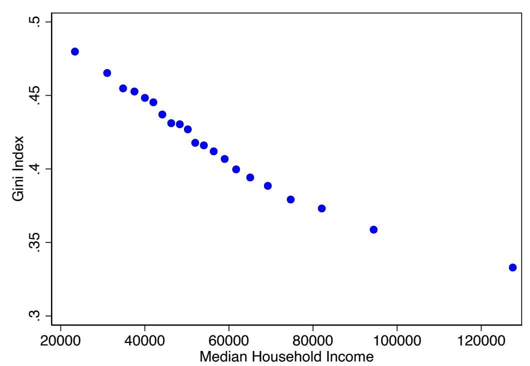
binscatter | binscatter

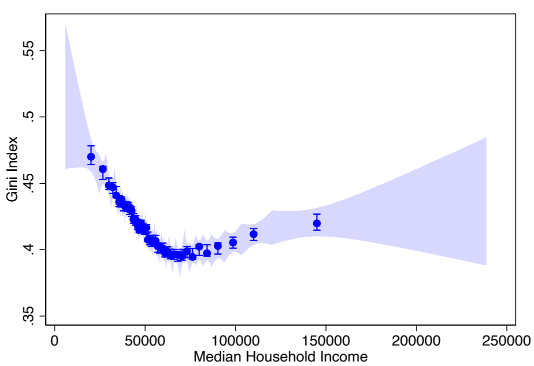


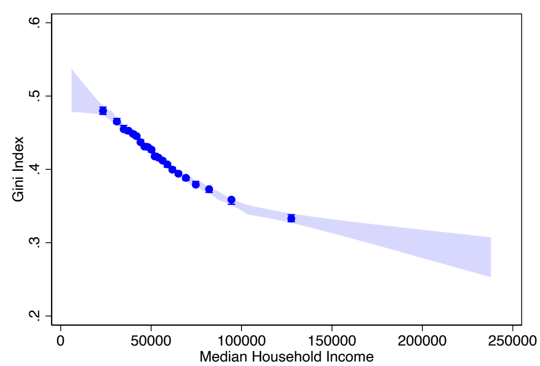


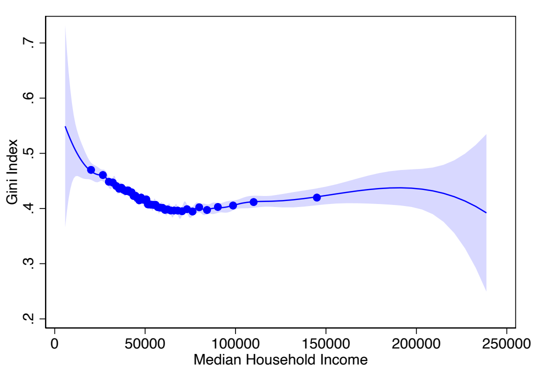
binsreg | binsreg











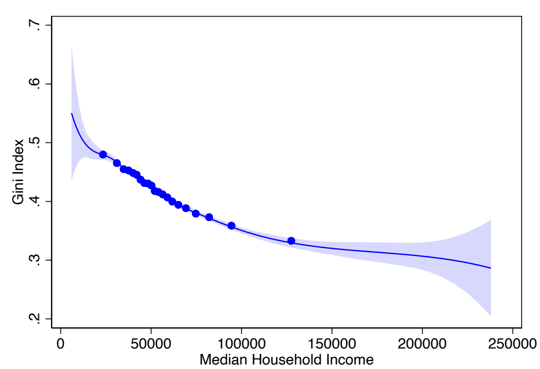
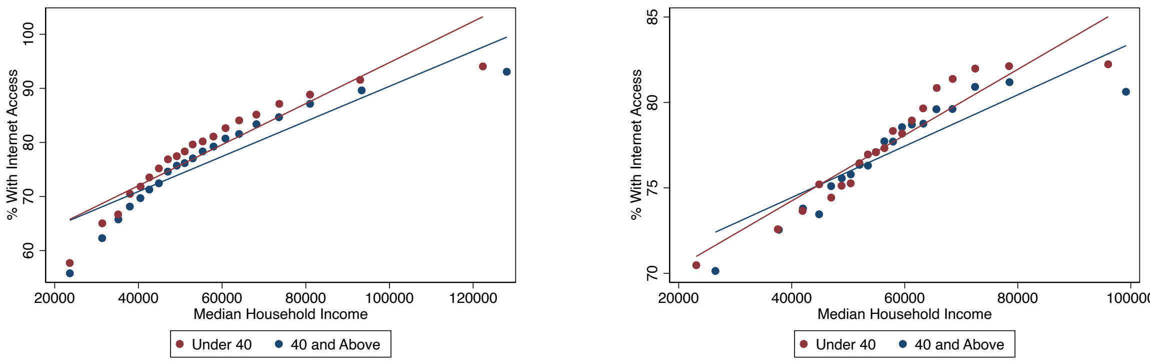


Figure SA-2: **Internet Access versus Household Income**

No Controls | **With Controls**

binscatter | binscatter



binsreg | binsreg

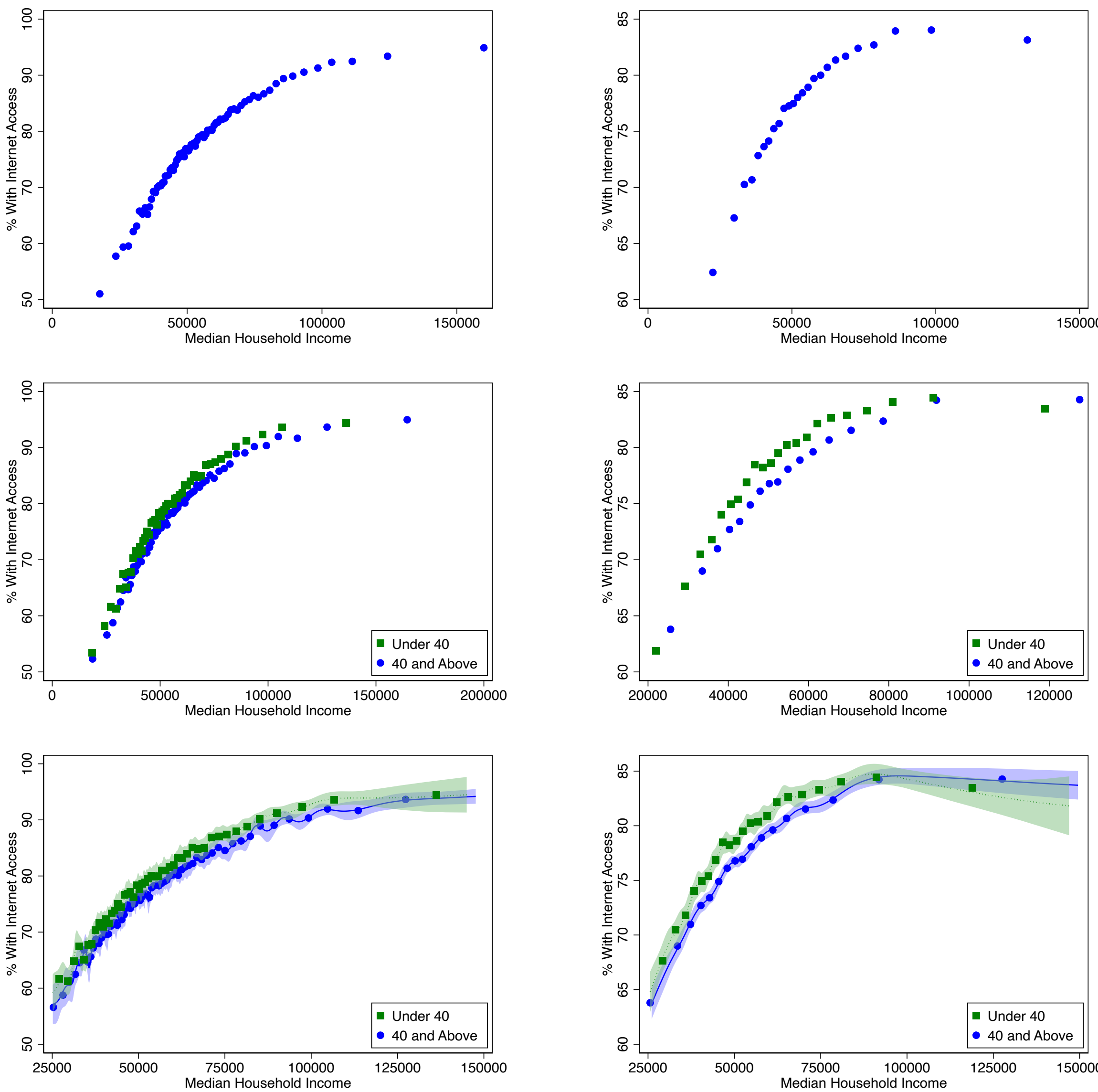
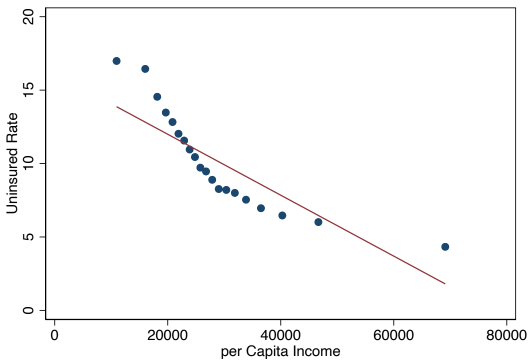
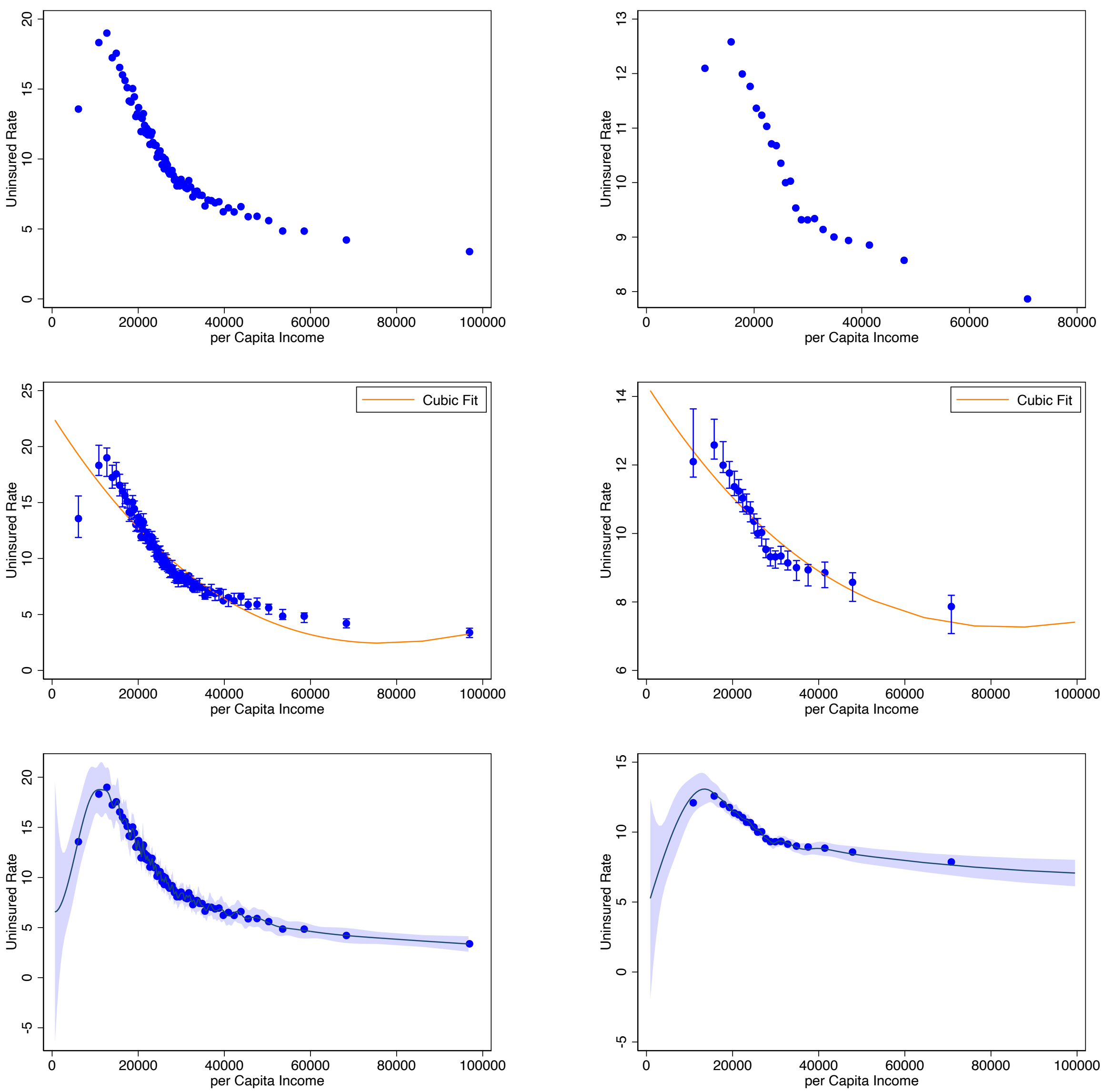


Figure SA-3: **Uninsured Rate versus per Capita Income**

No Controls	**With Controls**
binscatter	binscatter



binsreg	binsreg



SA-7 Proof

SA-7.1 Proof for Section SA-1.3

SA-7.1.1 Proof of Lemma SA-1.1

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_0 based on the population quantiles. By the mean value theorem,

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

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

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

Then,

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

and the desired result follows. □

SA-7.1.2 Proof of Lemma SA-1.2

Proof. For $s = 0$, the result is trivial. For $0 < s \leq p$, $\widehat{\mathbf{b}}_s(x)$ 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 $(\xi_\ell, \xi_{\ell+p+1})$ 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

$$\varphi_{j,\alpha}(x) = \sqrt{J} \cdot \mathbb{1}_{\widehat{\mathcal{B}}_j}(x) \Big(\frac{x - \hat{\tau}_{j-1}}{\hat{h}_j} \Big)^{\alpha}, \quad 1 \leq \alpha \leq p, \quad 1 \leq j \leq J. \tag{SA-7.1}$$

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

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

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-1.1, $\sup_{j,\alpha} |c_{j,\alpha}| \lesssim_{\mathbb{P}} 1$.

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 SA-1.1, and noticing the form of $c_{j,\alpha}$, $\max_{k,l} |(\widehat{\mathbf{T}}_s - \mathbf{T}_s)_{k,l}| \lesssim \sqrt{J\log J/n}$ 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, $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\|_\infty \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$ and $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$.

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

$$\widehat{b}_{s,\ell}(x) = (\hat{\tau}_{\ell+p+1} - \hat{\tau}_\ell) \sum_{j=\ell}^{\ell+p+1} \frac{(x-\hat{\tau}_j)_+^p}{\prod_{\substack{k=\ell \\ k\neq l}}^{\ell+p+1} (\hat{\tau}_k - \hat{\tau}_j)}, \tag{SA-7.2}$$

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 $\varphi_{j,\alpha}(x)$, $j = \ell+1, \ldots, \ell+p+1, \alpha = 0, \ldots, m-1$:

$$\widehat{b}_{s,\ell}(x) = \sum_{\alpha=0}^{p} \sum_{j=\ell+1}^{\ell+p+1} c_{j,\alpha}\varphi_{j,\alpha}(x), \quad \text{for some } c_{j,\alpha}. \tag{SA-7.3}$$

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-7.3), combined with its coefficient in Equation (SA-7.2), 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-7.1.3 Proof of Lemma SA-1.3

Proof. The sparsity of the basis follows by construction. To show the bound on $\|\widehat{\mathbf{b}}_s^{(v)}(x)\|$, notice that when $s = 0$, for any $x \in \mathcal{X}$ and any $j = 1, \ldots, J(p+1)$, $0 \leq \widehat{b}_{0,j}(x) \leq \sqrt{J}$. Define $\varphi_{j,\alpha}(x)$ as in Equation (SA-7.1). 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-1.1 and Lemma SA-1.2. $\square$

SA-7.1.4 Proof of Lemma SA-1.4

Proof. By Lemma SA-1.1, it suffices to establish the approximation power of $\mathbf{b}_s(x;\Delta)$ for all $\Delta \in \Pi$. For $v = 0$, by Theorem 6.27 of Schumaker (2007), $\max_{\Delta\in\Pi}\min_{\boldsymbol{\beta}\in\mathbb{R}^{K_s}}\sup_{x\in\mathcal{X}}|\mu_0(x)-\mathbf{b}_s(x;\Delta)'\boldsymbol{\beta}| \lesssim J^{-p-1}$. By Huang (2003) and Assumption SA-DGP, 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 $v > 0$, again, we only need to consider the case where Δ belongs to Π. For any $\Delta \in \Pi$, we can take the best L_∞-approximation: for some $\boldsymbol{\beta}_\infty(\Delta) \in \mathbb{R}^{K_s}$, $\|\mu_0(\cdot)-\mathbf{b}_s(\cdot;\Delta)'\boldsymbol{\beta}_\infty(\Delta)\|_\infty \lesssim J^{-p-1}$, and $\|\mu_0^{(v)}(\cdot)-\mathbf{b}_s^{(v)}(\cdot;\Delta)'\boldsymbol{\beta}_\infty(\Delta)\|_\infty \lesssim J^{-p-1+v}$. Such a construction exists by Lemma SA-6.1 of Cattaneo, Farrell, and Feng (2020). Then, $\|\mu_0^{(v)}(\cdot)-\mathbf{b}_s^{(v)}(\cdot;\Delta)'\boldsymbol{\beta}_0(\Delta)\|_\infty \lesssim \|\mu_0^{(v)}(\cdot)-\mathbf{b}_s^{(v)}(\cdot;\Delta)'\boldsymbol{\beta}_\infty(\Delta)\|_\infty + \|\mathbf{b}_s^{(v)}(\cdot;\Delta)'(\boldsymbol{\beta}_\infty(\Delta)-\boldsymbol{\beta}_0(\Delta))\|_\infty \lesssim J^{-p-1+v} + \|\mathbf{b}_s^{(v)}(\cdot;\Delta)'(\boldsymbol{\beta}_\infty(\Delta)-\boldsymbol{\beta}_0(\Delta))\|_\infty$. By definition of $\boldsymbol{\beta}_0(\Delta)$,

$$\boldsymbol{\beta}_0(\Delta)-\boldsymbol{\beta}_\infty(\Delta) = \mathbb{E}[\mathbf{b}_s(x_i;\Delta)\mathbf{b}_s(x_i;\Delta)']^{-1}\mathbb{E}[\mathbf{b}_s(x_i;\Delta)r_\infty(x_i;\Delta)],$$

where $r_\infty(x_i;\Delta) = \mu_0(x_i)-\mathbf{b}_s(x_i;\Delta)'\boldsymbol{\beta}_\infty(\Delta)$. By the argument given later in the proof of Lemma SA-2.1 in Section SA-2, we have $\|\mathbb{E}[\mathbf{b}_s(x_i;\Delta)\mathbf{b}_s(x_i;\Delta)']^{-1}\|_\infty \lesssim 1$ uniformly over $\Delta \in \Pi$. Since $\mathbf{b}_s(x_i;\Delta)$ is supported on a finite number of bins, $\|\mathbb{E}[\mathbf{b}_s(x_i;\Delta)r_\infty(x_i;\Delta)]\|_\infty \lesssim J^{-p-1-1/2}$. Then the desired result follows. $\square$

SA-7.2 Proof for Section SA-2

SA-7.2.1 Proof of Lemma SA-2.1

Proof. The upper bound on the maximum eigenvalue of $\mathbf{Q}_0$ follows from Lemma SA-1.2, and the quasi-uniformity property of population quantiles shown in the proof of Lemma SA-1.1. Also, in view of Lemma SA-1.1, the lower bound on the minimum eigenvalue of $\mathbf{Q}_0$ follows from Theorem 4.41 of Schumaker (2007), by which the minimum eigenvalue of $\mathbf{Q}_0/J$ (the scaling factor dropped) is bounded by $\min_{1\leq j\leq J} h_j$ up to some universal constant.

Now, we prove the convergence of $\widehat{\mathbf{Q}}$. In view of Lemma SA-1.2, 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-1.1, with probability approaching one, $\widehat{\Delta}$ 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 $\mathcal{A}_n$,

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

By the relation between matrix norms, the right-hand-side of the above inequality is further bounded by $\sup_{\Delta\in\Pi}\|\mathbb{E}_n[\mathbf{b}_0(x_i;\Delta)\mathbf{b}_0(x_i;\Delta)'] - \mathbb{E}[\mathbf{b}_0(x_i;\Delta)\mathbf{b}_0(x_i;\Delta)']\|_\infty$. Let a_{kl} be a generic (k,l)th entry of the matrix inside the matrix norm, i.e.,

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

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

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

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 $b_{0,k}(\cdot;\Delta)$ and $b_{0,l}(\cdot;\Delta)$ 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

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

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),

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

implying that

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

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

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

Next, Let α_{kl} be a generic (k,l)th entry of $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)']/J - \mathbb{E}[\mathbf{b}_0(x_i)\mathbf{b}_0(x_i)']/J$, 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

$$\begin{aligned}
\alpha_{kl} =& \int_{\widehat{\mathcal{B}}_j}\Big(\frac{x-\hat{\tau}_j}{\hat{h}_j}\Big)^{\ell} f_X(x)dx - \int_{\widehat{\mathcal{B}}_j}\Big(\frac{x-\tau_j}{h_j}\Big)^{\ell} f_X(x)dx \\
=& \hat{h}_j\int_0^1 z^{\ell} f_X(z\hat{h}_j+\hat{\tau}_j)dz - h_j\int_0^1 z^{\ell} f_X(zh_j+\tau_j)dz \\
=& (\hat{h}_j - h_j)\int_0^1 z^{\ell} f_X(z\hat{h}_j+\hat{\tau}_j)dz + h_j\int_0^1 z^{\ell}\Big(f_X(z\hat{h}_j+\hat{\tau}_j) - f_X(zh_j+\tau_j)\Big)dz
\end{aligned} \qquad \text{(SA-7.4)}$$

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

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

Since $f_X(\cdot)$ is uniformly continuous on $\mathcal{X}$, the second term in (SA-7.4) 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 $\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'] - \mathbf{Q}_0\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$.

Given the above fact, it follows that $\|\widehat{\mathbf{Q}}^{-1}\| \lesssim_{\mathbb{P}} 1$. Notice that $\widehat{\mathbf{Q}}$ and $\mathbf{Q}_0$ are banded matrices with finite band width. Then the bounds on $\|\widehat{\mathbf{Q}}\|_\infty$ and $\|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\|_\infty$ hold by Theorem 2.2 of Demko (1977). This completes the proof. □

SA-7.2.2 Proof of Lemma SA-2.2

Proof. Since $\mathbb{E}[\epsilon_i^2|x_i = x]$ is bounded and bounded away from zero uniformly over $x \in \mathcal{X}$, we have $\widehat{\mathbf{Q}} \lesssim \bar{\mathbf{\Sigma}} \lesssim \widehat{\mathbf{Q}}$. Then, by Lemma SA-2.1, $1 \lesssim_{\mathbb{P}} \lambda_{\min}(\bar{\mathbf{\Sigma}}) \lesssim \lambda_{\max}(\bar{\mathbf{\Sigma}}) \lesssim_{\mathbb{P}} 1$. The upper bound on $\bar{\Omega}(x)$ immediately follows by Lemmas SA-1.3 and SA-2.1.

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 in which $\widehat{\Delta} \in \Pi$. Introduce an auxiliary function $\varrho(x) = (x - x_0)^v/h_{x_0}^v$ 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 $\Delta \in \Pi$. Let $\{\varphi_j\}_{j=1}^{K_s}$ be the dual basis for B-splines $\breve{\mathbf{b}}(x) := \mathbf{b}_s(x;\Delta)/\sqrt{J}$, 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 reproduces polynomials,

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

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 $\breve{b}_{s,j}(x)$, we have $|\varphi_j\varrho| \lesssim \|\varrho\|_{L_\infty[\mathcal{I}_j]}$ where $\mathcal{I}_j$ denotes the support of $\breve{b}_{s,j}(x)$ and $\|\cdot\|_{L_\infty[\mathcal{I}_j]}$ denotes the sup-norm on $\mathcal{I}_j$. All points within such $\mathcal{I}_j$ should be no greater than $(p+1)\max_{1\le j\le J} h_j(\Delta)$ away from x_0 where $h_j(\Delta)$ denotes the length of the jth bin in Δ. Hence, $\|\varrho\|_{L_\infty[\mathcal{I}_j]} \lesssim 1$. The desired lower bound follows. The bound on $\Omega(x)$ can be established similarly. □

SA-7.2.3 Proof of Lemma SA-2.3

Proof. By Lemmas SA-1.2, SA-1.3 and SA-2.1, $\sup_{x\in\mathcal{X}} \|\widehat{\mathbf{b}}^{(v)}(x)'\|_\infty \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$. Define a function class

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

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-1.3, $\mathcal{G}$ 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-7.2.4 Proof of Lemma SA-2.4

Proof. Note that $\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_0(x_i)] = A_1(x) + A_2(x)$, with $A_1(x) := \widehat{\mathbf{b}}_s^{(v)}(x)'(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1})\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_0(x_i)]$ and $A_2(x) := \widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_0(x_i)]$. By definition of $\widehat{r}_0(\cdot)$, we have $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_0(x_i)] = 0$. Define the following function class

$$\mathcal{G} := \Big\{x \mapsto b_{s,l}(x; \Delta)r_0(x; \Delta) : 1 \le l \le K_s, \Delta \in \Pi\Big\}.$$

By Lemma SA-1.4, $\sup_{\Delta\in\Pi} |r_0(x; \Delta)|_\infty \lesssim J^{-p-1}$. Then, $\sup_{g\in\mathcal{G}} |g|_\infty \lesssim J^{-p-1+1/2}$, 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

$$g(x) = b_{s,l}(x; \Delta)(\mu_0(x) - \mathbf{b}_s(x; \Delta)'\boldsymbol{\beta}_0(\Delta)) = b_{s,l}(x; \Delta)\mu_0(x) - \sum_{k=\underline{k}}^{\underline{k}+p} b_{s,l}(x; \Delta)b_{s,k}(x; \Delta)\beta_{0,k}(\Delta)$$

for some $1 \le l, \underline{k} \le K_s$ where $\beta_{0,k}(\Delta)$ denotes the kth element in $\boldsymbol{\beta}_0(\Delta)$. Here we use the sparsity property of the partitioning basis: the summand in the second term is nonzero only if $b_{s,l}(x; \Delta)$ and $b_{s,k}(x; \Delta)$ have overlapping supports. For each l, there are only a finite number of such $b_{s,k}(x; \Delta)$

functions. Then, using the same argument given in the proof of Lemma SA-2.1,

$$N(\mathcal{G}, L_2(\mathbb{Q}), \varepsilon\|\bar{G}\|_{L_2(\mathbb{Q})}) \leq \Big(\frac{J^l}{\varepsilon}\Big)^z$$

for some finite l and z and the envelop $\bar{G} = CJ^{-p-1+1/2}$ for $C > 0$ large enough. By Theorem 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

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

and, by Lemma SA-2.1, $\|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\|_\infty \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$. Then, using the bound on the basis given in Lemma SA-1.3,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}|A_1(x)| \lesssim_{\mathbb{P}} J^v\sqrt{J}\sqrt{\frac{J\log J}{n}}J^{-p-1}\sqrt{\frac{\log J}{n}} = J^{-p-1+v}\frac{J\log J}{n}, \quad \text{and}\\
&\sup_{x\in\mathcal{X}}|A_2(x)| \lesssim_{\mathbb{P}} J^v\sqrt{J}J^{-p-1}\sqrt{\frac{\log J}{n}} = J^{-p-1+v}\sqrt{\frac{J\log J}{n}}.
\end{aligned}$$

These results complete the proof. □

SA-7.2.5 Proof of Lemma SA-2.5

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,

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 = \Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}\mathbf{w}_i\mathbf{w}_j'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{w}_i M_{ij}(\mu_0(x_j) + \epsilon_j)\Big).$$

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

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

We have

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

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, $\frac{\mathbf{H}'\mathbf{M}_{\mathbf{B}}\mathbf{H}}{n} \geq 0$, and $\frac{\mathbf{H}'\mathbf{M}_{\mathbf{B}}\mathbf{V}}{n}$ has mean zero conditional on $\mathbf{X}$ and by Lemma SA-1 of Cattaneo, Jansson, and Newey (2018b),

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

where $\|\cdot\|_F$ denotes the Frobenius norm. Therefore, we conclude that $\frac{\mathbf{W}'\mathbf{M}_{\mathbf{B}}\mathbf{W}}{n} \gtrsim_{\mathbb{P}} 1$.

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 $\mathbf{G} = (\mu_0(x_1), \ldots, \mu_0(x_n))'$ and note that

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

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

To show the second result, note that by Lemmas SA-1.2, SA-1.3 and SA-2.1, $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_s^{(v)}(x)'\|_\infty \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ 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-2.1 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 $\mathbf{w}$, $l = 1, \ldots, d$,

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

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-2.3 and the assumption that $\sup_{x\in\mathcal{X}} \mathbb{E}[|w_{i,l}|^\nu|x_i = x] \lesssim 1$ for all l, we have $\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)(w_{i,l} - \mathbb{E}[w_{i,l}|x_i])]\|_\infty \lesssim_\mathbb{P} \sqrt{\log J/n}$. Thus, the last result follows. □

SA-7.2.6 Proof of Theorem SA-2.1

Proof. The result follows by Lemmas SA-1.4, SA-2.4 and SA-2.5. □

SA-7.2.7 Proof of Corollary SA-2.1

Proof. The result follows by Theorem SA-2.1 and Lemma SA-2.3. □

SA-7.2.8 Proof of Theorem SA-2.2

Proof. Since $\widehat{\epsilon}_i := y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}} = \epsilon_i + \mu_0(x_i) - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) =: \epsilon_i + u_i$, we can write

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

Now we bound each term in the following.

Step 1: For $\mathbf{V}_1$, we further write $u_i = (\mu_0(x_i) - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}) - \mathbf{w}_i'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) =: u_{i1} - u_{i2}$. 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 $\|2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'u_{i1}u_{i2}]\| \le \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(u_{i1}^2+u_{i2}^2)]\|$, it suffices to bound $\mathbf{V}_{11}$ and $\mathbf{V}_{12}$. For $\mathbf{V}_{11}$,

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

where the last inequality holds by Lemma SA-2.1 and Corollary SA-2.1. On the other hand,

$$\|\mathbf{V}_{12}\| = \Big\|\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Big(\sum_{\ell=1}^d w_{i\ell}^2(\widehat{\gamma}_\ell - \gamma_{0,\ell})^2 + \sum_{\ell\neq\ell'} w_{i\ell}w_{i\ell'}(\widehat{\gamma}_\ell - \gamma_{0,\ell})(\widehat{\gamma}_{\ell'} - \gamma_{0,\ell'})\Big)\Big]\Big\|$$

$$\lesssim \Big\|\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Big(\sum_{\ell}^{d} w_{i\ell}^2(\widehat{\gamma}_\ell-\gamma_{0,\ell})^2\Big)\Big]\Big\|$$

by CR-inequality. By Lemma SA-2.5, $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|^2=o_{\mathbb{P}}(J/n)$. 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,

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

where the last step follows by Lemma SA-2.1 and the result given in the next step. In addition,

$$\|\mathbf{V}_{22}\|=\Big\|2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i\sum_{\ell=1}^{d}w_{i,\ell}(\widehat{\gamma}_\ell-\gamma_{0,\ell})]\Big\|.$$

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

Step 3: For $\mathbf{V}_3$, in view of Lemma SA-1.1 and SA-1.2, it suffices to show that

$$\sup_{\Delta\in\Pi}\Big\|\mathbb{E}_n[\mathbf{b}_0(x_i;\Delta)\mathbf{b}_0(x_i;\Delta)'(\epsilon_i^2-\sigma^2(x_i))]\Big\|\lesssim_{\mathbb{P}}\Big(\frac{J\log J}{n^{\frac{\nu-2}{\nu}}}\Big)^{1/2}.$$

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

$$\mathcal{G}=\Big\{(x_1,\varphi_1)\mapsto b_{0,l}(x_1;\Delta)b_{0,k}(x_1;\Delta)\varphi_1:1\leq l\leq J(p+1),1\leq k\leq J(p+1),\Delta\in\Pi\Big\}.$$

Then for $g\in\mathcal{G}$, $\sum_{i=1}^n g(x_1,\varphi_1)=\sum_{i=1}^n g(x_1,\varphi_1^+)+\sum_{i=1}^n g(x_1,\varphi_1^-)$.

Now, for the truncated piece, we have $\sup_{g\in\mathcal{G}}|g(x_1,\varphi_1^-)|\lesssim JM$, and

$$\begin{aligned}\sup_{g\in\mathcal{G}}\mathbb{V}[g(x_1,\varphi_1^-)]&\lesssim\sup_{x\in\mathcal{X}}\mathbb{E}[(\varphi_1^-)^2|x_1=x]\sup_{\Delta\in\Pi}\sup_{1\leq l,k\leq J(p+1)}\mathbb{E}[b_{0,l}^2(x_1;\Delta)b_{0,k}^2(x_1;\Delta)]\\&\lesssim JM\sup_{x\in\mathcal{X}}\mathbb{E}\Big[|\varphi_1|\Big|x_i=x\Big]\lesssim JM.\end{aligned}$$

The VC condition holds by the same argument given in the proof of Lemma SA-2.1. Then, by Proposition 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big|\mathbb{E}_n[g(x_i,\varphi_i^-)]\Big|\Big] \lesssim \Big(\frac{JM\log(JM)}{n}\Big)^{1/2} + \frac{JM\log(JM)}{n}.$$

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

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

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

$$\mathbb{E}[\max_{1\leq i\leq n}|\varphi_i^+|] \lesssim \mathbb{E}[\max_{1\leq i\leq n}\epsilon_i^2] \lesssim n^{2/\nu}, \quad \text{and} \quad \mathbb{E}[\mathbb{E}_n[|\varphi_i^+|]] \leq \mathbb{E}[|\varphi_1|^+|] \lesssim \frac{\mathbb{E}[|\epsilon|^\nu]}{M^{(\nu-2)/2}}.$$

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-LS, $\sup_{x\in\mathcal{X}}\mathbb{E}[\epsilon_i^2|x_i = x] \lesssim 1$. Then, by the same argument given in the proof of Lemma SA-2.1,

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

Then the proof is complete. □

SA-7.2.9 Proof of Theorem SA-2.3

Proof. We first show that for each fixed $x\in\mathcal{X}$,

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

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,

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

Now, using Lemmas SA-1.3, SA-2.1 and SA-2.2,

$$\begin{aligned}\frac{1}{n^{3/2}}\sum_{i=1}^n\mathbb{E}\Big[|a_i\epsilon_i|^3\Big|\mathbf{X}\Big]&\lesssim\bar{\Omega}(x)^{-3/2}\frac{1}{n^{3/2}}\sum_{i=1}^n\mathbb{E}\Big[|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)\epsilon_i|^3\Big|\mathbf{X}\Big]\\&\lesssim\bar{\Omega}(x)^{-3/2}\frac{1}{n^{3/2}}\sum_{i=1}^n|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)|^3\\&\leq\bar{\Omega}(x)^{-3/2}\frac{\sup_{x\in\mathcal{X}}\sup_{z\in\mathcal{X}}|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(z)|}{n^{3/2}}\sum_{i=1}^n|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)|^2\\&\lesssim_{\mathbb{P}}\frac{1}{J^{3/2+3v}}\cdot\frac{J^{1+v}}{\sqrt{n}}\cdot J^{1+2v}\to 0\end{aligned}$$

since $J/n=o(1)$. By Theorem SA-2.2, the above weak convergence still holds if $\bar{\Omega}(x)$ is replaced by $\widehat{\Omega}(x)$. Now, the desired result follows by Lemmas SA-1.4, SA-2.4 and SA-2.5. □

SA-7.2.10 Proof of Corollary SA-2.2

Proof. Note that for a given p, by Theorem SA-2.6, $J_{\texttt{IMSE}}\asymp n^{\frac{1}{2p+3}}$. Then, for $(p+q)$th-order binscatter estimator, $nJ_{\texttt{IMSE}}^{-2p-2q-3}=o(1)$ and $\frac{J_{\texttt{IMSE}}^2\log^2 J_{\texttt{IMSE}}}{n}=o(1)$. Then the conclusion of Theorem SA-2.3 holds for the $(p+q)$th-order binscatter estimator. Then the result immediately follows.

□

SA-7.2.11 Proof of Theorem SA-2.4

Proof. The proof is divided into several steps.

Step 1: Note that

$$\begin{aligned}&\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}-\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\Omega(x)/n}}\right|\\\leq&\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\Omega(x)/n}}\right|\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\Omega}(x)^{1/2}-\Omega(x)^{1/2}}{\widehat{\Omega}(x)^{1/2}}\right|\end{aligned}$$

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

where the last step uses Lemma SA-2.2, Corollary SA-2.1 and Theorem SA-2.2. Then, in view of Lemmas SA-1.4, SA-2.4, SA-2.5 and Theorem SA-2.2 and the rate restriction given in the lemma, we have

$$\sup_{x\in\mathcal{X}} \Big| \frac{\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}} - \frac{\widehat{\mathbf{b}}_s(x)'\widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}} \mathbb{G}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i] \Big| = o_{\mathbb{P}}(a_n^{-1}).$$

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 $\{\sigma^2_{[i]}\}_{i=1}^n$ where $\sigma^2_{[i]} = \sigma^2(x_{(i)})$. Clearly, conditional on $\mathbf{X}$, $\{\epsilon_{[i]}\}_{i=1}^n$ is still an independent mean zero sequence. Then by Assumptions SA-DGP, SA-LS and the result of Sakhanenko (1991), there exists a sequence of i.i.d. standard normal random variables $\{\zeta_{[i]}\}_{i=1}^n$ such that

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

Then, using summation by parts,

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

By Lemmas SA-1.3, SA-2.1 and SA-2.2, $\sup_{x\in\mathcal{X}} \sup_{x_i\in\mathcal{X}} |\mathscr{K}(x, x_i)| \lesssim_{\mathbb{P}} \sqrt{J}$, 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

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

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

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

where we recover the original ordering. Since $\mathbb{G}_n[\widehat{\mathbf{b}}(x_i)\zeta_i\sigma_i]=_{d|\mathbf{X}}\mathbf{N}(0,\bar{\mathbf{\Sigma}})$ ($=_{d|\mathbf{X}}$ denotes "equal in distribution conditional on $\mathbf{X}$"), the above steps construct the following approximating process:

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

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

Step 3: Note that

$$\begin{aligned}\sup_{x\in\mathcal{X}}|\bar{Z}_p(x)-Z_p(x)|\leq&\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}^{(v)}(x)'(\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1})}{\sqrt{\Omega(x)}}\bar{\mathbf{\Sigma}}^{1/2}\mathbf{N}_{K_s}\Big|\\&+\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}^{(v)}(x)'\mathbf{Q}_0^{-1}}{\sqrt{\Omega(x)}}\Big(\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\Big)\mathbf{N}_{K_s}\Big|\\&+\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}_0^{(v)}(x)'(\widehat{\mathbf{T}}_s-\mathbf{T}_s)\mathbf{Q}_0^{-1}}{\sqrt{\Omega(x)}}\mathbf{\Sigma}_0^{1/2}\mathbf{N}_{K_s}\Big|,\end{aligned}$$

where each term on the right-hand side is a mean-zero Gaussian process conditional on $\mathbf{X}$. By Lemmas SA-1.2 and SA-2.1, $\|\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$ 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-2.1 and Theorem X.3.8 of Bhatia (2013), $\|\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$. By Gaussian Maximal Inequality (see, e.g., van der vaart and Wellner, 1996, Corollary 2.2.8),

$$\mathbb{E}\Big[\sup_{x\in\mathcal{X}}|\bar{Z}_p(x)-Z_p(x)|\Big|\mathbf{X}\Big]\lesssim_{\mathbb{P}}\sqrt{\log J}\Big(\|\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\|+\|\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1}\|+\|\widehat{\mathbf{T}}_s-\mathbf{T}_s\|\Big)=o_{\mathbb{P}}(a_n^{-1}),$$

where the last line follows from the imposed rate restriction. The proof is complete. □

SA-7.2.12 Proof of Theorem SA-2.5

Proof. This conclusion follows from Lemmas SA-1.3 and SA-2.1, Theorem SA-2.2 and Gaussian Maximal Inequality as applied in Step 3 in the proof of Theorem SA-2.4. □

SA-7.2.13 Proof of Theorem SA-2.6

Proof. We rely on the following decomposition:

$$\begin{aligned}\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)=&\,\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i]+\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_0(x_i)]+\\&\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0-\mu_0^{(v)}(x)\Big)-\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i'](\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0).\end{aligned}\qquad\text{(SA-7.5)}$$

The proof is divided into several steps.

Step 1: By Lemma SA-2.5, 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-2.1,

$$\int_{\mathcal{X}}\mathbb{V}[A(x)|\mathbf{X}]\omega(x)dx=\frac{1}{n}\,\mathrm{trace}\,\Big(\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\int_{\mathcal{X}}\widehat{\mathbf{b}}_s^{(v)}(x)\widehat{\mathbf{b}}_s^{(v)}(x)'\omega(x)dx\Big)+o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n}\Big).$$

In fact, using the argument given in the proof of Lemma SA-1.3, we also have

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

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

$$\mathscr{V}_n(p,s,v)=J^{-(1+2v)}\,\mathrm{trace}\,\Big(\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\Big)\asymp 1.$$

Step 2: By decomposition (SA-7.5),

$$\begin{aligned}\mathbb{E}[\widehat{\mu}^{(v)}(x)|\mathbf{X},\mathbf{W}]-\mu_0^{(v)}(x)=&\,\widehat{\mathbf{b}}_s(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_0(x_i)]+\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0-\mu_0^{(v)}(x)\Big)\\&-\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i']\mathbb{E}[(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)|\mathbf{X},\mathbf{W}]\\=:&\,\mathfrak{B}_1(x)+\mathfrak{B}_2(x)+\mathfrak{B}_3(x).\end{aligned}$$

By Lemma SA-2.4, $\int_{\mathcal{X}} \mathfrak{B}_1(x)^2\omega(x)dx = o_{\mathbb{P}}(J^{-2p-2+2v})$. By Lemma SA-2.5, $\int_{\mathcal{X}} \mathfrak{B}_3(x)^2\omega(x)dx = o_{\mathbb{P}}(J^{-2p-2+2v})$. By Lemma SA-1.4, $\int_{\mathcal{X}} \mathfrak{B}_2(x)^2\omega(x)dx \lesssim_{\mathbb{P}} J^{-2p-2+2v}$. 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

$$J^{2p+2-2v}\int_{\mathcal{X}}\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0-\mu_0^{(v)}(x)\Big)^2\omega(x)dx \lesssim_{\mathbb{P}} 1.$$

Then, by Lemma SA-6.1 of Cattaneo, Farrell, and Feng (2020), for $s=p$,

$$\sup_{x\in\mathcal{X}}\left|\mu_0^{(v)}(x)-\widehat{\mathbf{b}}_p^{(v)}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta})-\frac{\mu^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}_x^L}{\hat{h}_x}\Big)\right|=o_{\mathbb{P}}(J^{-(p+1-v)}), \quad \text{(SA-7.6)}$$

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 $\mu_0^{(v)}(x)$ in the larger space in terms of sup-norm. Then it follows that

$$\begin{aligned}
&\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0-\mu_0^{(v)}(x)\\
&=\widehat{\mathbf{b}}_s^{(v)}(x)'\Big(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\Big)^{-1}\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\mu_0(x_i)]-\mu_0^{(v)}(x)\\
&=\widehat{\mathbf{b}}_s^{(v)}(x)'\Big(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\Big)^{-1}\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_s(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!}\hat{h}_{x_i}^{p+1}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}_{x_i}^L}{\hat{h}_{x_i}}\Big)\Big]\\
&\qquad-\frac{\mu_0^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}_x^L}{\hat{h}_x}\Big)+o_{\mathbb{P}}(J^{-p-1+v})\\
&=J^{-p-1}\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_0(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}_{x_i}^L}{\hat{h}_{x_i}}\Big)\Big]\\
&\qquad-\frac{J^{-p-1+v}\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}_x^L}{\hat{h}_x}\Big)+o_{\mathbb{P}}(J^{-p-1+v}),
\end{aligned} \quad \text{(SA-7.7)}$$

where the last step uses Lemmas SA-1.1-SA-1.3 and SA-2.1, and $o_{\mathbb{P}}(\cdot)$ holds uniformly over $x\in\mathcal{X}$. Taking integral of the squared bias and using Assumption SA-DGP and Lemmas SA-1.1-SA-1.3 and SA-2.1 again, we have three leading terms:

$$M_1(x):=\int_{\mathcal{X}}\Big(\frac{J^{-p-1+v}\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}_x^L}{\hat{h}_x}\Big)\Big)^2\omega(x)dx$$

$$
\begin{aligned}
&= \frac{J^{-2p-2+2v}|\mathscr{E}_{2p+2-2v}|}{(2p+2-2v)!}\int_{\mathcal{X}}\Big[\frac{\mu_0^{(p+1)}(x)}{f_X(x)^{p+1-v}}\Big]^2\omega(x)dx + o_{\mathbb{P}}(J^{-2p-2+2v}),\\
M_2(x) :=& J^{-2p-2}\int_{\mathcal{X}}\bigg(\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_0(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}^L_{x_i}}{\hat{h}_{x_i}}\Big)\Big]\bigg)^2\omega(x)dx\\
=& J^{-2p-2}\boldsymbol{\xi}_{0,f}'\mathbf{T}_s'\mathbf{Q}_0^{-1}\Big(\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\Big)\mathbf{Q}_0^{-1}\mathbf{T}_s\boldsymbol{\xi}_{0,f} + o_{\mathbb{P}}(J^{-2p-2+2v}),\\
M_3(x) :=& J^{-2p-2+v}\int_{\mathcal{X}}\bigg\{\bigg(\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_0(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}^L_{x_i}}{\hat{h}_{x_i}}\Big)\Big]\bigg)\\
&\qquad\times\frac{\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}^L_x}{\hat{h}_x}\Big)\bigg\}\omega(x)dx\\
=& J^{-2p-2+v}\boldsymbol{\xi}_{0,f}'\mathbf{T}_s'\mathbf{Q}_0^{-1}\mathbf{T}_s\boldsymbol{\xi}_{v,\omega} + o_{\mathbb{P}}(J^{-2p-2+2v}),
\end{aligned}
$$

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 $f_X(\cdot)$ and $\omega(\cdot)$ respectively), we define

$$
\boldsymbol{\xi}_{v,\lambda} = \int_{\mathcal{X}}\mathbf{b}_0^{(v)}(x)\frac{\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\tau^L_x}{h_x}\Big)\lambda(x)dx.
$$

τ_x and h_x are defined the same way as $\hat{\tau}_x$ and $\hat{h}_x$, but are based on Δ_0, the partition using population quantiles. Therefore, the leading terms now only rely on the non-random partition Δ_0 as well as other deterministic functions, which are simply equivalent to the leading bias if we repeat the above derivation but set $\widehat{\Delta} = \Delta_0$. Then the proof is complete. □

SA-7.2.14 Proof of Corollary SA-2.3

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: $\mathbf{Q}_0$, $\boldsymbol{\Sigma}_0$ 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 $\omega(x)$, $f_X(x)$ and $\sigma^2(x)$, these matrices can be further approximated:

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

where

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

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

"$o_{\mathbb{P}}(\cdot)$" 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

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

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

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

$$\widetilde{\mathbf{Q}}_l = h_l^{-2v}\Big(\int_0^1 \boldsymbol{\psi}(z)\boldsymbol{\psi}(z)'dz\Big)^{-1}\int_0^1 \boldsymbol{\psi}^{(v)}(z)\boldsymbol{\psi}^{(v)}(z)'dz$$

for $\boldsymbol{\psi}(z) = (1, z, \ldots, z^p)$. Employing Lemma SA-1.1 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 $\mathscr{V}(p,0,v) := \operatorname{trace}\Big\{\Big(\int_0^1 \boldsymbol{\psi}(z)\boldsymbol{\psi}(z)'dz\Big)^{-1}\int_0^1 \boldsymbol{\psi}^{(v)}(z)\boldsymbol{\psi}^{(v)}(z)'dz\Big\}\int_{\mathcal{X}}\sigma^2(x)f_X(x)^{2v}\omega(x)dx$.

Step 2: Now, consider the special case in which $s = 0$. By Lemma A.3 of Cattaneo, Farrell, and Feng (2020), we can construct an L_∞ approximation error

$$r_\infty^{(v)}(x;\widehat{\Delta}) := \mu_0^{(v)}(x) - \widehat{\mathbf{b}}_0^{(v)}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta}) = \frac{\mu_0^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{B}_{p+1-v}\Big(\frac{x-\hat{\tau}_x^L}{\hat{h}_x}\Big) + o_{\mathbb{P}}(J^{-(p+1-v)}),$$

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,

$$
\begin{aligned}
&\max_{1\leq j\leq J(p+1)} |\mathbb{E}_{\widehat{\Delta}}[\widehat{b}_{0,j}(x) r_\infty(x;\widehat{\Delta})]| \\
&= \max_{1\leq j\leq J(p+1)} \Big| \int_{\mathcal{X}} \widehat{b}_{0,j}(x) r_\infty(x;\widehat{\Delta}) f_X(x) dx \Big| \\
&= \max_{1\leq j\leq J(p+1)} \Big| \int_{\hat{\tau}_x^L}^{\hat{\tau}_x^L+\hat{h}_x} \widehat{b}_{0,j}(x) r_\infty(x;\widehat{\Delta}) f_X(\hat{\tau}_x^L) dx \Big| + o_{\mathbb{P}}(J^{-p-1-1/2}) \\
&= \max_{1\leq j\leq J(p+1)} \Big| f_X(\hat{\tau}_x^L) \frac{\mu_0^{(p+1)}(x) J^{-p-1}}{(p+1)!} \int_{\hat{\tau}_x^L}^{\hat{\tau}_x^L+\hat{h}_x} \widehat{b}_{0,j}(x) \mathscr{B}_{p+1}\Big(\frac{x-\hat{\tau}_x^L}{\hat{h}_x}\Big) dx \Big| + o_{\mathbb{P}}(J^{-p-1-1/2}) \\
&= o_{\mathbb{P}}(J^{-p-1-1/2}),
\end{aligned}
$$

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 $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x) r_0(x;\widehat{\Delta})] = 0$,

$$
\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}(x)(r_0(x;\widehat{\Delta}) - r_\infty(x;\widehat{\Delta})]\| = \|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}(x)\widehat{\mathbf{b}}(x)'(\boldsymbol{\beta}_\infty(\widehat{\Delta}) - \boldsymbol{\beta}_0(\widehat{\Delta}))]\| = o_{\mathbb{P}}(J^{-p-1}).
$$

By Lemma SA-2.1, $\lambda_{\min}(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)]') \gtrsim_{\mathbb{P}} 1$, and thus $\|\boldsymbol{\beta}_\infty(\widehat{\Delta}) - \boldsymbol{\beta}_0(\widehat{\Delta})\| = o_{\mathbb{P}}(J^{-p-1})$. Then,

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

Therefore, we can represent the leading term in the integrated squared bias by L_∞ approximation error: $\int_{\mathcal{X}} \mathfrak{B}_2(x)^2\omega(x)dx = \int_{\mathcal{X}}(\mu_0^{(v)}(x) - \widehat{\mathbf{b}}^{(v)}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta}))^2\omega(x)dx + o_{\mathbb{P}}(J^{-2p-2+2v})$. Finally, using the results given in Lemma SA-1.1, change of variables and the definition of Riemann integral, we conclude that

$$
\int_{\mathcal{X}} \Big(\mathbb{E}[\widehat{\mu}^{(v)}(x)|\mathbf{X},\mathbf{W}] - \mu_0^{(v)}(x)\Big)^2 \omega(x) dx = J^{-2(p+1-v)}\mathscr{B}(p,0,v) + o_{\mathbb{P}}(J^{-2p-2+2v})
$$

where

$$\mathscr{B}(p,0,v)=\frac{\int_0^1[\mathscr{B}_{p+1-v}(z)]^2dz}{((p+1-v)!)^2}\int_{\mathcal{X}}\frac{[\mu_0^{(p+1)}(x)]^2}{f_X(x)^{2p+2-2v}}\omega(x)dx.$$

Then the proof is complete. □

SA-7.2.15 Proof of Corollary SA-2.4

Proof. For $v>0$, the desired result is equivalent to that given in Theorem SA-2.6. For $v=0$, we will have two additional terms $\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$ and $(\widehat{\mathbf{w}}-\mathbf{w})'\boldsymbol{\gamma}_0$ in Equation (SA-7.5). By Assumption, $\widehat{\mathbf{w}}-\mathbf{w}=o_{\mathbb{P}}(\sqrt{J/n}+J^{-p-1})$, and thus $(\widehat{\mathbf{w}}-\mathbf{w})'\boldsymbol{\gamma}_0$ as a (conditional) bias term is of higher order. The term $\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$ can treated the same way as that we analyze $\widehat{\mathbf{b}}_s(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i'](\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$. By Lemma SA-2.5, it is also of higher order. Then, the proof is complete. □

SA-7.3 Proof for Section SA-3

SA-7.3.1 Proof of Lemma SA-3.1

Proof. We write $\Psi_{i,1}:=\Psi_1(x_i,\mathbf{w}_i;\eta_i)$. By Assumption SA-GL(iv) and (v), $\Psi_{i,1}\eta_{i,1}^2$ is bounded and bounded away from zero uniformly over $1\leq i\leq n$. Thus, $\mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)']\lesssim\mathbf{Q}_0\lesssim\mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)']$. Then, the bounds on the minimum and maximum eigenvalues of $\mathbf{Q}_0$ follow from Lemma SA-2.1.

Next, we prove the convergence of $\bar{\mathbf{Q}}$. Again, in view of Lemma SA-1.2, it suffices to show the convergence for $s=0$. Let $\mathcal{A}_n$ denote the event on which $\widehat{\Delta}\in\Pi$. By Lemma SA-1.1, $\mathbb{P}(\mathcal{A}_n^c)=o(1)$. On $\mathcal{A}_n$, as in the proof of Lemma SA-2.1,

$$\begin{aligned}&\Big\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'\Psi_{i,1}\eta_{i,1}^2]-\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'\Psi_{i,1}\eta_{i,1}^2]\Big\|\\&\leq\sup_{\Delta\in\Pi}\|\mathbb{E}_n[\mathbf{b}_0(x_i;\Delta)\mathbf{b}_0(x_i;\Delta)'\Psi_{i,1}\eta_i^2]-\mathbb{E}[\mathbf{b}_0(x_i;\Delta)\mathbf{b}_0(x_i;\Delta)'\Psi_{i,1}\eta_i^2]\|_\infty.\end{aligned}$$

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

$$|a_{kl}|=\Big|\mathbb{E}_n[b_{0,k}(x_i;\Delta)b_{0,l}(x_i;\Delta)'\Psi_{i,1}\eta_{i,1}^2]-\mathbb{E}[b_{0,k}(x_i;\Delta)b_{0,l}(x_i;\Delta)'\Psi_{i,1}\eta_{i,1}^2]\Big|.$$

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

$$\mathcal{G}=\Big\{(x_1,\mathbf{w}_1)\mapsto b_{0,k}(x_1;\Delta)b_{0,l}(x_1;\Delta)\Psi_i\eta_{i,1}^2:1\leq k,l\leq J(p+1),\Delta\in\Pi\Big\}.$$

We have $\sup_{g\in\mathcal{G}}|g|_\infty\lesssim J$ and $\sup_{g\in\mathcal{G}}\mathbb{V}[g]\leq\sup_{g\in\mathcal{G}}\mathbb{E}[g^2]\lesssim J$, by Assumption SA-GL. Then, by the same argument given in the proof of Lemma SA-2.1,

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

implying $\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'\Psi_{i,1}\eta_{i,1}^2]-\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'\Psi_{i,1}\eta_{i,1}^2]\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$.

Now, let α_{kl} be a generic (k,l)th entry of $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'\Psi_{i,1}\eta_{i,1}^2]/J-\mathbb{E}[\mathbf{b}_0(x_i)\mathbf{b}_0(x_i)'\Psi_{i,1}\eta_{i,1}^2]/J$. By definition, it is either equal to zero or

$$\begin{aligned}\alpha_{kl}=&\int_{\widehat{\mathcal{B}}_j}\Big(\frac{x-\hat{\tau}_j}{\hat{h}_j}\Big)^\ell\varphi(x_i)f_X(x)dx-\int_{\mathcal{B}_j}\Big(\frac{x-\tau_j}{h_j}\Big)^\ell\varphi(x_i)f_X(x)dx\\=&\hat{h}_j\int_0^1 z^\ell\varphi(z\hat{h}_j+\hat{\tau}_j)f_X(z\hat{h}_j+\hat{\tau}_j)dz-h_j\int_0^1 z^\ell\varphi(zh_j+\tau_j)f_X(zh_j+\tau_j)dz\\=&(\hat{h}_j-h_j)\int_0^1 z^\ell\varphi(z\hat{h}_j+\hat{\tau}_j)f_X(z\hat{h}_j+\hat{\tau}_j)dz\\&+h_j\int_0^1 z^\ell\Big(\varphi(z\hat{h}_j+\hat{\tau}_j)f_X(z\hat{h}_j+\hat{\tau}_j)-\varphi(zh_j+\tau_j)f_X(zh_j+\tau_j)\Big)dz\end{aligned}$$

for some $1\leq j\leq J$ and $0\leq\ell\leq 2p$ and $\varphi(x_i)=\mathbb{E}[\varkappa(x_i,\mathbf{w}_i)|x_i]$. By Assumptions SA-DGP and SA-GL and the argument in the proof of Lemma SA-2.1,

$$\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'\Psi_{i,1}\eta_{i,1}^2]-\mathbf{Q}_0\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}.$$

Given the above fact, it follows that $\|\bar{\mathbf{Q}}^{-1}\|\lesssim_{\mathbb{P}}1$. Notice that $\bar{\mathbf{Q}}$ and $\mathbf{Q}_0$ are banded matrices with finite band width. Then, the bounds on the elements of $\bar{\mathbf{Q}}^{-1}$, $\|\bar{\mathbf{Q}}\|_\infty$ and $\|\bar{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1}\|_\infty$ hold by Theorem 2.2 of Demko (1977). This completes the proof. □

SA-7.3.2 Proof of Lemma SA-3.2

Proof. Since $\mathbb{E}[\psi(\epsilon_i)^2|x_i=x,\mathbf{w}_i=\mathbf{w}]$ and $(\eta^{(1)}(\mu_0(x)+\mathbf{w}'\boldsymbol{\gamma}_0))^2$ is bounded and bounded away from zero uniformly over $x\in\mathcal{X}$ and $\mathbf{w}\in\mathcal{W}$, $\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\lesssim\bar{\mathbf{\Sigma}}\lesssim\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']$. Then, the desired results follow by the same argument given in the proof of Lemma SA-2.2. $\square$

SA-7.3.3 Proof of Lemma SA-3.3

Proof. By Lemmas SA-1.2, SA-1.3 and SA-3.1, $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}^{(v)}(x)'\|_\infty \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\bar{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$. Define the following function class

$$\mathcal{G} = \Big\{(x_1, \mathbf{w}_1, \epsilon_1) \mapsto b_{0,l}(x_1;\Delta)\eta^{(1)}(\mu_0(x_1) + \mathbf{w}_1'\boldsymbol{\gamma}_0)\psi(\epsilon_1) : 1 \le l \le J(p+1), \Delta \in \Pi\Big\}.$$

The desired result follows by the same argument in the proof of Lemma SA-2.3. $\square$

SA-7.3.4 Proof of Lemma SA-3.4

Proof. Let $\tilde{\epsilon}_i = y_i - \eta(\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)$. We write $\mathfrak{r}(x_i,\mathbf{w}_i,y_i) := \mathfrak{r}(x_i,\mathbf{w}_i,y_i;\widehat{\Delta}) := \eta_{i,1}\psi(\epsilon_i) - \eta^{(1)}(\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)\psi(\widetilde{\epsilon}_i) = A_1(x_i,\mathbf{w}_i,y_i) + A_2(x_i,\mathbf{w}_i,y_i)$ where $A_1(x_i,\mathbf{w}_i,y_i) := A_1(x_i,\mathbf{w}_i,y_i;\widehat{\Delta}) := (\eta_{i,1} - \eta^{(1)}(\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0))\psi(\epsilon_i)$ and $A_2(x_i,\mathbf{w}_i,y_i) := A_2(x_i,\mathbf{w}_i,y_i;\widehat{\Delta}) := \eta^{(1)}(\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)(\psi(\epsilon_i) - \psi(\tilde{\epsilon}_i))$.

First, by Assumption SA-GL and Lemma SA-1.4, $\sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}}|\eta^{(1)}(\mu_0(x)+\mathbf{w}'\boldsymbol{\gamma}_0) - \eta^{(1)}(\widehat{\mathbf{b}}_s(x)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}'\boldsymbol{\gamma}_0)| \lesssim J^{-p-1}$. Also, for every $1 \le l \le K_s$ and $\Delta \in \Pi$,

$$\begin{aligned}
&b_{s,l}(x;\Delta)\Big(\eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0) - \eta(\mathbf{b}_s(x;\Delta)'\boldsymbol{\beta}_0(\Delta) + \mathbf{w}'\boldsymbol{\gamma}_0)\Big)\\
&= b_{s,l}(x;\Delta)\eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0) - b_{s,l}(x;\Delta)\eta\Big(\sum_{k=\underline{k}_l}^{\underline{k}_l+p} b_{s,k}(x;\Delta)\beta_{0,k}(\Delta) + \mathbf{w}'\boldsymbol{\gamma}_0\Big)
\end{aligned}$$

for some $1 \le \underline{k}_l \le K_s$ where $\beta_{0,k}(\Delta)$ denotes the kth element in $\boldsymbol{\beta}_0(\Delta)$. For the function class $\mathcal{G} = \{(x,\mathbf{w},y) \mapsto b_{s,l}(x;\Delta)A_1(x,\mathbf{w},y;\Delta) : 1 \le l \le K_s, \Delta \in \Pi\}$, by the same argument given in the proof of Lemma SA-2.3,

$$\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)A_1(x_i,\mathbf{w}_i,y_i)]\|_\infty \lesssim_{\mathbb{P}} J^{-p-1}\Big(\frac{\log J}{n}\Big)^{1/2}.$$

Next, let $\mathscr{F}_{XW}$ be the σ-field generated by $\{(x_i, \mathbf{w}_i)\}_{i=1}^n$. Note that

$$\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)A_2(x_i, \mathbf{w}_i, y_i)] = \mathbb{E}_n[\mathbb{E}[\widehat{\mathbf{b}}_s(x_i)A_2(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW}]] + \\ \mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)A_2(x_i, \mathbf{w}_i, y_i) - \mathbb{E}[\widehat{\mathbf{b}}_s(x_i)A_2(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW}]\Big].$$

By Assumption SA-GL and Lemma SA-1.4,

$$\mathbb{E}[A_2(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW}] = -\eta^{(1)}(\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)\Psi(x_i; \eta(\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)) \lesssim J^{-p-1}$$

a.s. on $\mathscr{F}_{XW}$. Then, $\|\mathbb{E}_n[\mathbb{E}[\widehat{\mathbf{b}}_s(x_i)A_2(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW}]]\|_\infty \lesssim_{\mathbb{P}} J^{-p-1-1/2}$ by the same argument in the proof of Lemma SA-2.1. On the other hand, define the following function class

$$\mathcal{G} := \Big\{(x, \mathbf{w}, y) \mapsto b_{s,l}(x; \Delta)A_2(x, \mathbf{w}, y; \Delta) : 1 \leq l \leq K_s, \Delta \in \Pi\Big\}.$$

By Assumption SA-GL, $\sup_{g\in\mathcal{G}} \|g\|_\infty \lesssim J^{1/2}$, and $\sup_{g\in\mathcal{G}} \mathbb{V}[g(x_i, \mathbf{w}_i, y_i)] \lesssim_{\mathbb{P}} J^{-p-1}$. By a similar argument given above, this function class is of VC-type. Then, as in the proof of Lemma SA-2.3, by Proposition 6.1 of Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019),

$$\sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^n (g(x_i, \mathbf{w}_i, y_i) - \mathbb{E}[g(x_i, \mathbf{w}_i, y_i)])\Big| \lesssim J^{-\frac{p+1}{2}}\sqrt{\frac{\log J}{n}} + \frac{J^{1/2}\log J}{n}.$$

Collecting these results, we conclude that

$$\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\mathbb{E}[\widehat{\mathbf{b}}_s(x_i)\mathfrak{r}(x_i, \mathbf{w}_i, y_i)] \lesssim_{\mathbb{P}} J^{-p-1+v} + J^{\frac{2v-p-1}{2}}\Big(\frac{J\log J}{n}\Big)^{1/2} + \frac{J^{1+v}\log J}{n}.$$

The proof is complete. □

SA-7.3.5 Proof of Lemma SA-3.5

Proof. By convexity of $\rho(y; \eta(\cdot))$, we only need to consider $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_0 + \varepsilon\boldsymbol{\alpha}/\sqrt{J}$ for any sufficiently small fixed $\varepsilon > 0$ and $\boldsymbol{\alpha} \in \mathbb{R}^{K_s}$ such that $\|\boldsymbol{\alpha}\| = 1$. For notational simplicity, let $\widehat{\mathbf{b}}_i := \widehat{\mathbf{b}}_s(x_i)$. For such choice of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} \in \mathbb{R}^d$,

$$\begin{aligned}\delta_i(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \rho(y_i;\eta(\widehat{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\boldsymbol{\gamma})) - \rho(y_i;\eta(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}))\\ &= \int_0^{\varepsilon\widehat{\mathbf{b}}_i'\boldsymbol{\alpha}/\sqrt{J}} \psi(x_i,\mathbf{w}_i;\eta(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}+t))\eta^{(1)}(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}+t)dt.\end{aligned}$$

Let $\mathscr{F}_{XW}$ be the σ-field generated by $\{(x_i,\mathbf{w}_i)\}_{i=1}^n$. We have

$$\mathbb{E}_n[\delta_i(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})] = \mathbb{E}_n[\delta_i(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})] = \frac{1}{\sqrt{n}}\mathbb{G}_n[\delta_i(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})] + \mathbb{E}_n[\mathbb{E}[\delta_i(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})|\mathscr{F}_{XW}]],$$

where $\mathbb{G}_n[\cdot]$ denotes $\sqrt{n}(\mathbb{E}_n[\cdot]-\mathbb{E}[\cdot|\mathscr{F}_{XW}])$ and $\mathbb{E}[\delta(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})|\mathscr{F}_{XW}] := \mathbb{E}[\delta(\boldsymbol{\beta},\boldsymbol{\gamma})|\mathscr{F}_{XW}]|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}}$, that is, the conditional expectation with $\widehat{\boldsymbol{\gamma}}$ viewed as fixed. By Assumption SA-GL,

$$\begin{aligned}\mathbb{E}[\delta_i(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})|\mathscr{F}_{XW}] &= \int_0^{\varepsilon\widehat{\mathbf{b}}_i'\boldsymbol{\alpha}/\sqrt{J}} \Psi(x_i,\mathbf{w}_i;\eta(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}+t))\eta^{(1)}(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}+t)dt\\ = \int_0^{\varepsilon\widehat{\mathbf{b}}_i'\boldsymbol{\alpha}/\sqrt{J}} &\Psi_1(x_i,\mathbf{w}_i;\xi_{i,t})(\eta(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}+t)-\eta_{i,1})\eta^{(1)}(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}+t)dt \gtrsim_{\mathbb{P}} \varepsilon^2\boldsymbol{\alpha}'\widehat{\mathbf{b}}_i\widehat{\mathbf{b}}_i'\boldsymbol{\alpha}/J,\end{aligned}$$

where $\xi_{i,t}$ is between $\eta(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}+t)$ and $\eta(\mu_0(x_i)+\mathbf{w}_i'\boldsymbol{\gamma}_0)$. Note that we use the fact that $\Psi(x,\mathbf{w}_i,\eta_i)=0$, $\eta(\cdot)$ is strictly monotonic and $\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0=o(1)$. Thus, $\mathbb{E}_n[\mathbb{E}[\delta_i(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})|\mathscr{F}_{XW}]] \gtrsim_{\mathbb{P}} J^{-1}\varepsilon^2$.

On the other hand, let $\mathcal{H} := \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \leq C\mathfrak{r}_\gamma\}$ and define the following function class

$$\mathcal{G} := \{(x_i,\mathbf{w}_i,y_i)\mapsto \delta_i(\boldsymbol{\beta},\boldsymbol{\gamma}) : \boldsymbol{\alpha}\in\mathcal{S}^{K_s}, \boldsymbol{\gamma}\in\mathcal{H}\}.$$

By Assumption SA-GL, $\sup_{g\in\mathcal{G}}|g| \lesssim \varepsilon$, $\sup_{g\in\mathcal{G}}\mathbb{E}_n[\mathbb{E}[g^2|\mathscr{F}_{XW}]] \lesssim J^{-1}\varepsilon^2$, and VC-index of $\mathcal{G}$ is bounded by CK_s for an absolute constant $C>0$. Therefore, by Lemma 22 of Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019) and the rate restriction,

$$\sup_{g\in\mathcal{G}}\Big|\frac{1}{\sqrt{n}}\mathbb{G}_n[\delta_i(\boldsymbol{\beta},\boldsymbol{\gamma})]\Big| \lesssim_{\mathbb{P}} J^{-1}\Big(\frac{J^2\log J}{n}\Big)^{1/2}\varepsilon + J^{-1}\frac{J^2\log J}{n}\varepsilon = o(\varepsilon/J).$$

Thus, for any fixed (sufficiently small) $\varepsilon>0$, $\mathbb{E}_n[\delta_i(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}})]>0$ when n is sufficiently large. Thus, $\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_0\| = o_{\mathbb{P}}(J^{-1/2})$, implying $\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_0\|_\infty = o_{\mathbb{P}}(J^{-1/2})$ immediately. $\square$

SA-7.3.6 Proof of Theorem SA-3.1

Proof. The proof is long. We divide it into several steps.

Step 0: We first prepare some notation and useful facts. To simplify the presentation, in this proof we drop the scaling factor $\sqrt{J}$ in the basis by defining $\breve{\mathbf{b}}_i := (\widehat{b}_{s,1}(x_i), \cdots, \widehat{b}_{s,K_s}(x_i))'/\sqrt{J} := \widehat{\mathbf{b}}_s(x_i)/\sqrt{J}$ and $\breve{\boldsymbol{\beta}}_0 = \sqrt{J}\widehat{\boldsymbol{\beta}}_0$. Throughout the proof, $C, c, C_1, c_1, C_2, c_2, \cdots$ denote (strictly positive) absolute constants, $\mathscr{F}_{XW}$ denotes the σ-field generated by $\{(x_i, \mathbf{w}_i)\}_{i=1}^n$, and $\operatorname{supp}(g(\cdot))$ denotes the support of a generic function $g(\cdot)$.

Moreover, define

$$\mathcal{V} = \{(v_1, \cdots, v_{K_s})' : \exists k \in \{1, \cdots, K_s\}, |v_\ell| \leq \varrho^{|k-\ell|}\varepsilon_n \text{ for } |\ell - k| \leq M_n \text{ and } v_\ell = 0 \text{ otherwise}\},$$
$$\mathcal{H}_l = \{\mathbf{v} \in \mathbb{R}^{K_s} : \|\mathbf{v}\|_\infty \leq r_{l,n}\} \text{ for } l = 1, 2, \quad \text{and} \quad \mathcal{H}_3 = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| \leq r_{3,n}\},$$

where $\varrho \in (0,1)$ is the constant given in Lemma SA-3.1, $r_{1,n} = C_1[(J\log n/n)^{1/2} + J^{-p-1}]$, $\varepsilon_n \leq r_{2,n} = \mathfrak{z}_n[(\frac{J\log n}{n})^{3/4}\sqrt{\log n} + J^{-\frac{p+1}{2}}\sqrt{\frac{J}{n}}\log n + \mathfrak{r}_\gamma]$ for $\mathfrak{z}_n > 0$ such that $r_{2n} \leq c$ for some sufficiently small constant $c > 0$, $r_{3,n} = C\mathfrak{r}_\gamma$, and $M_n = c_1 \log n$. Note that by Assumption SA-GL, $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \in \mathcal{H}_3$ with probability approaching one for C large enough, and by Lemma SA-3.5, $\sqrt{J}\widehat{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_0 \leq c$ with probability approaching one.

For any $\boldsymbol{\beta}_1 \in \mathcal{H}_1, \boldsymbol{\beta}_2 \in \mathcal{H}_2$, $\boldsymbol{\upsilon} \in \mathcal{V}$ and $\boldsymbol{\gamma} := \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1$ with $\boldsymbol{\gamma}_1 \in \mathcal{H}_3$, define

$$\begin{aligned}
\delta_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma}) &= \rho\Big(y_i; \eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}_i'\boldsymbol{\gamma})\Big) - \rho\Big(y_i; \eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\upsilon}) + \mathbf{w}_i'\boldsymbol{\gamma})\Big) \\
&\quad - \Big[\eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}_i'\boldsymbol{\gamma}) - \eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\upsilon}) + \mathbf{w}_i'\boldsymbol{\gamma})\Big] \\
&\qquad \times \psi(y_i; \eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)) \\
&= \int_{-\breve{\mathbf{b}}_i'\boldsymbol{\upsilon}}^0 \Big[\psi\Big(y_i; \eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}_i'\boldsymbol{\gamma} + t)\Big) - \psi\Big(y_i; \eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)\Big)\Big] \\
&\qquad \times \eta^{(1)}\Big(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}_i'\boldsymbol{\gamma} + t\Big)dt.
\end{aligned}$$

Note that $\delta_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma}) \neq 0$ only if $\breve{\mathbf{b}}_i'\boldsymbol{\upsilon} \neq 0$. For each $\boldsymbol{\upsilon} \in \mathcal{V}$, let $\mathcal{J}_{\boldsymbol{\upsilon}} = \{j : \upsilon_j \neq 0\}$. By construction, the cardinality of $\mathcal{J}_{\boldsymbol{\upsilon}}$ is bounded by $2M_n$. We have $\delta_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma}) \neq 0$ only if $\breve{b}_j(x_i) \neq 0$ for some $j \in \mathcal{J}_\upsilon$, which happens only when $x_i \in \operatorname{supp}(\breve{b}_j(\cdot))$ for some $j \in \mathcal{J}_{\boldsymbol{\upsilon}}$. Let $\mathcal{I}_{\boldsymbol{\upsilon}} = \cup_{j\in\mathcal{J}_{\boldsymbol{\upsilon}}} \operatorname{supp}(\breve{b}_j(\cdot))$. Since the basis functions are locally supported, $\mathcal{I}_{\boldsymbol{\upsilon}}$ includes at most c_2M_n

(connected) intervals for all $\boldsymbol{\upsilon} \in \mathcal{V}$. Moreover, at most $c_3 M_n$ basis functions in $\breve{\mathbf{b}}(\cdot)$ have supports overlapping with $\mathcal{I}_{\boldsymbol{\upsilon}}$. Denote the set of indices for such basis functions by $\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}$. Let $\breve{\beta}_{0,j}$, $\beta_{1,j}$ and $\beta_{2,j}$ be the jth entries of $\breve{\boldsymbol{\beta}}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ respectively, and υ_j be the jth entry of $\boldsymbol{\upsilon}$. Based on the above observations, we have $\delta_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma}) \equiv \delta_i(\boldsymbol{\beta}_{1,\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}}, \boldsymbol{\beta}_{2,\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}}, \boldsymbol{\upsilon}, \boldsymbol{\gamma})$ where

$$\begin{aligned}\delta_i(\boldsymbol{\beta}_{1,\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}}, \boldsymbol{\beta}_{2,\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}}, \boldsymbol{\upsilon}, \boldsymbol{\gamma}) := \int_{-\sum\limits_{j\in\mathcal{J}_{\boldsymbol{\upsilon}}} \breve{b}_{i,j}\upsilon_j}^{0} &\Big[\psi\Big(y_i; \eta\Big(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\breve{\beta}_{0,l} + \beta_{1,l} + \beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma} + t\Big)\Big) \\ &-\psi\Big(y_i;\eta\Big(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}\breve{\beta}_{0,l} + \mathbf{w}_i'\boldsymbol{\gamma}_0\Big)\Big)\Big] \times \eta^{(1)}\Big(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\breve{\beta}_{0,l} + \beta_{1,l} + \beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma} + t\Big) dt \mathbb{1}_{i,\boldsymbol{\upsilon}},\end{aligned}$$

$\mathbb{1}_{i,\boldsymbol{\upsilon}} = \mathbb{1}(x_i \in \mathcal{I}_{\boldsymbol{\upsilon}})$, and $\boldsymbol{\beta}_{1,\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}}$ and $\boldsymbol{\beta}_{2,\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}}$ respectively denote the subvectors of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ whose indices belong to $\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}$. Accordingly, for $\widetilde{\boldsymbol{\beta}}_1 \in \mathbb{R}^{c_3 M_n}, \widetilde{\boldsymbol{\beta}}_2 \in \mathbb{R}^{c_3 M_n}$, define the following function class

$$\mathcal{G} = \Big\{(x_i, \mathbf{w}_i, y_i) \mapsto \delta_i(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma}) : \boldsymbol{\upsilon} \in \mathcal{V}, \|\widetilde{\boldsymbol{\beta}}_1\|_\infty \leq r_{1,n}, \|\widetilde{\boldsymbol{\beta}}_2\|_\infty \leq r_{2,n}, \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 \in \mathcal{H}_3\Big\}.$$

Step 1: We bound $\sup_{g\in\mathcal{G}} |\mathbb{E}_n[g(x_i, \mathbf{w}_i, y_i)] - \mathbb{E}[g(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW}]|$ in this step. Let $a_i(t) := \eta(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}'\breve{\beta}_{0,l} + \mathbf{w}_i'\boldsymbol{\gamma}_0 + t)$. Define $\underline{a}_i = \min\{a_i(0), a_i(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\beta_{1,l}+\beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma}_1), a_i(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\beta_{1,l}+\beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma}_1 + \sum_{j\in\mathcal{J}_{\boldsymbol{\upsilon}}} \breve{b}_{i,j}\upsilon_j)\}$, and $\bar{a}_i = \max\{a_i(0), a_i(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\beta_{1,l}+\beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma}_1), a_i(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\beta_{1,l}+\beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma}_1 + \sum_{j\in\mathcal{J}_{\boldsymbol{\upsilon}}} \breve{b}_{i,j}\upsilon_j)\}$. Consider the following two cases.

First, suppose that $(y_i - \bar{a}_i, y_i - \underline{a}_i)$ does not contain any discontinuity points. By Assumption SA-GL,

$$\Big|\psi(y_i; a_i(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\beta_{1,l} + \beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma} + t)) - \psi(y_i; a_i(0))\Big| \lesssim r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n}$$

for any t in the designated range. Second, if $(y_i - \bar{a}_i, y_i - \underline{a}_i)$ contains at least one discontinuity point, say $\jmath$. For any t on the interval of integration, by Assumption SA-DGP,

$$\Big|\psi(y_i; a_i(\sum_{l\in\bar{\mathcal{J}}_{\boldsymbol{\upsilon}}} \breve{b}_{i,l}(\beta_{1,l} + \beta_{2,l}) + \mathbf{w}_i'\boldsymbol{\gamma} + t)) - \psi(y_i; a_i(0))\Big| \lesssim 1 + r_{3,n}$$

for any $(x_i, \mathbf{w}_i, y_i)$, and in this case $y_i \in (\jmath + \underline{a}, \jmath + \bar{a})$. Note that by construction, for each $\boldsymbol{\upsilon} \in \mathcal{V}$, there exists some $k_{\boldsymbol{\upsilon}}$ such that $|\upsilon_l| \leq \varrho^{|l-k_{\boldsymbol{\upsilon}}|}\varepsilon_n$ for $|l - k_{\boldsymbol{\upsilon}}| \leq M_n$. Therefore, we can further write $\mathbb{1}_{i,\boldsymbol{\upsilon}} = \sum_{j:\widehat{\mathcal{B}}_j \subset \mathcal{I}_{\boldsymbol{\upsilon}}} \mathbb{1}_{i,\boldsymbol{\upsilon},j}$ where each $\mathbb{1}_{i,\boldsymbol{\upsilon},j}$ is an indicator of the subinterval involved in $\mathcal{I}_{\boldsymbol{\upsilon}}$, and the above facts imply that for any $x_i \in \widehat{\mathcal{B}}_l \subset \mathcal{I}_{\boldsymbol{\upsilon}}$ for some l,

$$\mathbb{V}[\delta_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma})|\mathscr{F}_{XW}] \lesssim \varrho^{2|l-k_{\boldsymbol{\upsilon}}|}\varepsilon_n^2(r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n})(|\eta_{i,1}| + r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n}).$$

In addition, since $\delta_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma}) \neq 0$ only if $x_i \in \mathcal{I}_{\boldsymbol{\upsilon}}$, for all $g \in \mathcal{G}$,

$$\mathbb{E}_n[\mathbb{V}[g(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW}]] \lesssim \varepsilon_n^2(r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n}) \sum_{l:\widehat{\mathcal{B}}_l \subset \mathcal{I}_{\boldsymbol{\upsilon}}} \mathbb{E}_n[\mathbb{1}_{i,\boldsymbol{\upsilon},l}]\varrho^{2|l-k_{\boldsymbol{\upsilon}}|}.$$

Note that this inequality holds for any event in $\mathscr{F}_{XW}$. Define an event $\mathcal{A}_1$ on which $\sup_{1\leq j\leq J} \mathbb{E}_n[\mathbb{1}_{i,j}] \leq C_2 J^{-1}$ for some large enough $C_2 > 0$ where $\mathbb{1}_{i,j} = \mathbb{1}(x_i \in \widehat{\mathcal{B}}_j)$. By the argument in Lemma SA-2.1, $\mathbb{P}(\mathcal{A}_1^c) \to 0$. On $\mathcal{A}_1$,

$$\bar{\sigma}^2 := \sup_{g\in\mathcal{G}} \mathbb{E}_n[\mathbb{V}[g(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW}]] \lesssim \varepsilon_n^2 J^{-1}(r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n}).$$

On the other hand,

$$\bar{G} := \sup_{g\in\mathcal{G}} |g(x_i, \mathbf{w}_i, y_i)| \lesssim \varepsilon_n(1 + r_{3,n})(|\eta_{i,1}| + r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n}).$$

Also, for any $g, \tilde{g} \in \mathcal{G}$, denote the corresponding parameters defining g and $\tilde{g}$ by $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\upsilon}, \boldsymbol{\gamma})$ and $(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \tilde{\boldsymbol{\upsilon}}, \tilde{\boldsymbol{\gamma}})$. We have

$$\begin{aligned}
\tilde{g}(x_i, \mathbf{w}_i, y_i) - g(x_i, \mathbf{w}_i, y_i) = &\int_0^{\varLambda_1} \Big[\psi(y_i; \eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}_i'\boldsymbol{\gamma} + t)) \\
&\quad - \psi(y_i; \eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\gamma_0))\Big] \times \eta^{(1)}(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}_i'\boldsymbol{\gamma} + t)dt \\
&- \int_0^{\varLambda_2} \Big[\psi(y_i; \eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\upsilon}) + \mathbf{w}_i'\boldsymbol{\gamma} + t)) \\
&\quad - \psi(y_i; \eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\gamma_0))\Big] \times \eta^{(1)}(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\upsilon}) + \mathbf{w}_i'\boldsymbol{\gamma} + t)dt \\
\lesssim\ &(1 + \Lambda_1 + \Lambda_2)(|\eta_{i,1}| + r_{1,n} + r_{2,n} + \varLambda_1 + \varLambda_2 + r_{3,n}) \\
&\times (\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_\infty + \|\tilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)\|_\infty + \|\tilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}\|_\infty + \|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|),
\end{aligned}$$

where $\Lambda_1 = \breve{\mathbf{b}}_i'(\tilde{\boldsymbol{\beta}}_1+\tilde{\boldsymbol{\beta}}_2-\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)+\mathbf{w}_i'(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma})$ and $\Lambda_2 = \Lambda_1-\breve{\mathbf{b}}_i'(\tilde{\boldsymbol{v}}-\boldsymbol{v})$. Based on these observations,

$$\|\bar{G}\|_{\bar{\mathbb{P}},2}\int_0^{\frac{\bar{\sigma}}{\|\bar{G}\|_{\bar{\mathbb{P}},2}}}\sqrt{1+\sup_{\mathbb{Q}}\log N(\mathcal{G},L_2(\mathbb{Q}),t\|\bar{G}\|_{\mathbb{Q},2})}dt \lesssim \bar{\sigma}\Big(\sqrt{\log J}+\sqrt{\log n\log\frac{1}{\bar{\sigma}}}\Big)\lesssim \bar{\sigma}\log n,$$

where the supremum is taken over all finite discrete probability measures $\mathbb{Q}$. Then, by Lemma SA-3.6,

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}}|\mathbb{G}_n[g(x_i,\mathbf{w}_i,y_i)]|\Big|\mathscr{F}_{XW}\Big]\lesssim \bar{\sigma}\log n+\frac{\sqrt{\mathbb{E}[\bar{\bar{G}}^2]}\log^2 n}{\sqrt{n}},$$

where $\bar{\bar{G}} = \max_{1\le i\le n}\bar{G}(x_i,\mathbf{w}_i,y_i)$. Note that $(\mathbb{E}[\bar{\bar{G}}^2])^{1/2}\lesssim \varepsilon_n$.

Therefore, on $\mathcal{A}_1$ (whose probability approaches one),

$$\begin{aligned}&\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V},\boldsymbol{\gamma}_1\in\mathcal{H}_3}\Big|\mathbb{E}_n[\delta_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v},\boldsymbol{\gamma})]-\mathbb{E}_n[\mathbb{E}[\delta_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v},\boldsymbol{\gamma})|\mathscr{F}_{XW}]]\Big|\\ &\lesssim\Big(J^{-1}\varepsilon_n\sqrt{\mathfrak{L}_n}\sqrt{\frac{J}{n}}\log n+\frac{\varepsilon_n\log^2 n}{n}\Big)\end{aligned}$$

for $\mathfrak{L}_n = r_{1,n}+r_{2,n}+r_{3,n}+\varepsilon_n$.

Step 2: For $\widetilde{\mathbf{Q}} := \mathbb{E}_n[\breve{\mathbf{b}}_i\breve{\mathbf{b}}_i'\Psi^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)(\eta^{(1)}(\breve{\mathbf{b}}_i\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))^2]$, by Assumption SA-GL and the same argument in the proof of Lemma SA-3.1, $\|\bar{\mathbf{Q}}-\widetilde{\mathbf{Q}}\|_\infty\vee\|\bar{\mathbf{Q}}-\widetilde{\mathbf{Q}}\|\lesssim J^{-p-1}J^{-1}$. Therefore,

$$\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V}}|\boldsymbol{v}'(\widetilde{\mathbf{Q}}-\bar{\mathbf{Q}})(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)|\lesssim J^{-p-2}\varepsilon_n(r_{1,n}+r_{2,n}).$$

In addition, by Lemmas SA-3.3 and SA-3.4,

$$\bar{\boldsymbol{\beta}} = -\bar{\mathbf{Q}}^{-1}\mathbb{E}_n[\breve{\mathbf{b}}_i\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))]\le r_{1,n}$$

with probability approaching one for C_1 large enough.

Step 3: By Taylor expansion, we have

$$\begin{aligned}&\mathbb{E}_n[\mathbb{E}[\delta_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v},\boldsymbol{\gamma})|\mathscr{F}_{XW}]]\\ &=\mathbb{E}_n\Big[\int_{-\breve{\mathbf{b}}_i'\boldsymbol{v}}^0\Big\{\Psi(x_i,\mathbf{w}_i;\eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}+t))\\ &\qquad-\Psi(x_i,\mathbf{w}_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\Big\}\times\eta^{(1)}\Big(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}+t\Big)dt\Big]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_n\Big[\int_{-\breve{\mathbf{b}}_i'\boldsymbol{\upsilon}}^{0} \Big\{\Psi_1(x_i,\mathbf{w}_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\Big(\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}_1+t)\\
&\qquad\qquad\qquad\qquad\qquad\qquad + \eta^{(2)}(\xi_{i,t})(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}_1+t)^2\Big)\\
&\qquad\qquad + \Psi_2(\tilde{\xi}_{i,t})\Big(\eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}+t)-\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)\Big)^2\Big\}\\
&\qquad\qquad\quad \times\Big(\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)+\eta^{(2)}(\check{\xi}_{i,t})(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}_1+t)\Big)\Big]\\
&= \boldsymbol{\upsilon}'\widetilde{\mathbf{Q}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\boldsymbol{\upsilon}'\mathbb{E}_n[\mathbf{b}_i\widetilde{\varkappa}_i\mathbf{w}_i']\boldsymbol{\gamma}_1-\frac{1}{2}\boldsymbol{\upsilon}\widetilde{\mathbf{Q}}\boldsymbol{\upsilon}+\mathrm{I}+\mathrm{II}+\mathrm{III},
\end{aligned}
$$

where $\xi_{i,t}$ and $\check{\xi}_{i,t}$ are between $\breve{\mathbf{b}}_i\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0$ and $\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}+t$, $\tilde{\xi}_{i,t}$ is between $\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)$ and $\eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}+t)$, $\Psi_2(x,\mathbf{w};\tau)=\frac{\partial^2}{\partial\tau^2}\Psi(x,\mathbf{w};\tau)$, $\widetilde{\varkappa}_i=\Psi_1(x_i,\mathbf{w}_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)$, $\boldsymbol{\upsilon}'\mathbb{E}_n[\mathbf{b}_i\widetilde{\varkappa}_i\mathbf{w}_i']\boldsymbol{\gamma}_1\lesssim\varepsilon_n r_{3,n}/J$, $-\frac{1}{2}\boldsymbol{\upsilon}\widetilde{\mathbf{Q}}\boldsymbol{\upsilon}\lesssim\varepsilon_n^2/J$, and I,II, and III are defined and bounded as follows:

$$
\begin{aligned}
\mathrm{I} &= \mathbb{E}_n\Big[\int_{-\breve{\mathbf{b}}_i'\boldsymbol{\upsilon}}^{0}\Psi_1(x_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)\\
&\qquad\qquad\times\eta^{(2)}(\check{\xi}_{i,t})(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}_1+t)^2 dt\mathbb{1}_{i,\boldsymbol{\upsilon}}\Big]\lesssim\varepsilon_n J^{-1}(r_{1,n}+r_{2,n}+\varepsilon_n+r_{3,n})^2,\\
\mathrm{II} &= \mathbb{E}_n\Big[\int_{-\breve{\mathbf{b}}_i'\boldsymbol{\upsilon}}^{0}\Psi_1(x_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0)+\mathbf{w}_i'\boldsymbol{\gamma}_0)\eta^{(2)}(\xi_{i,t})(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i\boldsymbol{\gamma}_1+t)^2\\
&\qquad\qquad\times\eta^{(1)}\Big(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}+t\Big)dt\mathbb{1}_{i,\boldsymbol{\upsilon}}\Big]\lesssim\varepsilon_n J^{-1}(r_{1,n}+r_{2,n}+\varepsilon_n+r_{3,n})^2,\\
\mathrm{III} &= \mathbb{E}_n\Big[\int_{-\breve{\mathbf{b}}_i'\boldsymbol{\upsilon}}^{0}\Psi_2(\tilde{\xi}_{i,t})\Big(\eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i\boldsymbol{\gamma}+t)-\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)\Big)^2\\
&\qquad\qquad\times\eta^{(1)}\Big(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}+t\Big)dt\mathbb{1}_{i,\boldsymbol{\upsilon}}\Big]\lesssim\varepsilon_n J^{-1}(r_{1,n}+r_{2,n}+\varepsilon_n+r_{3,n})^2.
\end{aligned}
$$

These bounds hold uniformly for $\boldsymbol{\upsilon}\in\mathcal{V}$, $\boldsymbol{\beta}_1\in\mathcal{H}_1$, $\boldsymbol{\beta}_2\in\mathcal{H}_2$ and $\boldsymbol{\gamma}_1\in\mathcal{H}_3$ (that is, uniformly over the function class $\mathcal{G}$), and on an event $\mathcal{A}_1\cap\mathcal{A}_2$ where $\mathcal{A}_2=\{\lambda_{\max}(\widetilde{\mathbf{Q}})\leq c_4J^{-1}\}$ for some large enough $c_4>0$. Note that $\mathbb{P}(\mathcal{A}_1\cap\mathcal{A}_2)\to 1$ by Lemma SA-3.1.

Step 4: By Assumption SA-GL and Taylor's expansion,

$$
\begin{aligned}
\mathrm{IV} &= \mathbb{E}_n\Big[\Big(\eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma})-\eta(\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\upsilon})+\mathbf{w}_i'\boldsymbol{\gamma})\Big)\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\Big]\\
&\qquad-\mathbb{E}_n\Big[\boldsymbol{\upsilon}'\breve{\mathbf{b}}_i\psi(y_i,\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)\Big]\\
&= \mathbb{E}_n\Big[\boldsymbol{\upsilon}'\breve{\mathbf{b}}_i\psi(y_i,\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\Big(\eta^{(2)}(\xi_i)(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\upsilon})+\mathbf{w}_i\boldsymbol{\gamma}_1)+\eta^{(2)}(\tilde{\xi}_i)\boldsymbol{\upsilon}'\breve{\mathbf{b}}_i\Big)\Big]
\end{aligned}
$$

$$\lesssim J^{-1}((J\log n/n)^{1/2}+J^{-p-1})(\varepsilon_n+r_{1,n}+r_{2,n}+r_{3,n})\varepsilon_n,$$

where ξ_i is between $\widehat{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0$ and $\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\upsilon})+\mathbf{w}_i'\boldsymbol{\gamma}$ and $\tilde{\xi}_i$ is between $\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\mathbf{w}_i'\boldsymbol{\gamma}$ and $\breve{\mathbf{b}}_i'(\breve{\boldsymbol{\beta}}_0+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\upsilon})+\mathbf{w}_i'\boldsymbol{\gamma}$. The last line holds on the event

$$\begin{aligned}\mathcal{A}_3=\Big\{\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{\upsilon}\in\mathcal{V},\boldsymbol{\gamma}_1\in\mathcal{H}_3}\bigg(&\Big\|\mathbb{E}_n\Big[\breve{\mathbf{b}}_i\breve{\mathbf{b}}_i'\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\eta^{(2)}(\xi_i)\Big]\Big\|_\infty+\\&\Big\|\mathbb{E}_n\Big[\breve{\mathbf{b}}_i\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\eta^{(2)}(\tau_i)\mathbf{w}_i\Big]\Big\|_\infty\bigg)\lesssim J^{-1}\Big(\Big(\frac{J\log n}{n}\Big)^{1/2}+J^{-p-1}\Big)\Big\},\end{aligned}$$

where $\tau_i=\xi_i$ (or $\tilde{\xi}_i$), which only depends on x_i and $\mathbf{w}_i$ for each $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\upsilon}$, and $\boldsymbol{\gamma}$. Note that $\mathbb{E}[\psi(y_i,\eta_i)|\mathscr{F}_{XW}]=0$ and $\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0-\mu_0(x_i)\lesssim J^{-p-1}$. Then, we can use the argument in the proof of Lemmas SA-3.3 and SA-3.4 again to obtain $\mathbb{P}(\mathcal{A}_3)\to 1$ by choosing $C_3>0$ sufficiently large.

Step 5: Let $\bar{\boldsymbol{\upsilon}}=c_5\varepsilon_nJ^{-1}[\bar{\mathbf{Q}}^{-1}]_{k\cdot}$ for some k such that $|\beta_{2,k}|=\|\boldsymbol{\beta}_2\|_\infty$ for some $c_5>0$ where $[\bar{\mathbf{Q}}^{-1}]_{k\cdot}$ denotes the kth row of $\bar{\mathbf{Q}}^{-1}$. Take $\boldsymbol{\upsilon}=(\upsilon_1,\cdots,\upsilon_{K_s})$ where $\upsilon_j=\bar{\upsilon}_j$ for $|j-k|\leq M_n$ and zero otherwise. Clearly, $\boldsymbol{\upsilon}\in\mathcal{V}$ on an event $\mathcal{A}_4$ with $\mathbb{P}(\mathcal{A}_4)\to 1$. On $\mathcal{A}_2\cap\mathcal{A}_4$,

$$|(\boldsymbol{\upsilon}-\bar{\boldsymbol{\upsilon}})'\bar{\mathbf{Q}}\boldsymbol{\beta}_2|\lesssim\varepsilon_nJ^{-1}r_{2,n}n^{-c_5}$$

for some large $c_5>0$ if we let c_1 be sufficiently large.

Step 6: Finally, partition the whole parameter space into shells: $\mathcal{O}=\cup_{\ell=-\infty}^{\bar{L}}\mathcal{O}_\ell$ where $\mathcal{O}_\ell=\{\boldsymbol{\beta}\in\mathbb{R}^{K_s}:2^{\ell-1}r_{2,n}\leq\|\boldsymbol{\beta}-\breve{\boldsymbol{\beta}}_0-\bar{\boldsymbol{\beta}}\|_\infty\leq 2^\ell r_{2,n}\}$ for the smallest $\bar{L}$ such that $\bar{L}r_{2,n}\geq c$, and $\bar{\mathbf{Q}}\bar{\boldsymbol{\beta}}=-\mathbb{E}_n[\breve{\mathbf{b}}_i\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0)\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))]$. Define $\mathcal{A}=\cap_{j=1}^4\mathcal{A}_j$. Then, for some constant $L\leq\bar{L}$, we have by Lemma SA-3.5 and the results given in the previous steps,

$$\begin{aligned}&\mathbb{P}(\|\breve{\boldsymbol{\beta}}-\breve{\boldsymbol{\beta}}_0-\bar{\boldsymbol{\beta}}\|_\infty\geq 2^Lr_{2,n}|\mathscr{F}_{XW})\\
\leq{}&\mathbb{P}\Big(\bigcup_{\ell=L}^{\bar{L}}\Big\{\inf_{\boldsymbol{\beta}\in\mathcal{O}_\ell}\sup_{\boldsymbol{\upsilon}\in\mathcal{V}}\mathbb{E}_n[\rho(y_i;\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))-\rho(y_i;\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{\upsilon})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))]<0\Big\}\Big|\mathscr{F}_{XW}\Big)+o_{\mathbb{P}}(1)\\
={}&\mathbb{P}\Big(\bigcup_{\ell=L}^{\bar{L}}\Big\{\inf_{\boldsymbol{\beta}\in\mathcal{O}_\ell}\sup_{\boldsymbol{\upsilon}\in\mathcal{V}}\Big\{\mathbb{E}\Big[\rho(y_i;\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))-\rho(y_i;\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{\upsilon})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))\\
&\qquad-[\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})-\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{\upsilon})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})]\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))|\mathscr{F}_{XW}\Big]+\\
&\qquad\mathbb{E}_n\Big[(\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})-\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{\upsilon})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))|\mathscr{F}_{XW}\Big]+\end{aligned}$$

$$\frac{1}{\sqrt{n}}\mathbb{G}_n\Big[\rho(y_i;\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))-\rho(y_i;\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{\upsilon})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))-$$

$$[\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})-\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{\upsilon})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})]\psi(y_i;\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))\Big]\Big\}<0\Big\}\Big|\mathscr{F}_{XW}\Big)+o_{\mathbb{P}}(1)$$

$$\leq\mathbb{P}\Big(\bigcup_{\ell=L}^{\bar{L}}\Big\{\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1}\sup_{\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell}}\sup_{\boldsymbol{\gamma}_1\in\mathcal{H}_3}\sup_{\boldsymbol{\upsilon}\in\mathcal{V}}\frac{1}{\sqrt{n}}\Big|(\mathbb{1}(\mathcal{A}_1)+\mathbb{1}(\mathcal{A}_1^c))\mathbb{G}_n[\delta_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\upsilon},\boldsymbol{\gamma})]\Big|>$$

$$C_4J^{-1}2^{\ell}r_{2,n}\varepsilon_n\Big\}\cap\mathcal{A}\Big|\mathscr{F}_{XW}\Big)+o_{\mathbb{P}}(1)$$

$$\leq\sum_{\ell=L}^{\bar{L}}(C_6J^{-1}2^{\ell}r_{2,n}\varepsilon_n)^{-1}\mathbb{1}(\mathcal{A}_1)\mathbb{E}\Big[\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1}\sup_{\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell}}\sup_{\boldsymbol{\gamma}_1\in\mathcal{H}_3}\sup_{\boldsymbol{\upsilon}\in\mathcal{V}}\frac{1}{\sqrt{n}}\mathbb{G}_n[\delta_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\upsilon},\boldsymbol{\gamma})]\Big|\mathscr{F}_{XW}\Big]+o_{\mathbb{P}}(1),$$

where $\mathbb{G}_n[\cdot]$ is understood as $\sqrt{n}(\mathbb{E}_n[\cdot]-\mathbb{E}[\cdot|\mathscr{F}_{XW}])$ in the above, we let $\varepsilon_n=2^Lr_{2,n}$, and $\mathbb{1}(\mathcal{A}_1)$ is an indicator of the event $\mathcal{A}_1$. Using the result in Step 1 and the rate condition, the first term in the last line can be made arbitrarily small by choosing L large enough, when n is sufficiently large. Then, the proof is complete. □

SA-7.3.7 Proof of Theorem SA-3.2

Proof. Since $\widehat{\epsilon}_i:=\epsilon_i+\eta_i-\widehat{\eta}_i=:\epsilon_i+u_i$, we can write

$$\begin{aligned}
&\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\eta}_{i,1}^2\psi(\widehat{\epsilon}_i)^2]-\mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\eta_{i,1}^2\sigma^2(x_i,\mathbf{w}_i)]\\
=&\,\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\eta}_{i,1}^2\Big(\psi(\epsilon_i+u_i)^2-\psi(\epsilon_i)^2\Big)\Big]+\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Big(\widehat{\eta}_{i,1}^2-\eta_{i,1}^2\Big)\psi(\epsilon_i)^2\Big]\\
&+\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2(\psi(\epsilon_i)^2-\sigma^2(x_i,\mathbf{w}_i))]\\
&+\Big(\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2\sigma^2(x_i,\mathbf{w}_i)]-\mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\eta_{i,1}^2\sigma^2(x_i,\mathbf{w}_i)]\Big)\\
=:&\mathbf{V}_1+\mathbf{V}_2+\mathbf{V}_3+\mathbf{V}_4.
\end{aligned}$$

Now, we bound each term in the following.

Step 1: For $\mathbf{V}_1$, we further write $\mathbf{V}_1=\mathbf{V}_{11}+\mathbf{V}_{12}$ where

$$\begin{aligned}
\mathbf{V}_{11}&:=\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2\Big(\psi(\epsilon_i+u_i)^2-\psi(\epsilon_i)^2\Big)\Big],\\
\mathbf{V}_{12}&:=\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Big((\widehat{\eta}_{i,1}^2-\eta_{i,1}^2\Big)\Big(\psi(\epsilon_i+u_i)^2-\psi(\epsilon_i)^2\Big)\Big].
\end{aligned}$$

Let $r_{1,n}=C_1(J\log n/n)^{1/2}+J^{-p-1}$ for a constant $C_1>0$. By Assumtion SA-GL and Lemma SA-3.1, $\max_{1\leq i\leq n}|u_i|\leq r_{1,n}$ with arbitrarily large probability for C_1 sufficiently large. For $\mathbf{V}_{11}$, let

$\mathcal{J}$ be the set of all the discontinuity points of $\psi(\cdot)$. Define $\mathbb{1}_{i,\mathcal{D}} := \mathbb{1}(\epsilon_i \in \mathcal{D})$ and $\mathbb{1}_{i,\mathcal{D}^c} := (1 - \mathbb{1}_{i,\mathcal{D}})$ where $\mathcal{D} := \{a : |a - \jmath| \leq r_{1,n} \text{ for some } \jmath \in \mathcal{J}\}$. Define

$$\mathbf{V}_{111} := \mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2\Big(\psi(\epsilon_i + u_i)^2 - \psi(\epsilon_i)^2\Big)\mathbb{1}_{i,\mathcal{D}}\Big],$$
$$\mathbf{V}_{112} := \mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2\Big(\psi(\epsilon_i + u_i)^2 - \psi(\epsilon_i)^2\Big)\mathbb{1}_{i,\mathcal{D}^c}\Big].$$

On the one hand, by definition of $\mathcal{D}$ and Assumption SA-GL,

$$\|\mathbf{V}_{111}\| \lesssim \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\mathbb{E}[\mathbb{1}_{i,\mathcal{D}}|\mathbf{X}]]\| + \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(\mathbb{1}_{i,\mathcal{D}} - \mathbb{E}[\mathbb{1}_{i,\mathcal{D}}|\mathbf{X}])]\|.$$

By Assumption SA-GL and Lemma SA-2.1, the first term on the right hand side is $O_{\mathbb{P}}(r_{1,n})$. For the second term, conditional on $\mathbf{X}$, it is an independent sequence with mean zero. Thus, we can apply the argument given in Step 3 below and conclude that the second term is $O_{\mathbb{P}}(\sqrt{r_{1,n}J\log J/n} + J\log J/n)$. Note that in this case, the indicator $\mathbb{1}_{i,\mathcal{D}}$ is trivially bounded uniformly.

On the other hand, by Assumption SA-GL,

$$\|\mathbf{V}_{112}\| \lesssim r_{1,n}\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2|\psi(\epsilon_i + u_i) + \psi(\epsilon_i)|]\|.$$

Since $|c| \leq \frac{1}{2}|1 + c^2|$ for any scalar c, we have

$$\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2|\psi(\epsilon_i)|\Big] \leq \mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2(1 + \psi(\epsilon_i)^2)\Big] \lesssim_{\mathbb{P}} 1,$$

by Lemma SA-3.1 and the result in Step 3. In addition, we further write

$$\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2|\psi(\epsilon_i + u_i)|\Big] = \mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2|\psi(\epsilon_i) + (\psi(\epsilon_i + u_i) - \psi(\epsilon_i))|\Big].$$

Repeat the previous argument to bound this term. We conclude that $\|\mathbf{V}_{11}\| \lesssim_{\mathbb{P}} r_{1,n}$.

$\mathbf{V}_{12}$ can be treated using the previous argument combined with the argument given in Step 2 and the result in Step 3. It leads to $\|\mathbf{V}_{12}\| \lesssim_{\mathbb{P}} r_{1,n}$.

Step 2: For $\mathbf{V}_2$, by Assumption SA-GL, Corollary SA-3.1 and the argument given later in Step

3, we have

$$\|\mathbf{V}_2\| \leq \max_{1\leq i\leq n} |\widehat{\eta}_{i,1}^2 - \eta_{i,1}^2| \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\psi(\epsilon_i)^2]\| \lesssim_{\mathbb{P}} (J\log n/n)^{1/2} + J^{-p-1}.$$

Step 3: For $\mathbf{V}_3$, in view of Lemmas SA-1.1 and SA-1.2, it suffices to show that

$$\sup_{\Delta\in\Pi} \Big\| \mathbb{E}_n[\mathbf{b}_0(x_i;\Delta)\mathbf{b}_0(x_i;\Delta)'\eta_{i,1}^2(\psi(\epsilon_i)^2 - \sigma^2(x_i,\mathbf{w}_i))] \Big\| \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n^{\frac{\nu-2}{\nu}}}\Big)^{1/2}.$$

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

$$\mathcal{G} = \Big\{(x_1,\mathbf{w}_1,\varphi_1) \mapsto b_{0,l}(x_1;\Delta)b_{0,k}(x_1;\Delta)\eta_{i,1}^2\varphi_1 : 1 \leq l \leq J(p+1), 1 \leq k \leq J(p+1), \Delta \in \Pi\Big\}.$$

For $g \in \mathcal{G}$, $\sum_{i=1}^n g(x_i,\mathbf{w}_i,\varphi_i) = \sum_{i=1}^n g(x_i,\mathbf{w}_i,\varphi_i^+) + \sum_{i=1}^n g(x_i,\mathbf{w}_i,\varphi_i^-)$.

Now, for the truncated piece, we have $\sup_{g\in\mathcal{G}} |g(x_i,\mathbf{w}_i,\varphi_i^-)| \lesssim JM$, and

$$\begin{aligned}\sup_{g\in\mathcal{G}} \mathbb{V}[g(x_1,\mathbf{w}_1,\varphi_1^-)] &\lesssim \sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}} \mathbb{E}[(\varphi_i^-)^2|x_i = x, \mathbf{w}_i = \mathbf{w}] \sup_{\Delta\in\Pi} \sup_{1\leq l,k\leq J(p+1)} \mathbb{E}[b_{0,l}^2(x_i;\Delta)b_{0,k}^2(x_i;\Delta)\eta_{i,1}^4] \\ &\lesssim JM \sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}} \mathbb{E}\Big[|\varphi_1|\Big|x_i = x\Big] \lesssim JM.\end{aligned}$$

The VC condition holds by the same argument given in the proof of Lemma SA-2.1. Then, using Proposition 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

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

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

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

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

$$\mathbb{E}[\max_{1\le i\le n}|\varphi_i^+|]\lesssim\mathbb{E}[\max_{1\le i\le n}\psi(\epsilon_i)^2]\lesssim n^{2/\nu}\quad\text{and}\quad\mathbb{E}[\mathbb{E}_n[|\varphi_i^+|]]\le\mathbb{E}[|\varphi_1|^+|]\lesssim\frac{\mathbb{E}[|\psi(\epsilon_1)|^\nu]}{M^{(\nu-2)/2}}.$$

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-GL, $\sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}}\mathbb{E}[\psi(\epsilon_i)^2|x_i=x]\lesssim 1$. Then, by the same argument given in the proof of Lemma SA-3.1,

$$\sup_{\Delta\in\Pi}\left\|\frac{1}{\sqrt{n}}\mathbb{G}_n[\mathbf{b}_s(x_i;\Delta)\mathbf{b}_s(x_i;\Delta)'\eta_{i,1}^2\sigma^2(x_i,\mathbf{w}_i)]\right\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}\quad\text{and}$$
$$\left\|\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\eta_{i,1}^2\psi(\epsilon_i)^2\Big]-\mathbb{E}\Big[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\eta_{i,1}^2\psi(\epsilon_i)^2\Big]\right\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}.$$

The proof for the first conclusion is complete.

Step 5: The second result follows by Lemmas SA-1.3, SA-3.1 and Assumption SA-GL(vi). The proof is complete.

□

SA-7.3.8 Proof of Theorem SA-3.3

Proof. We first show that for each fixed $x\in\mathcal{X}$,

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

is asymptotically normal. Conditional on $\mathscr{F}_{XW}$, the σ-field generated by $\{(x_i,\mathbf{w}_i)\}_{i=1}^n$, it is a mean zero independent sequence over i with variance equal to 1. Then by Berry-Esseen inequality,

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}(\mathbb{G}_n[a_i\psi(\epsilon_i)]\le u|)-\Phi(u)\Big|\le\min\left(1,\ \frac{\sum_{i=1}^n\mathbb{E}[|a_i\psi(\epsilon_i)|^3|\mathscr{F}_{XW}]}{n^{3/2}}\right).$$

By Lemmas SA-1.3, SA-3.1 and SA-3.2,

$$\begin{aligned}\frac{1}{n^{3/2}}\sum_{i=1}^n\mathbb{E}\Big[|a_i\psi(\epsilon_i)|^3\Big|\mathscr{F}_{XW}\Big]&\lesssim\bar{\Omega}(x)^{-3/2}\frac{1}{n^{3/2}}\sum_{i=1}^n\mathbb{E}\Big[|\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)|^3\Big|\mathscr{F}_{XW}\Big]\\&\lesssim\bar{\Omega}(x)^{-3/2}\frac{1}{n^{3/2}}\sum_{i=1}^n|\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)|^3\end{aligned}$$

$$
\begin{aligned}
&\leq \bar{\Omega}(x)^{-3/2}\frac{\sup_{x\in\mathcal{X}}\sup_{z\in\mathcal{X}}|\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(z)|}{n^{3/2}}\sum_{i=1}^{n}|\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)|^2\\
&\lesssim_{\mathbb{P}} \frac{1}{J^{3/2+3v}}\cdot\frac{J^{1+v}}{\sqrt{n}}\cdot J^{1+2v}\to 0
\end{aligned}
$$

since $J/n=o(1)$. By Theorem SA-3.2, the above weak convergence still holds if $\bar{\Omega}(x)$ is replaced by $\widehat{\Omega}(x)$. Now, the desired result follows by Theorem SA-3.1. □

SA-7.3.9 Proof of Theorem SA-3.4

Proof. The proof is divided into several steps.

Step 1: Note that

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}-\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\Omega(x)/n}}\right|\\
\leq&\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\Omega(x)/n}}\right|\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\Omega}(x)^{1/2}-\Omega(x)^{1/2}}{\widehat{\Omega}(x)^{1/2}}\right|\\
\lesssim_{\mathbb{P}}&\Big(\sqrt{\log n}+\sqrt{n}J^{-p-1-1/2}\Big)\Big(J^{-p-1}+\sqrt{\frac{J\log n}{n^{1-\frac{2}{\nu}}}}\Big)
\end{aligned}
$$

where the last step uses Lemma SA-3.2 and Corollary SA-3.1. Then, in view of Lemmas SA-1.4, SA-3.4, Theorems SA-3.1, SA-3.2 and the rate restriction given in the lemma, we have

$$
\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}-\frac{\widehat{\mathbf{b}}_s(x)'\bar{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}}\mathbb{G}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)]\right|=o_{\mathbb{P}}(a_n^{-1}).
$$

Step 2: Let us write $\mathscr{K}(x,x_i)=\Omega(x)^{-1/2}\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\mathbf{b}_s(x_i)$. 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, $\{\epsilon_i\}_{i=1}^n$, $\{\mathbf{w}_i\}_{i=1}^n$ and $\{\sigma^2(x_i,\mathbf{w}_i)\}_{i=1}^n$ are ordered as concomitants $\{\epsilon_{[i]}\}_{i=1}^n$, $\{\mathbf{w}_{[i]}\}$ and $\{\sigma^2_{[i]}\}_{i=1}^n$ where $\sigma^2_{[i]}=\sigma^2(x_{(i)},\mathbf{w}_{[i]})$. Clearly, conditional on $\mathscr{F}_{XW}$ (the σ-field generated by $\{(x_i,\mathbf{w}_i)\}$), $\{\psi(\epsilon_{[i]})\}_{i=1}^n$ is still an independent mean zero sequence. Then by Assumptions SA-DGP, SA-GL and the result of Sakhanenko (1991), there exists a sequence of i.i.d. standard normal random variables $\{\zeta_{[i]}\}_{i=1}^n$ such that

$$
\max_{1\leq\ell\leq n}|S_\ell|:=\max_{1\leq\ell\leq n}\left|\sum_{i=1}^{l}\eta^{(1)}(\mu_0(x_{(i)})+\mathbf{w}_{[i]}'\gamma_0)\psi(\epsilon_{[i]})-\sum_{i=1}^{l}\eta^{(1)}(\mu_0(x_{(i)})+\mathbf{w}_{[i]}'\gamma_0)\sigma_{[i]}\zeta_{[i]}\right|\lesssim_{\mathbb{P}} n^{\frac{1}{\nu}}.
$$

Then, using summation by parts,

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^{n}\mathscr{K}(x,x_{(i)})\eta^{(1)}(\mu_0(x_{(i)})+\mathbf{w}_{[i]}'\boldsymbol{\gamma}_0)(\psi(\epsilon_{[i]})-\sigma_{[i]}\zeta_{[i]})\right| \\
&=\sup_{x\in\mathcal{X}}\left|\mathscr{K}(x,x_{(n)})S_n-\sum_{i=1}^{n-1}S_i\left(\mathscr{K}(x,x_{(i+1)})-\mathscr{K}(x,x_{(i)})\right)\right| \\
&\leq \sup_{x\in\mathcal{X}}\max_{1\leq i\leq n}|\mathscr{K}(x,x_i)||S_n|+\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}}\sum_{i=1}^{n-1}S_i\Big(\widehat{\mathbf{b}}_s(x_{(i+1)})-\widehat{\mathbf{b}}_s(x_{(i)})\Big)\right| \\
&\leq \sup_{x\in\mathcal{X}}\max_{1\leq i\leq n}|\mathscr{K}(x,x_i)||S_n|+\sup_{x\in\mathcal{X}}\left\|\frac{\widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}}\right\|_\infty\left\|\sum_{i=1}^{n-1}S_i\Big(\widehat{\mathbf{b}}_s(x_{(i+1)})-\widehat{\mathbf{b}}_s(x_{(i)})\Big)\right\|_\infty
\end{aligned}
$$

By Lemmas SA-1.3, SA-3.1 and SA-2.2, $\sup_{x\in\mathcal{X}}\sup_{x_i\in\mathcal{X}}|\mathscr{K}(x,x_i)|\lesssim_{\mathbb{P}}\sqrt{J}$, and

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

Then, notice that

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

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 $\xi>0$,

$$
\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\mathbb{G}_n[\mathscr{K}(x,x_i)\eta^{(1)}(\mu_0(x_i)+\mathbf{w}_i'\boldsymbol{\gamma}_0)(\psi(\epsilon_i)-\sigma_i\zeta_i)]>\xi a_n^{-1}|\mathscr{F}_{XW}\Big)=o_{\mathbb{P}}(1),
$$

where we recover the original ordering. Since $\mathbb{G}_n[\widehat{\mathbf{b}}(x_i)\zeta_i\sigma_i\eta_{i,1}]=_{d|\mathscr{F}_{XW}}\mathbf{N}(0,\bar{\mathbf{\Sigma}})$ ($=_{d|\mathscr{F}_{XW}}$ denotes "equal in distribution conditional on $\mathscr{F}_{XW}$"), the above steps construct the following approximating process:

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

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

Step 3: Note that

$$\begin{aligned}\sup_{x\in\mathcal{X}}|\bar{Z}_p(x)-Z_p(x)|\leq \sup_{x\in\mathcal{X}}\Big|&\frac{\widehat{\mathbf{b}}^{(v)}(x)'(\bar{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1})}{\sqrt{\Omega(x)}}\bar{\mathbf{\Sigma}}^{1/2}\mathbf{N}_{K_s}\Big|\\ &+\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}^{(v)}(x)'\mathbf{Q}_0^{-1}}{\sqrt{\Omega(x)}}\Big(\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\Big)\mathbf{N}_{K_s}\Big|\\ &+\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}_0^{(v)}(x)'(\widehat{\mathbf{T}}_s-\mathbf{T}_s)\mathbf{Q}_0^{-1}}{\sqrt{\Omega(x)}}\mathbf{\Sigma}_0^{1/2}\mathbf{N}_{K_s}\Big|,\end{aligned}$$

where each term on the right-hand side is a mean-zero Gaussian process conditional on $\mathscr{F}_{XW}$. By Lemma SA-1.2 and SA-3.1, $\|\bar{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$ 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-1.3 and Theorem X.3.8 of Bhatia (2013), $\|\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$. By Gaussian Maximal Inequality (van der vaart and Wellner, 1996, Corollary 2.2.8),

$$\mathbb{E}\Big[\sup_{x\in\mathcal{X}}|\bar{Z}_p(x)-Z_p(x)|\Big|\mathscr{F}_{XW}\Big]\lesssim_{\mathbb{P}}\sqrt{\log J}\Big(\|\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\|+\|\bar{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1}\|+\|\widehat{\mathbf{T}}_s-\mathbf{T}_s\|\Big)=o_{\mathbb{P}}(a_n^{-1})$$

where the last line follows from the imposed rate restriction. Then the proof is complete. □

SA-7.3.10 Proof of Theorem SA-3.5

Proof. This conclusion follows from Lemmas SA-1.3, SA-3.1, Theorem SA-3.2 and Gaussian Maximal Inequality as applied in Step 3 in the proof of Theorem SA-3.4. □

SA-7.3.11 Proof of Theorem SA-3.6

Proof. By Lemmas SA-1.4, SA-3.1, SA-3.4 and Theorem SA-3.1, we immediately have

$$\begin{aligned}\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)=&\widehat{\mathbf{b}}_s(x_i)'(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_0)-\widehat{r}_{0,v}(x)\\ =&-\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)]-\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\Psi(x_i,\mathbf{w}_i;\check{\eta}_i)]\\ &-\widehat{r}_{0,v}(x)+O_{\mathbb{P}}\Big(J^v\Big\{\Big(\frac{J\log n}{n}\Big)^{3/4}\sqrt{\log n}+J^{-\frac{p+1}{2}}\Big(\frac{J\log^2 n}{n}\Big)^{1/2}+\mathfrak{r}_\gamma\Big\}\Big),\end{aligned}$$

where $\check{\eta}_i = \eta(\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)$. Note that the $O_{\mathbb{P}}(\cdot)$ holds uniformly over $x \in \mathcal{X}$, and thus the integral of the squared remainder is $o_{\mathbb{P}}(J^{1+2v}/n + J^{-2(p+1-v)})$ by the rate condition. Then,

$$\texttt{AISE} = \int_{\mathcal{X}} \Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)] + \widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\Psi(x_i, \mathbf{w}_i; \check{\eta}_i)] + \widehat{r}_{0,v}(x)\Big)^2 \omega(x)dx.$$

Next, taking conditional expectation given $\mathbf{X}$ and $\mathbf{W}$ and using the argument in the proof of Lemma SA-2.1 again, we have

$$\begin{aligned}
\mathbb{E}[\texttt{AISE}|\mathbf{X}, \mathbf{W}] = &\frac{1}{n}\,\mathrm{trace}\Big(\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\Big) + o_{\mathbb{P}}(J^{2v+1}/n)\\
&+ \int_{\mathcal{X}}\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0 - \mu_0^{(v)}(x)\Big)^2\omega(x)dx\\
&+ \int_{\mathcal{X}}\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\Psi(x_i, \mathbf{w}_i; \check{\eta}_i)]\Big)^2\omega(x)dx\\
&+ 2\int_{\mathcal{X}}\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\Psi(x_i, \mathbf{w}_i; \check{\eta}_i)]\widehat{r}_{0,v}(x)\omega(x)dx.
\end{aligned}$$

Note that by Assumption SA-GL, $\Psi(x_i, \mathbf{w}_i; \check{\eta}_i) = -\Psi_1(x_i, \mathbf{w}_i; \eta_{i,0})\eta_{i,1}\widehat{r}_0(x_i) + O_{\mathbb{P}}(J^{-2p-2})$ where $O_{\mathbb{P}}(\cdot)$ holds uniformly over i. The terms in the last three lines correspond to the integrated squared bias. We can use the expression of $\widehat{r}_{0,v}$ in Equation (SA-7.7) and repeat the argument in the proof of Theorem SA-2.6 to approximate the integrated squared bias in terms of the analogues based on the non-random partition Δ_0. □

SA-7.4 Proof for Section SA-4

SA-7.4.1 Proof of Theorem SA-4.1

Proof. We first show that

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

By Theorem SA-2.4 or Theorem SA-3.4, there exists a sequence of constants ξ_n such that $\xi_n = o(1)$ and

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

Then,

$$
\begin{aligned}
\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\widehat{T}(x)|\le u\Big) &\le \mathbb{P}\Big(\Big\{\sup_{x\in\mathcal{X}}|\widehat{T}(x)|\le u\Big\}\cap\Big\{\Big|\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)|-\sup_{x\in\mathcal{X}}|Z_p(x)|\Big|\le \xi_n/a_n\Big\}\Big)+o(1)\\
&\le \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)|\le u+\xi_n/a_n\Big)+o(1)\\
&\le \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)|\le u\Big)+\sup_{u\in\mathbb{R}}\mathbb{E}\Big[\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|Z_p(x)|-u\Big|\le \xi_n/a_n\Big|\mathbf{X}\Big)\Big]\\
&\le \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)|\le u\Big)+\mathbb{E}\Big[\sup_{u\in\mathbb{R}}\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|Z_p(x)|-u\Big|\le \xi_n/a_n\Big|\mathbf{X}\Big)\Big]+o(1).
\end{aligned}
$$

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

$$
\begin{aligned}
\sup_{u\in\mathbb{R}}\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|Z_p(x)|-u\Big|\le \xi_n/a_n\Big|\mathbf{X}\Big) &\le 4\xi_n a_n^{-1}\mathbb{E}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)|\Big|\mathbf{X}\Big]+o(1)\\
&\lesssim_{\mathbb{P}} \xi_n a_n^{-1}\sqrt{\log J}+o(1)\to 0
\end{aligned}
$$

where the last step uses Gaussian Maximal Inequality (see van der vaart and Wellner, 1996, Corollary 2.2.8). By Dominated Convergence Theorem,

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

The other side of the inequality follows similarly.

By similar argument, using Theorem SA-2.5 or Theorem SA-3.5, we have

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

Then it remains to show that

$$
\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)|\le u\Big)-\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)|\le u|\mathbf{X}\Big)\Big|=o_{\mathbb{P}}(1). \tag{SA-7.8}
$$

Now, 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 $\mathbf{V}_0 = \mathbf{T}_s'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0\mathbf{Q}_0^{-1}\mathbf{T}_s$ and $\breve{\mathbf{N}}_{K_s} := \mathbf{T}_s'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0^{1/2}\mathbf{N}_{K_s}$ is a K_s-dimensional normal random vector. Importantly, by this construction, $\breve{\mathbf{N}}_{K_s}$ and $\mathbf{V}_0$ do not depend on $\widehat{\Delta}$ and x, and they are only determined by the deterministic partition Δ_0.

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: $\{\omega : \sup_{x\in\mathcal{X}} |Z_p(x; \Delta_1)| \leq u\} = \{\omega : \sup_{x\in\mathcal{X}} |Z_p(x; \Delta_2)| \leq u\}$ for any u. Thus,

$$\mathbb{E}\Big[\mathbb{P}\Big(\sup_{x\in\mathcal{X}} |Z_p(x)| \leq u\Big|\mathbf{X}\Big)\Big] = \mathbb{P}\Big(\sup_{x\in\mathcal{X}} |Z_p(x)| \leq u\Big|\mathbf{X}\Big) + o_{\mathbb{P}}(1).$$

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-1.2, 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, e.g., the proof of Theorems SA-2.4 and SA-2.5, 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. □

SA-7.4.2 Proof of Corollary SA-4.1

Proof. Given $J = J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$, the rate restrictions required in Theorem SA-4.1 are satisfied. Let $\xi_{1,n} = o(1)$, $\xi_{2,n} = o(1)$ and $\xi_{3,n} = o(1)$. Then,

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

where $c^0(1-\alpha+\xi_{3,n})$ denotes the $(1-\alpha+\xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}}|Z_{p+q}(x)|$, the first inequality holds by Theorem SA-2.4 or Theorem SA-3.4, 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 (2014b). The other side of the bound follows similarly. $\square$

SA-7.4.3 Proof of Theorem SA-4.2

Proof. Throughout this proof, we let $\xi_{1,n}=o(1)$, $\xi_{2,n}=o(1)$ and $\xi_{3,n}=o(1)$ be sequences of vanishing constants. Moreover, let A_n be a sequence of diverging constants such that $\sqrt{\log J}A_n \lesssim \sqrt{\frac{n}{J^{1+2v}}}$. Note that under $\ddot{\mathsf{H}}_0$,

$$\sup_{x\in\mathcal{X}}|\ddot{T}_p(x)|\leq \sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}\right|+\sup_{x\in\mathcal{X}}\left|\frac{\mu_0^{(v)}(x)-m^{(v)}(x;\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\ddot{T}_p(x)|>\mathfrak{c}\Big] &\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)|>\mathfrak{c}-\sup_{x\in\mathcal{X}}\Big|\frac{\mu_0^{(v)}(x)-m^{(v)}(x;\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big|\Big]\\
&\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)|>\mathfrak{c}-\xi_{1,n}/a_n-\sup_{x\in\mathcal{X}}\Big|\frac{\mu_0^{(v)}(x)-m^{(v)}(x;\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big|\Big]+o(1)\\
&\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)|>c^0(1-\alpha-\xi_{3,n})-(\xi_{1,n}+\xi_{2,n})/a_n-\\
&\qquad \sup_{x\in\mathcal{X}}\Big|\frac{\mu_0^{(v)}(x)-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big|\Big]+o(1)\\
&\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)|>c^0(1-\alpha-\xi_{3,n})\Big]+o(1)\\
&=\alpha+o(1)
\end{aligned}$$

where $c^0(1-\alpha-\xi_{3,n})$ denotes the $(1-\alpha-\xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}}|Z_p(x)|$, the second inequality holds by Theorems SA-2.4 (or Theorem SA-3.4), the third by Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), the fourth by the condition that $\sup_{x\in\mathcal{X}}\big|\frac{\mu_0^{(v)}(x)-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\big|=o_{\mathbb{P}}(\frac{1}{\sqrt{\log J}})$ and Anti-Concentration Inequality in Chernozhukov, Chetverikov, and Kato (2014b). The other side of the bound follows similarly.

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

$$
\begin{aligned}
&\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\ddot{T}_p(x)|>\mathfrak{c}\Big]\\
=&\mathbb{P}\Big[\sup_{x\in\mathcal{X}}\Big|\widehat{T}_p(x)+\frac{\mu_0^{(v)}(x)-m^{(v)}(x,\bar{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}+\frac{m^{(v)}(x,\bar{\boldsymbol{\theta}})-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big|>\mathfrak{c}\Big]\\
\geq&\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)|\leq\sup_{x\in\mathcal{X}}\Big|\frac{\mu_0^{(v)}(x)-m^{(v)}(x,\bar{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}+\frac{m^{(v)}(x,\bar{\boldsymbol{\theta}})-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big|-\mathfrak{c}\Big]-o(1)\\
\geq&\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)|\leq\sqrt{\log J}A_n-\xi_{1,n}/a_n\Big]-o(1)\\
\geq&1-o(1).
\end{aligned}
$$

where the third line holds by Lemma SA-2.2 (or Lemma SA-3.2), Theorem SA-2.2 (or Theorem SA-3.2), 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-2.4 (or Theorem SA-3.4), and the last by the Talagrand-Samorodnitsky Concentration Inequality (van der vaart and Wellner, 1996, Proposition A.2.7). □

SA-7.4.4 Proof of Theorem SA-4.3

Proof. The definitions of A_n, $\xi_{1,n},\xi_{2,n}$ and $\xi_{3,n}$ are the same as in the proof of Theorem SA-4.2. 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,

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

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

where $c^0(1-\alpha-\xi_{3,n})$ denotes the $(1-\alpha-\xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}} Z_p(x)$, the second line holds by Theorem SA-2.4 (or Theorem SA-3.4), the third by Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), the fourth by Anti-Concentration Inequality in Chernozhukov, Chetverikov, and Kato (2014b).

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

$$\begin{aligned}
\mathbb{P}\Big[\sup_{x\in\mathcal{X}} \dot{T}_p(x) > \mathfrak{c}\Big] &= \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \Big(\widehat{T}_p(x) + \frac{\mu_0^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}\Big) > 0\Big] \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |\widehat{T}_p(x)| < \sup_{x\in\mathcal{X}} \frac{\mu_0^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c},\ \sup_{x\in\mathcal{X}} \frac{\mu_0^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} > \mathfrak{c}\Big] \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |\widehat{T}_p(x)| < \sup_{x\in\mathcal{X}} \frac{\mu_0^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}\Big] - o(1) \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |\widehat{T}_p(x)| < \sqrt{\log J} A_n\Big] - o(1) \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |Z_p(x)| < \sqrt{\log J} A_n - \xi_{1,n}/a_n\Big] - o(1) \\
&\geq 1 - o(1)
\end{aligned}$$

where the third line holds by Lemma SA-2.2 (or Lemma SA-3.2), Theorem SA-2.2 (or Theorem SA-3.2), 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 Theorem SA-2.4 (or Theorem SA-3.4), and the last by Proposition A.2.7 in van der vaart and Wellner (1996).

□

SA-7.4.5 Proof of Theorem SA-4.4

Proof. By Taylor expansion and Theorem SA-3.1,

$$\begin{aligned}
&\eta(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}) - \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0) \\
&= \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\Big(\widehat{\mathbf{b}}_s(x)'\widehat{\boldsymbol{\beta}} - \mu_0(x)\Big) + O_{\mathbb{P}}\Big(\|\widehat{\mathbf{w}} - \mathbf{w}\| + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| + \frac{J\log n}{n} + J^{-2p-2}\Big)
\end{aligned}$$

$$= \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\widehat{\mathbf{b}}_s(x)'\bar{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)]$$

$$+ O_{\mathbb{P}}\Big(J^{-p-1} + \Big(\frac{J\log n}{n}\Big)^{3/4}\sqrt{\log n} + J^{-\frac{p+1}{2}}\Big(\frac{J\log^2 n}{n}\Big)^{1/2} + \mathfrak{r}_\gamma + \|\widehat{\mathbf{w}} - \mathbf{w}\|\Big).$$

Note that $\frac{\partial}{\partial x}\vartheta(x;\mathbf{w}) = \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\mu_0^{(1)}(x)$. Then,

$$\begin{aligned}
&\eta^{(1)}(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}})\widehat{\mu}^{(1)}(x) - \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\mu_0^{(1)}(x)\\
=\,&\eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\Big(\widehat{\mu}^{(1)}(x) - \mu_0^{(1)}(x)\Big) + O_{\mathbb{P}}\Big(\Big(\frac{J\log n}{n}\Big)^{1/2} + J^{-p-1} + \mathfrak{r}_\gamma + \|\widehat{\mathbf{w}} - \mathbf{w}\|\Big)\\
=\,&\eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\widehat{\mathbf{b}}_s^{(1)}(x)'\bar{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(\epsilon_i)] +\\
&O_{\mathbb{P}}\Big(J^{-p-1+v} + \Big(\frac{J\log n}{n}\Big)^{1/2} + J^{v}\Big(\frac{J\log n}{n}\Big)^{3/4}\sqrt{\log n} + J^{-\frac{p+1-2v}{2}}\Big(\frac{J\log^2 n}{n}\Big)^{1/2} + \mathfrak{r}_\gamma + \|\widehat{\mathbf{w}} - \mathbf{w}\|\Big).
\end{aligned}$$

Then, the strong approximation can be constructed based on the same argument given in the proof of Theorem SA-3.4.

□

References

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernandez-Val (2019): "Conditional Quantile Processes based on Series or Many Regressors," *Journal of Econometrics*, 213(1), 4–29.

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 (2020): "Large sample properties of partitioning-based series estimators," *Annals of Statistics*, 48(3), 1718–1741.

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 (2014a): "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, 42(4), 1564–1597.

——— (2014b): "Anti-Concentration and Honest Adaptive Confidence Bands," *Annals of Statistics*, 42(5), 1787–1818.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139(1), 4–14.

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.

Koenker, R. (2005): *Quantile Regression*, Econometric Society Monographs. Cambridge University Press.

Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur Representation for Local Polynomial Estimates of M-Regression and Its Application to the Additive Model," *Econometric Theory*, 26(5), 1529–1564.

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.

Shen, X., D. Wolfe, and S. Zhou (1998): “Local Asymptotics for Regression Splines and Confidence Regions,” *Annals of Statistics*, 26(5), 1760–1782.

van der vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Application to Statistics*. Springer.