Nonlinear Binscatter Methods
Supplemental Appendix*

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

July 19, 2024

Abstract

This supplement collects all technical proofs for more general theoretical results than those reported in the main paper. Several of our new theoretical results for nonlinear partitioning-based series estimation may be of independent interest. More details on methodological aspects of nonlinear binscatter are also provided. Companion general-purpose software and replication files are available at https://nppackages.github.io/binsreg/.

Contents

SA-1   Introduction

This supplemental appendix is a comprehensive collection of all our new theoretical results for nonlinear binscatter estimators with semi-linear covariate-adjustment and random partitioning. Many of our results contribute to the broader literature on nonparametric estimation and inference, particularly when using series estimators, and are thus of independent interest outside of binned scatter plots. To help place our results in the literature, we include a remark labelled "Improvements over literature" at the end of each technical subsection that discusses in detail the technical improvements presented in that subsection and gives related references.

Here we give a brief summary of this appendix, including pointers to some of the major new results. We proceed as follows. The next subsection lists notation used throughout; further notation is defined throughout Section SA-2 and at the outset of Section SA-3. Then, Section SA-2 describes the setup for nonlinear binscatter methods, including the statistical model, parameters of interest, and assumptions, as well as the (random) partitioning and estimation details. Specifically, Assumption SA-DGP imposes some basic conditions on the data generating process. Assumption SA-SM imposes some technical conditions that characterizes and restricts the statistical model of interest. The loss function specified there is general enough to cover many practically important examples such as mean regression, quantile regression, logit/probit estimation, and Huber regression. Assumption SA-HLE imposes some mild high-level conditions on the estimation procedure. Assumption SA-RP summarizes the key conditions on the partitioning scheme used in our theory. We allow for a large class of random partitions. Importantly, the "convergence" of the random partition (Assumption SA-RP(ii)) is not necessary for most of our main theoretical results, thereby allowing for flexible data-driven partitioning methods, including certain recursive adaptive partitioning methods: see Devroye et al. (2013), Zhang and Singer (2010), and references therein.

Section SA-3.1 presents some preliminary technical lemmas for analyzing nonlinear binscatter (and thus also partitioning-based estimators more broadly). New results include precise non-asymptotic concentration results related to Gram matrices (Lemma SA-3.1), asymptotic variances (Lemmas SA-3.2 and SA-3.3), approximation errors (Lemma SA-3.4), and uniform convergence (Lemma SA-3.5). Sharp control of these objects is a crucial ingredient for obtaining results under weak conditions, as below.

Section SA-3.2 presents a tight uniform (in x) Bahadur representation for nonlinear binscatter (Theorem SA-3.1). This is our first main result. We allow for random partitions and much weaker rate restrictions (on the tuning parameter J) than previously imposed in the literature, in addition to additional controls. The data-dependent partitioning means our series estimator uses random basis functions, and this is entirely new. In terms of tuning parameter rate restrictions, previous results required $J^4/n \to 0$ (up to $\log(n)$ terms) or something stricter, while our restriction is that $J^{\frac{2\nu}{\nu-1}}/n \to 0$ (up to $\log(n)$ terms), with $\nu > 2$ denoting the number of finite moments of the "score", and thus may be substantially weaker. Note that our class of models is often broader than prior work also. Importantly, our results can now allow for piecewise constant binscatters, i.e., with degree $p = 0$, which is excluded by prior results in the literature (i.e., for previous technical results there was no sequence $J \to \infty$ such that the bias and variance are simultaneously controlled). In addition, employing our novel uniform Bahadur representation, we can establish the uniform convergence rates of nonlinear binscatter (Corollary SA-3.1) and variance estimators (Theorem SA-3.2) under similarly weak restrictions.

Section SA-3.3 studies the pointwise distributional approximation for nonlinear binscatter estimators. These results are omitted from the main paper to save space, but are standard properties of interest in the nonparametrics literature and thus are included for completeness. The main result is Theorem SA-3.3, which establishes pointwise asymptotic Normality for our point estimators, again allowing for random (and possibly "non-convergent") partitions, and under mild rate restrictions similar to those for the (uniform) Bahadur representation.

Section SA-3.4 presents a new Nagar-type approximate IMSE expansion for nonlinear binscatter estimators with semi-linear covariate-adjustment and random partitions (Theorem SA-3.4), which has no antecedent in the literature. Our results can be used to design data-driven procedures for selecting IMSE-optimal choices of tuning parameters for nonlinear binscatter. Again, these results are novel in their breadth, the weakness of the assumptions, and the conditions on the partitioning. Here we do require an extra assumption on the partitioning in order to characterize the leading terms in the expansion: intuitively, the random partitioning must "settle" to a population partition so that the leading constants of the expansion can be expressed. For example, sample quantiles converge to population quantiles, so this assumption is satisfied.

Uniform inference is dealt with in the next several sections of this appendix. First, Section SA-3.5

establishes a uniform (in x) distributional approximation for nonlinear binscatter estimators. The two main results, which are combined into one in the main text (Theorem 2), are the (conditional) strong approximation in Theorem SA-3.5 and the feasible implementation thereof in Theorem SA-3.6. Again, We allow for a large class of random partitions, a broad class of (possibly) nonlinear and nonsmooth models, and additional controls. Here the partitions do not need to be "convergent" in any sense. These results are obtained under weak assumptions, including in particular mild rate restrictions, as in the case of the uniform Bahadur representation, all of which improves on the literature in various directions as explained in the text below. Finally, Theorem SA-3.7 shows a distributional approximation for the suprema of the t-statistic processes in the case of the convergent partition (as in the previous paragraph).

Sections SA-3.6–SA-3.8 employ the strong approximation results to study uniform inference for various parameters of interest in the specific context of nonlinear binscatter. These results rely on, and inherit the novelty of, Theorems SA-3.5 and SA-3.6. New results include valid uniform confidence bands (Theorem SA-3.8), consistent hypothesis tests about parametric specification (Theorem SA-3.9) and tests for shape restrictions (Theorem SA-3.10). All these results explicitly account for the possibly random partitioning scheme and semi-linear covariate-adjustment with random evaluation points.

Section SA-4 discusses implementation details for nonlinear binscatter, including standard error computation, feasible data-driven number of bins selector, and choices of polynomial orders given a fixed number of bins. For a more explicit treatment of the package `binsreg` per se, see Cattaneo et al. (2024a) and https://nppackages.github.io/binsreg/.

Finally, Section SA-5 contains the proofs for all the technical results in Section SA-3.

SA-1.1   Notation

See van der Vaart and Wellner (1996), Bhatia (2013), Giné and Nickl (2016), and references therein, for background definitions.

Matrices and Norms. For (column) vectors, $\|\cdot\|$ denotes the Euclidean norm, $\|\cdot\|_1$ denotes the L_1 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

4

$\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$ 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 $\mathcal{S}^L$ to denote the unit circle in $\mathbb{R}^L$, i.e., $\|\mathbf{a}\| = 1$ for any $\mathbf{a} \in \mathcal{S}^L$. 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 if g is a univariate function, let $g^{(v)}(z) = d^v g(z)/dz^v$ be the vth derivative for $v \geq 0$.

Asymptotics. For sequences of numbers or random variables, we use $l_n \lesssim m_n$ to denote that $\limsup_n |l_n/m_n|$ is finite, $l_n \lesssim_{\mathbb{P}} m_n$ or $l_n = O_{\mathbb{P}}(m_n)$ to denote $\limsup_{\varepsilon\to\infty} \limsup_n \mathbb{P}[|l_n/m_n| \geq \varepsilon] = 0$, $l_n = o(m_n)$ implies $l_n/m_n \to 0$, and $l_n = o_{\mathbb{P}}(m_n)$ implies that $l_n/m_n \to_{\mathbb{P}} 0$, where $\to_{\mathbb{P}}$ denotes convergence in probability. Accordingly, we write $l_n \gtrsim m_n$ if $m_n \lesssim l_n$, and $l_n \gtrsim_{\mathbb{P}} m_n$ if $m_n \lesssim_{\mathbb{P}} l_n$. $l_n \asymp m_n$ implies that $l_n \lesssim m_n$ and $m_n \lesssim l_n$.

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 $(A, \mathcal{A})$ and a suitably measurable class of functions $\mathcal{G}$ mapping A 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 $N(\mathcal{G}, L_2(\mathbb{Q}), \varepsilon\|\bar{G}\|_{\mathbb{Q},2})$.

Other. $\lceil z \rceil$ 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-2   Setup

Suppose that $(y_i, x_i, \mathbf{w}_i')$, $1 \leq i \leq n$, 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. Let $\mathbf{D} = [(y_i, x_i, \mathbf{w}_i')' : i = 1, 2, \ldots, n]$.

For a loss function $\rho(\cdot; \cdot)$ and a strictly monotonic transformation function $\eta(\cdot)$, define

$$(\mu_0(\cdot), \boldsymbol{\gamma}_0) = \underset{\mu\in\mathcal{M}, \boldsymbol{\gamma}\in\mathbb{R}^d}{\arg\min} \ \mathbb{E}\Big[\rho\Big(y_i; \eta(\mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma})\Big)\Big], \tag{SA-2.1}$$

where $\mathcal{M}$ is a space of functions satisfying certain smoothness conditions to be specified later.

This setup is general. For example, consider $\boldsymbol{\gamma}_0 = \mathbf{0}$. If $\rho(\cdot; \cdot)$ is a squared loss and $\eta(\cdot)$ is the 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 logistic regression for binary responses. When $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, the parametric and the nonparametric components are additively separable, and thus (SA-2.1) becomes a generalized partially linear model.

Binscatter estimators are typically constructed based on a (possibly random) partition of the support of the covariate x_i. Specifically, the relevant support of x_i is partitioned into J disjoint intervals, leading to the partitioning scheme $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \widehat{\mathcal{B}}_2, \ldots, \widehat{\mathcal{B}}_J\}$, where

$$
\widehat{\mathcal{B}}_j = \begin{cases} [\widehat{\tau}_{j-1}, \widehat{\tau}_j) & \text{if } j = 1, \cdots, J - 1 \\ [\widehat{\tau}_{J-1}, \widehat{\tau}_J] & \text{if } j = J \end{cases},
$$

One popular choice in binscatter applications is the quantile-based partition: $\widehat{\tau}_j = \widehat{F}_X^{-1}((j-1)/J)$ with $\widehat{F}_X(u) = n^{-1} \sum_{i=1}^n \mathbb{1}(x_i \leq u)$ the empirical cumulative distribution function and $\widehat{F}_X^{-1}$ its generalized inverse. Our theory is general enough to cover other partitioning schemes satisfying certain regularity conditions specified below. An innovation herein is accounting for the additional randomness from the partition $\widehat{\Delta}$. The number of bins J plays the role of the 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 piecewise polynomial basis of degree p, for some choice of $p = 0, 1, 2, \ldots$, is defined as

$$
\begin{bmatrix} \mathbb{1}_{\widehat{\mathcal{B}}_1}(x) & \mathbb{1}_{\widehat{\mathcal{B}}_2}(x) & \cdots & \mathbb{1}_{\widehat{\mathcal{B}}_J}(x) \end{bmatrix}' \otimes \begin{bmatrix} 1 & x & \cdots & x^p \end{bmatrix}',
$$

where $\mathbb{1}_{\mathcal{A}}(x) = \mathbb{1}(x \in \mathcal{A})$ and $\otimes$ is the Kronecker product operator. For convenience of later analysis, we use $\widehat{\mathbf{b}}_{p,0}(x)$ to denote a *standardized rotated* basis, the jth element of which is given by

$$
\sqrt{J} \times \mathbb{1}_{\widehat{\mathcal{B}}_{\bar{j}}}(x) \times \left(\frac{x - \widehat{\tau}_{\bar{j}-1}}{\hat{h}_{\bar{j}}} \right)^{j-1-(\bar{j}-1)(p+1)}, \quad j = 1, \cdots, (p+1)J,
$$

where $\bar{\bar{j}} = \lceil j/(p+1) \rceil$, $\lceil \cdot \rceil$ is the ceiling operator, and $\hat{h}_{\bar{j}} = \widehat{\tau}_{\bar{j}} - \widehat{\tau}_{\bar{j}-1}$. 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 helps simplify some expressions of our results. The standardized rotated basis $\widehat{\mathbf{b}}_{p,0}(x)$ is equivalent to the original piecewise polynomial basis 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 $1 \le s \le p$, we introduce the following basis

$$\widehat{\mathbf{b}}_{p,s}(x) = \left(\widehat{b}_{p,s,1}(x), \ldots, \widehat{b}_{p,s,K_{p,s}}(x) \right)' = \widehat{\mathbf{T}}_s \widehat{\mathbf{b}}_{p,0}(x), \qquad K_{p,s} = (p+1)J - s(J-1),$$

where $\widehat{\mathbf{T}}_s := \widehat{\mathbf{T}}_s(\widehat{\Delta})$ is a $K_{p,s} \times (p+1)J$ matrix depending on $\widehat{\Delta}$, which transforms a piecewise polynomial basis into a smoothed binscatter basis. Some useful properties of $\widehat{\mathbf{T}}_s$ are given in Lemma SA-5.3 in Section SA-5, and the explicit representation of $\widehat{\mathbf{T}}_s$ is available in the proof of Lemma SA-3.2 in Cattaneo, Crump, Farrell and Feng (2024b). When $s = 0$, we let $\widehat{\mathbf{T}}_0 = \mathbf{I}_{(p+1)J}$, the identity matrix of dimension $(p+1)J$. When $s = p$, $\widehat{\mathbf{b}}_{p,s}(x)$ is the well-known B-spline basis of order $p+1$ with simple knots, which is $(p-1)$-times continuously differentiable. 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 about spline functions. We require $s \le p$, since if $s = p+1$, $\widehat{\mathbf{b}}_{p,s}(x)$ reduces to a global polynomial basis of degree p.

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

$$\widehat{\mu}_{p,s}^{(v)}(x) := \widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \arg\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta\big(\widehat{\mathbf{b}}_{p,s}(x_i)'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\gamma}\big)\Big), \qquad \text{(SA-2.2)}$$

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

$$\widehat{\mu}_{p,s}^{(v)}(x) = \widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \quad \widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}) := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{K_{p,s}}} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta\big(\widehat{\mathbf{b}}_{p,s}(x_i)'\boldsymbol{\beta} + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}\big)\Big). \qquad \text{(SA-2.3)}$$

The representation (SA-2.3) 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-HLE below), which in nonlinear estimation models cases

can be justified in different ways, e.g., joint estimation, backfitting, profiling, and split-sampling, among other possibilities. Our software implementation (Cattaneo, Crump, Farrell and Feng, 2024a) relies on joint estimation, as done by the default base estimation packages in `Python`, `R`, and `Stata`.

In this supplement, we focus on estimation and inference of the following three parameters:

(i) the nonparametric component $\mu_0^{(v)}(x)$ for any $v \geq 0$,

(ii) the level function $\vartheta_0(x, \mathsf{w}) = \eta(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)$, and

(iii) the marginal effect $\zeta_0(x, \mathsf{w}) = \frac{\partial}{\partial x}\eta(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)$,

where w is a user-chosen evaluation point of the control variables, and thus these parameters are viewed as functions of x only in our theory. Nevertheless, all our results are readily applied to other linear or nonlinear transformations of $\mu_0(x)$, such as the higher-order derivatives $\frac{\partial^v}{\partial x^v}\eta(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)$. Given the binscatter estimates $\widehat{\mu}_{p,s}(x)$ and $\widehat{\boldsymbol{\gamma}}$ in (SA-2.2), the estimators of the three parameters defined above are given by

$$\widehat{\mu}_{p,s}^{(v)}(x), \quad \widehat{\vartheta}_{p,s}(x, \widehat{\mathsf{w}}) = \eta(\widehat{\mu}_{p,s}(x) + \widehat{\mathsf{w}}'\widehat{\boldsymbol{\gamma}}), \quad \text{and} \quad \widehat{\zeta}_{p,s}(x, \widehat{\mathsf{w}}) = \eta^{(1)}(\widehat{\mu}_{p,s}(x) + \widehat{\mathsf{w}}'\widehat{\boldsymbol{\gamma}})\widehat{\mu}_{p,s}^{(1)}(x)$$

respectively, for some consistent estimate $\widehat{\mathsf{w}}$ (non-random or generated based on $\{\mathsf{w}_i\}_{i=1}^n$) of the evaluation point w. As a reminder, we need to require $p \geq v$ to get $\widehat{\mu}_{p,s}^{(v)}(x)$, $p \geq 0$ to get $\widehat{\vartheta}_{p,s}(x, \widehat{\mathsf{w}})$, and $p \geq 1$ to get $\widehat{\zeta}_{p,s}(x, \widehat{\mathsf{w}})$.

Recall that in the main text we always set $s = p$ and omit the dependence of estimators on s. Thus, $\widehat{\mu}_p^{(v)}(x) = \widehat{\mu}_{p,p}^{(v)}(x)$, $\widehat{\vartheta}_p(x, \widehat{\mathsf{w}}) = \widehat{\vartheta}_{p,p}(x, \widehat{\mathsf{w}})$, and $\widehat{\zeta}_p(x, \widehat{\mathsf{w}}) = \widehat{\zeta}_{p,p}(x, \widehat{\mathsf{w}})$. In this supplement, however, all our results hold for a general choice of the degree and the smoothness of the basis. For ease of notation, the subscripts p and s of the above estimators are dropped hereafter:

$$\widehat{\mu}^{(v)}(x) := \widehat{\mu}_{p,s}^{(v)}(x), \quad \widehat{\vartheta}(x, \widehat{\mathsf{w}}) := \widehat{\vartheta}_{p,s}(x, \widehat{\mathsf{w}}), \quad \text{and} \quad \widehat{\zeta}(x, \widehat{\mathsf{w}}) := \widehat{\zeta}_{p,s}(x, \widehat{\mathsf{w}}).$$

Remark SA-2.1 (Smoothness and Bias Correction)**.** This supplemental appendix presents *all* results under general choices of the number of bins J, the degree of the basis p, and the smoothness of the basis s. By contrast, for simplicity, the main paper employs the basis with the maximum

smoothness, i.e. choosing $s = p$, and considers the special case in which J is taken to be the IMSE-optimal choice for a fixed p (see Theorem SA-3.4), and inference is conducted based on the binscatter basis of degree $(p + 1)$. Such a choice of J guarantees that the smoothing bias of the binscatter estimator is negligible in inference under mild conditions and thus can be viewed as a bias correction strategy. ⌟

We first assume the following basic conditions on the data generating process.

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

(i) $\{(y_i, x_i, \mathbf{w}_i') : 1 \leq i \leq n\}$ *are i.i.d. random vectors satisfying* (SA-2.1) *and supported on* $\mathcal{Y} \times \mathcal{X} \times \mathcal{W}$*, where* $\mathcal{X}$ *is a compact interval and* $\mathcal{W}$ *is a compact set.*

(ii) $F_X(x) := \mathbb{P}[x_i \leq x]$ *has a Lipschitz continuous (Lebesgue) density* $f_X(x)$ *bounded away from zero on* $\mathcal{X}$*.*

(iii) $F_{Y|XW}(y|x_i, \mathbf{w}_i) := \mathbb{P}[y_i \leq y | x_i, \mathbf{w}_i]$ *has a (conditional) density* $f_{Y|XW}(y|x_i, \mathbf{w}_i)$ *supported on* $\mathcal{Y}_{x\mathbf{w}}$ *with respect to some sigma-finite measure, and* $\sup_{x \in \mathcal{X}, \mathbf{w} \in \mathcal{W}} \sup_{y \in \mathcal{Y}_{x\mathbf{w}}} f_{Y|XW}(y|x, \mathbf{w}) \lesssim 1$*.*

Next, we impose several technical conditions related to the statistical model of interest.

Assumption SA-SM (Statistical Model)**.**

(i) $\rho(y; \eta)$ *is absolutely continuous with respect to* $\eta \in \mathbb{R}$ *and admits a derivative* $\psi(y, \eta) := \psi^\dagger(y - \eta)\psi^\ddagger(\eta)$ *almost everywhere.* $\psi^\ddagger(\cdot)$ *is continuously differentiable and strictly positive or negative.* $\psi^\dagger(\cdot)$ *is Lipschitz continuous if* $F_{Y|XW}(y|x_i, \mathbf{w}_i)$ *does not have a Lebesgue density, or piecewise Lipschitz with finitely many discontinuity points otherwise.*

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

(iii) $\mathbb{E}[\psi(y_i, \eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0))|x_i, \mathbf{w}_i] = 0$*. For* $\sigma^2(x, \mathbf{w}) := \mathbb{E}[\psi(y_i, \eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0))^2 | x_i = x, \mathbf{w}_i = \mathbf{w}]$*,* $\inf_{x \in \mathcal{X}, \mathbf{w} \in \mathcal{W}} \sigma^2(x, \mathbf{w}) \gtrsim 1$*.* $\mathbb{E}[\eta^{(1)}(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0)^2 \sigma^2(x_i, \mathbf{w}_i)|x_i = x]$ *is Lipschitz continuous on* $\mathcal{X}$*, and* $\sup_{x \in \mathcal{X}, \mathbf{w} \in \mathcal{W}} \mathbb{E}[|\psi(y_i, \eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0))|^\nu | x_i = x, \mathbf{w}_i = \mathbf{w}] \lesssim 1$ *for some* $\nu > 2$*.* $\mathbb{E}[\psi(y_i, \eta)|x_i = x, \mathbf{w}_i = \mathbf{w}]$ *is twice continuously differentiable with respect to* η*.*

(iv) $\inf_{x \in \mathcal{X}, \mathbf{w} \in \mathcal{W}} \varkappa(x, \mathbf{w}) \gtrsim 1$ *and* $\mathbb{E}[\varkappa(x_i, \mathbf{w}_i)|x_i = x]$ *is Lipschitz continuous on* $\mathcal{X}$ *where* $\varkappa(x, \mathbf{w}) :=$
$\Psi_1(x, \mathbf{w}; \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0))(\eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0))^2$, $\Psi_1(x, \mathbf{w}; \eta) := \frac{\partial}{\partial \eta}\Psi(x, \mathbf{w}; \eta)$, *and* $\Psi(x, \mathbf{w}; \eta) :=$
$\mathbb{E}[\psi(y_i, \eta)|x_i = x, \mathbf{w}_i = \mathbf{w}]$.

(v) $\mu_0(\cdot)$ *is* ς*-times continuously differentiable for some* $\varsigma \geq p + 1$.

Our next assumption imposes mild high-level conditions on the estimator $\widehat{\boldsymbol{\gamma}}$ of the coefficient vector $\boldsymbol{\gamma}_0$, the estimator $\widehat{\mathbf{w}}$ of the evaluation point $\mathbf{w}$ for control variables, and the estimator of the function Ψ_1 defined previously in Assumption SA-SM(iv).

Assumption SA-HLE (High-Level Estimation Conditions)**.**

(i) $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \lesssim_{\mathbb{P}} \mathfrak{r}_\gamma$ *for* $\mathfrak{r}_\gamma = o(\sqrt{J/n} + J^{-p-1})$, *and* $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(1)$.

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

Note that $\Upsilon(x, \mathbf{w}) = \Psi_1(x, \mathbf{w}; \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0))$ in the main paper to streamline the presentation. Part (i) is a mild condition on the convergence of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\mathbf{w}}$. Part (ii) is a high-level condition that ensures we have a valid feasible estimator of the Gram matrix ($\bar{\mathbf{Q}}$ or $\mathbf{Q}_0$ defined at the outset of Section SA-3 below). Note that the convergence rate of $\eta^{(1)}(\widehat{\mu}(x_i) + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}})$ can be deduced from Corollary SA-3.1 below. Thus, part (ii) can be largely viewed as a restriction on $\widehat{\Psi}_1$ only. Note that $\widehat{\Psi}_1$ does not have to be consistent for Ψ_1 in any sense; it suffices that the estimator $\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\varkappa}(x_i, \mathbf{w}_i)]$ based on $\widehat{\Psi}_1$ as a whole is consistent. See Section SA-4 for several examples of the estimator $\widehat{\Psi}_1$.

SA-2.1 Partitions

We need some regularity conditions on the partitioning scheme, which can be verified in a case-by-case basis. We first define a family of "quasi-uniform" partitions for some absolute constant $C > 0$:

$$\Pi_C = \left\{\Delta : \frac{\max_{1 \leq j \leq J} h_j(\Delta)}{\min_{1 \leq j \leq J} h_j(\Delta)} \leq C\right\}, \tag{SA-2.4}$$

where $h_j(\Delta)$ denotes the length of the jth bin in the partition Δ. Roughly speaking, (SA-2.4) says that the bins in any $\Delta \in \Pi_C$ do not differ too much in length. Also, let $\mathbf{X} = [x_1, \dots, x_n]'$, $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_n]'$ and $\mathbf{Y} = [y_1, \cdots, y_n]'$.

Assumption SA-RP (Random Partition).

(i) $\widehat{\Delta} \perp\!\!\!\perp \mathbf{Y}|(\mathbf{X}, \mathbf{W})$ *and* $\widehat{\Delta} \in \Pi_C$ *w.p.a. 1 for some absolute constant* $C > 0$.

(ii) *There exists a non-random partition* $\Delta_0 = \{\mathcal{B}_1, \cdots, \mathcal{B}_J\}$ *with* $\mathcal{B}_j = [\tau_{j-1}, \tau_j)$ *for* $j \leq J - 1$ *and* $\mathcal{B}_J = [\tau_{J-1}, \tau_J]$ *such that* $\frac{\max_{1 \leq j \leq J} h_j}{\min_{1 \leq j \leq J} h_j} \leq c_{\mathtt{QU}}$ *for some absolute constant* $c_{\mathtt{QU}} > 0$, *and* $\max_{1 \leq j \leq J} |\hat{h}_j - h_j| \lesssim_{\mathbb{P}} J^{-1} \mathfrak{r}_{\mathtt{RP}}$ *for* $\mathfrak{r}_{\mathtt{RP}} = o(1)$.

Part (i) is the key condition for our main results and will be imposed throughout. First, it requires that the possibly random partition $\widehat{\Delta}$ be independent of the outcome $\mathbf{Y}$ given the covariates $(\mathbf{X}, \mathbf{W})$. This conditional independence assumption is trivially satisfied if $\widehat{\Delta}$ is deterministic (e.g., equally-spaced partition) or depends on $\mathbf{X}$ and $\mathbf{W}$ only (e.g., quantile-spaced partition based on $\mathbf{X}$). It also holds if a sample splitting scheme is used: a subsample (including the information about the outcome) is used for constructing the partition, and the other is employed to construct the binscatter estimator (so that $\widehat{\Delta}$ is independent of the data $(\mathbf{X}, \mathbf{W}, \mathbf{Y})$). Second, $\widehat{\Delta}$ is required to be "quasi-uniform" with large probability. It is trivially true for equally-spaced partitions and can be verified for quantile-spaced partitions under the mild conditions on the covariates density imposed before (see Lemma SA-5.2). However, this condition may be too restrictive for other modern machine-learning-based partitioning methods, in which case some additional regularization may be necessary to recover the quasi-uniformity property.

Part (ii) requires that the random partition $\widehat{\Delta}$ "stabilizes" to a fixed one in large samples. This is true if the partition is non-deterministic or generated by sample quantiles (since sample quantiles converge to population quantiles), but more generally, it is not always possible. Fortunately, this "convergence" requirement is not necessary for most of our main results (except Theorem SA-3.4 and Theorem SA-3.7). Thus, we will always make it clear if part (ii) of Assumption SA-RP is imposed.

Given the random partition $\widehat{\Delta}$, we use the notation $\mathbb{E}_{\widehat{\Delta}}[\cdot]$ to denote the expectation operator with the partition $\widehat{\Delta}$ viewed as fixed. To further simplify notation, let $\hat{h}_j = \hat{\tau}_j - \hat{\tau}_{j-1}$ be the width of the jth bin $\widehat{\mathcal{B}}_j$, and when the "limiting" partition $\Delta_0 = \{\mathcal{B}_1, \cdots, \mathcal{B}_J\}$ is defined (Assumption SA-RP(ii) holds), let h_j be the width of $\mathcal{B}_j$. Analogously to $\widehat{\mathbf{b}}_{p,s}(x)$, $\mathbf{b}_{p,s}(x)$ denotes the binscatter basis of degree p that is $(s-1)$-times continuously differentiable and is constructed based on the *nonrandom* partition Δ_0. We sometimes write $\mathbf{b}_{p,s}(x; \Delta) = (b_{p,s,1}(x; \Delta), \ldots, b_{p,s,K_{p,s}}(x; \Delta))'$ to emphasize a

binscatter basis is constructed based on a particular partition Δ. Therefore, $\widehat{\mathbf{b}}_{p,s}(x) = \mathbf{b}_{p,s}(x; \widehat{\Delta})$ and $\mathbf{b}_{p,s}(x) = \mathbf{b}_{p,s}(x; \Delta_0)$. Accordingly, we use $\mathbf{T}_s$ to denote the transformation matrix based on the non-random partition Δ_0 (which transforms $\mathbf{b}_{p,0}(x)$ to $\mathbf{b}_{p,s}(x)$).

SA-3   Main Results

We introduce the following quantities that will be extensively used throughout the supplement:

$$\eta_i = \eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0), \qquad \widehat{\eta}_i = \eta(\widehat{\mu}(x_i) + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}),$$

$$\eta_{i,1} = \eta^{(1)}(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0), \qquad \widehat{\eta}_{i,1} = \eta^{(1)}(\widehat{\mu}(x_i) + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}),$$

$$\eta_{0,1}(x, \mathsf{w}) = \eta^{(1)}(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0), \qquad \widehat{\eta}_{0,1}(x, \widehat{\mathsf{w}}) = \eta^{(1)}(\widehat{\mu}(x) + \widehat{\mathsf{w}}'\widehat{\boldsymbol{\gamma}}),$$

$$\widehat{\mu}(x_i) = \widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}}, \qquad \epsilon_i = y_i - \eta_i, \qquad \widehat{\epsilon}_i = y_i - \widehat{\eta}_i,$$

$$\widehat{\mathbf{Q}}_{p,s} := \widehat{\mathbf{Q}}_{p,s}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\Psi}_1(x_i, \mathbf{w}_i; \widehat{\eta}_i)\widehat{\eta}_{i,1}^2],$$

$$\bar{\mathbf{Q}}_{p,s} := \bar{\mathbf{Q}}_{p,s}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\Psi_1(x_i, \mathbf{w}_i; \eta_i)\eta_{i,1}^2],$$

$$\mathbf{Q}_{0,p,s} := \mathbf{Q}_{p,s}(\Delta_0) := \mathbb{E}[\mathbf{b}_{p,s}(x_i)\mathbf{b}_{p,s}(x_i)'\Psi_1(x_i, \mathbf{w}_i; \eta_i)\eta_{i,1}^2],$$

$$\widehat{\boldsymbol{\Sigma}}_{p,s} := \widehat{\boldsymbol{\Sigma}}_{p,s}(\widehat{\Delta}) := \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\psi(y_i, \widehat{\eta}_i)^2\widehat{\eta}_{i,1}^2],$$

$$\bar{\boldsymbol{\Sigma}}_{p,s} := \bar{\boldsymbol{\Sigma}}_{p,s}(\widehat{\Delta}) := \mathbb{E}_n\left[\mathbb{E}\left[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\psi(y_i, \eta_i)^2\eta_{i,1}^2\Big|\mathbf{X}, \mathbf{W}\right]\right],$$

$$\boldsymbol{\Sigma}_{0,p,s} := \boldsymbol{\Sigma}_{p,s}(\Delta_0) := \mathbb{E}\left[\mathbf{b}_{p,s}(x_i)\mathbf{b}_{p,s}(x_i)'\psi(y_i, \eta_i)^2\eta_{i,1}^2\right],$$

$$\widehat{\Omega}_{\mu^{(v)},p,s}(x) := \widehat{\Omega}_{\mu^{(v)},p,s}(x; \widehat{\Delta}) := \widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\mathbf{Q}}_{p,s}^{-1}\widehat{\boldsymbol{\Sigma}}_{p,s}\widehat{\mathbf{Q}}_{p,s}^{-1}\widehat{\mathbf{b}}_{p,s}^{(v)}(x),$$

$$\bar{\Omega}_{\mu^{(v)},p,s}(x) := \bar{\Omega}_{\mu^{(v)},p,s}(x; \widehat{\Delta}) := \widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\bar{\mathbf{Q}}_{p,s}^{-1}\bar{\boldsymbol{\Sigma}}_{p,s}\bar{\mathbf{Q}}_{p,s}^{-1}\widehat{\mathbf{b}}_{p,s}^{(v)}(x),$$

$$\Omega_{\mu^{(v)},p,s}(x) := \Omega_{\mu^{(v)},p,s}(x; \widehat{\Delta}) := \widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\mathbf{Q}_{0,p,s}^{-1}\boldsymbol{\Sigma}_{0,p,s}\mathbf{Q}_{0,p,s}^{-1}\widehat{\mathbf{b}}_{p,s}^{(v)}(x),$$

$$\widehat{\Omega}_{\vartheta,p,s}(x) := \widehat{\Omega}_{\vartheta,p,s}(x; \widehat{\Delta}) := [\eta^{(1)}(\widehat{\mu}(x) + \widehat{\mathsf{w}}'\widehat{\boldsymbol{\gamma}})]^2\widehat{\mathbf{b}}_{p,s}(x)'\widehat{\mathbf{Q}}_{p,s}^{-1}\widehat{\boldsymbol{\Sigma}}_{p,s}\widehat{\mathbf{Q}}_{p,s}^{-1}\widehat{\mathbf{b}}_{p,s}(x),$$

$$\bar{\Omega}_{\vartheta,p,s}(x) := \bar{\Omega}_{\vartheta,p,s}(x; \widehat{\Delta}) := [\eta^{(1)}(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)]^2\widehat{\mathbf{b}}_{p,s}(x)'\bar{\mathbf{Q}}_{p,s}^{-1}\bar{\boldsymbol{\Sigma}}_{p,s}\bar{\mathbf{Q}}_{p,s}^{-1}\widehat{\mathbf{b}}_{p,s}(x),$$

$$\Omega_{\vartheta,p,s}(x) := \Omega_{\vartheta,p,s}(x; \widehat{\Delta}) := [\eta^{(1)}(\mu(x) + \mathsf{w}'\boldsymbol{\gamma}_0)]^2\widehat{\mathbf{b}}_{p,s}(x)'\mathbf{Q}_{0,p,s}^{-1}\boldsymbol{\Sigma}_{0,p,s}\mathbf{Q}_{0,p,s}^{-1}\widehat{\mathbf{b}}_{p,s}(x),$$

$$\widehat{\Omega}_{\zeta,p,s}(x) := \widehat{\Omega}_{\zeta,p,s}(x; \widehat{\Delta}) := [\eta^{(1)}(\widehat{\mu}(x) + \widehat{\mathsf{w}}'\widehat{\boldsymbol{\gamma}})]^2\widehat{\mathbf{b}}_{p,s}^{(1)}(x)'\widehat{\mathbf{Q}}_{p,s}^{-1}\widehat{\boldsymbol{\Sigma}}_{p,s}\widehat{\mathbf{Q}}_{p,s}^{-1}\widehat{\mathbf{b}}_{p,s}^{(1)}(x),$$

$$\bar{\Omega}_{\zeta,p,s}(x) := \bar{\Omega}_{\zeta,p,s}(x; \widehat{\Delta}) := [\eta^{(1)}(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)]^2\widehat{\mathbf{b}}_{p,s}^{(1)}(x)'\bar{\mathbf{Q}}_{p,s}^{-1}\bar{\boldsymbol{\Sigma}}_{p,s}\bar{\mathbf{Q}}_{p,s}^{-1}\widehat{\mathbf{b}}_{p,s}^{(1)}(x), \quad \text{and}$$

$$\Omega_{\zeta,p,s}(x) := \Omega_{\zeta,p,s}(x; \widehat{\Delta}) := [\eta^{(1)}(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)]^2\widehat{\mathbf{b}}_{p,s}^{(1)}(x)'\mathbf{Q}_{0,p,s}^{-1}\boldsymbol{\Sigma}_{0,p,s}\mathbf{Q}_{0,p,s}^{-1}\widehat{\mathbf{b}}_{p,s}^{(1)}(x).$$

Recall that in the main text we always set $s = p$ and omit the dependence on s whenever there is no confusion. Thus,

$$\widehat{\mathbf{Q}}_p = \widehat{\mathbf{Q}}_{p,p}, \quad \bar{\mathbf{Q}}_p = \bar{\mathbf{Q}}_{p,p}, \quad \mathbf{Q}_{0,p} = \mathbf{Q}_{0,p,p},$$

$$\widehat{\mathbf{\Sigma}}_p = \widehat{\mathbf{\Sigma}}_{p,p}, \quad \bar{\mathbf{\Sigma}}_p = \bar{\mathbf{\Sigma}}_{p,p}, \quad \mathbf{\Sigma}_{0,p} = \mathbf{\Sigma}_{0,p,p},$$

$$\widehat{\Omega}_{\mu^{(v)},p}(x) = \widehat{\Omega}_{\mu^{(v)},p,p}(x), \quad \bar{\Omega}_{\mu^{(v)},p}(x) = \bar{\Omega}_{\mu^{(v)},p,p}(x), \quad \Omega_{\mu^{(v)},p}(x) = \Omega_{\mu^{(v)},p,p}(x),$$

$$\widehat{\Omega}_{\vartheta,p}(x) = \widehat{\Omega}_{\vartheta,p,p}(x), \quad \bar{\Omega}_{\vartheta,p}(x) = \bar{\Omega}_{\vartheta,p,p}(x), \quad \Omega_{\vartheta,p}(x) = \Omega_{\vartheta,p,p}(x),$$

$$\widehat{\Omega}_{\zeta,p}(x) = \widehat{\Omega}_{\zeta,p,p}(x), \quad \bar{\Omega}_{\zeta,p}(x) = \bar{\Omega}_{\zeta,p,p}(x), \quad \text{and} \quad \Omega_{\zeta,p}(x) = \Omega_{\zeta,p,p}(x).$$

In this supplement, however, all our results hold for a general choice of the degree and the smoothness of the basis. For ease of notation, the subscripts p and s of the above quantities are dropped hereafter:

$$\widehat{\mathbf{Q}} = \widehat{\mathbf{Q}}_{p,s}, \quad \bar{\mathbf{Q}} = \bar{\mathbf{Q}}_{p,s}, \quad \mathbf{Q}_0 = \mathbf{Q}_{0,p,s},$$

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_{p,s}, \quad \bar{\mathbf{\Sigma}} = \bar{\mathbf{\Sigma}}_{p,s}, \quad \mathbf{\Sigma}_0 = \mathbf{\Sigma}_{0,p,s},$$

$$\widehat{\Omega}_{\mu^{(v)}}(x) = \widehat{\Omega}_{\mu^{(v)},p,s}(x), \quad \bar{\Omega}_{\mu^{(v)}}(x) = \bar{\Omega}_{\mu^{(v)},p,s}(x), \quad \Omega_{\mu^{(v)}}(x) = \Omega_{\mu^{(v)},p,s}(x),$$

$$\widehat{\Omega}_{\vartheta}(x) = \widehat{\Omega}_{\vartheta,p,s}(x), \quad \bar{\Omega}_{\vartheta}(x) = \bar{\Omega}_{\vartheta,p,s}(x), \quad \Omega_{\vartheta}(x) = \Omega_{\vartheta,p,s}(x),$$

$$\widehat{\Omega}_{\zeta}(x) = \widehat{\Omega}_{\zeta,p,s}(x), \quad \bar{\Omega}_{\zeta}(x) = \bar{\Omega}_{\zeta,p,s}(x), \quad \text{and} \quad \Omega_{\zeta}(x) = \Omega_{\zeta,p,s}(x).$$

In addition, given the family Π_C of the quasi-uniform partitions defined in (SA-2.4), for any $\Delta \in \Pi$, we let $\boldsymbol{\beta}_0(\Delta) \in \mathbb{R}^{K_{p,s}}$ be any vector such that for every $v \leq p$,

$$\sup_{x \in \mathcal{X}} \left| \mu_0^{(v)}(x) - \mathbf{b}_{p,s}^{(v)}(x; \Delta)' \boldsymbol{\beta}_0(\Delta) \right| \lesssim J^{-p-1+v}.$$

Let $r_{0,v}(x; \Delta) = \mu_0^{(v)}(x) - \mathbf{b}_{p,s}^{(v)}(x; \Delta)' \boldsymbol{\beta}_0(\Delta)$ denote the corresponding approximation error. Accordingly, given the random partition $\widehat{\Delta}$, we let $\widehat{\boldsymbol{\beta}}_0 := \boldsymbol{\beta}_0(\widehat{\Delta})$, and $\widehat{r}_{0,v}(x) = \mu_0^{(v)}(x) - \widehat{\mathbf{b}}_{p,s}^{(v)}(x)' \widehat{\boldsymbol{\beta}}_0$ denote the corresponding approximation error. The existence of such vectors is guaranteed by Assumptions SA-DGP and SA-SM(v), and is verified in Lemma SA-5.5 in Section SA-5.

SA-3.1 Preliminary Lemmas

Lemma SA-3.1 (Gram). *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold. If $\frac{J \log J}{n} = o(1)$, then*

$$1 \lesssim \lambda_{\min}(\bar{\mathbf{Q}}) \leq \lambda_{\max}(\bar{\mathbf{Q}}) \lesssim 1, \quad [\bar{\mathbf{Q}}^{-1}]_{ij} \lesssim \varrho^{|i-j|} \quad w.p.a.\ 1, \quad and \quad \|\bar{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1,$$

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

If, in addition, Assumption SA-RP(ii) holds. Then,

$$1 \lesssim \lambda_{\min}(\mathbf{Q}_0) \leq \lambda_{\max}(\mathbf{Q}_0) \lesssim 1,$$

$$\|\bar{\mathbf{Q}} - \mathbf{Q}_0\| \lesssim_{\mathbb{P}} \left(\frac{J \log J}{n}\right)^{1/2} + \mathfrak{r}_{\mathsf{RP}}, \quad and \quad \|\bar{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\|_\infty \lesssim_{\mathbb{P}} \left(\frac{J \log J}{n}\right)^{1/2} + \mathfrak{r}_{\mathsf{RP}}.$$

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, SA-SM, SA-HLE and SA-RP(i) hold. If $\frac{J \log J}{n} = o(1)$, then w.p.a. 1,*

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

$$J \lesssim \inf_{x \in \mathcal{X}} \bar{\Omega}_{\vartheta}(x) \leq \sup_{x \in \mathcal{X}} \bar{\Omega}_{\vartheta}(x) \lesssim J,$$

$$J^3 \lesssim \inf_{x \in \mathcal{X}} \bar{\Omega}_{\zeta}(x) \leq \sup_{x \in \mathcal{X}} \bar{\Omega}_{\zeta}(x) \lesssim J^3.$$

If, in addition, Assumption SA-RP(ii) holds, then w.p.a. 1,

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

$$J \lesssim \inf_{x \in \mathcal{X}} \Omega_{\vartheta}(x) \leq \sup_{x \in \mathcal{X}} \Omega_{\vartheta}(x) \lesssim J,$$

$$J^3 \lesssim \inf_{x \in \mathcal{X}} \Omega_{\zeta}(x) \leq \sup_{x \in \mathcal{X}} \Omega_{\zeta}(x) \lesssim J^3.$$

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

Lemma SA-3.3 (Uniform Convergence: Variance). *Suppose that Assumptions SA-DGP, SA-SM,*

SA-HLE and *SA-RP(i)* hold. If $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} = o(1)$, then

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

Lemma SA-3.4 (Projection of Approximation Error). *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold. If $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} = o(1)$, then*

$$\sup_{x \in \mathcal{X}} \left| \widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\mathbb{E}_n \Big[\widehat{\mathbf{b}}_{p,s}(x_i)\Big(\eta_{i,1}\psi(y_i,\eta_i) - \eta^{(1)}(\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)\psi(y_i,\eta(\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0))) \Big) \Big] \right|$$

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

Lemma SA-3.5 (Uniform Consistency). *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold. If $\frac{J^{\frac{2\nu}{\nu-1}}(\log J)^{\frac{\nu}{\nu-1}}}{n} = o(1)$, then*

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0\|_\infty = o_{\mathbb{P}}(J^{-1/2}) \quad and \quad \sup_{x \in \mathcal{X}} |\widehat{\mu}(x) - \mu_0(x)| = o_{\mathbb{P}}(1).$$

Remark SA-3.1 (Side rate conditions). When $\nu \to \infty$, the rate restriction $\frac{J^{\frac{2\nu}{\nu-1}}(\log J)^{\frac{\nu}{\nu-1}}}{n} = o(1)$ tends to be $\frac{J^2\log J}{n} = o(1)$. We conjecture this rate restriction is stronger than needed. In fact, for piecewise polynomials (i.e., $s = 0$), we can show that $\frac{J^{\frac{\nu}{\nu-1}}(\log J)^{\frac{\nu}{\nu-1}}}{n} = o(1)$ suffices to establish the uniform consistency of $\widehat{\boldsymbol{\beta}}$, and this restriction is redundant in our main theorems in view of the condition $\frac{J^{\frac{\nu}{\nu-2}}(\log n)^{\frac{\nu}{\nu-2}}}{n} = o(1)$ imposed below. In other words, in this special case ($s = 0$), the condition $\frac{J^{\frac{2\nu}{\nu-1}}(\log J)^{\frac{\nu}{\nu-1}}}{n} = o(1)$ in all theorems below can be dropped. ⌟

Our result holds without imposing any smoothness restrictions on the estimation space. Specifically, the estimation procedure (SA-2.3) searches for solutions in $\mathbb{R}^{K_{p,s}}$, leading to an estimation space $\{\widehat{\mathbf{b}}_{p,s}(x)'\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^{K_{p,s}}\}$. 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) and references therein.

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), while it focuses on series-based quantile regression only. Furthermore, relative to prior work, our results allow for random partitioning schemes, formally taking into account both the potential randomness of the partition and the semi-linear regression estimation structure. Importantly, we highlight the key conditions on the possibly random partition (Assumptions SA-RP(i) and SA-RP(ii)) used to derive various properties of the Gram matrix, asymptotic variance and other quantities. ⌟

SA-3.2 Bahadur Representation

Theorem SA-3.1 (Bahadur Representation)**.** *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold and* $\frac{J^{\frac{\nu}{\nu-2}} \log n}{n} + \frac{J(\log n)^{7/3}}{n} + \frac{J^{\frac{2\nu}{\nu-1}}(\log n)^{\frac{\nu}{\nu-1}}}{n} + \frac{\log n}{J} = o(1)$. *Then,*

(i) $\widehat{\mu}^{(v)}(x)$ *satisfies that*

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

(ii) $\widehat{\vartheta}(x, \widehat{\mathsf{w}})$ *satisfies that*

$$
\sup_{x \in \mathcal{X}} \left| \widehat{\vartheta}(x, \widehat{\mathsf{w}}) - \vartheta_0(x, \mathsf{w}) + \eta^{(1)}(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)\widehat{\mathbf{b}}_{p,s}(x)' \bar{\mathbf{Q}}^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\eta_{i,1}\psi(y_i,\eta_i)] \right|
$$
$$
\lesssim_{\mathbb{P}} \left(\frac{J \log n}{n}\right)^{3/4} \log n + J^{-\frac{p+1}{2}} \left(\frac{J \log^2 n}{n}\right)^{1/2} + J^{-p-1} + \mathfrak{r}_\gamma + \|\widehat{\mathsf{w}} - \mathsf{w}\|.
$$

(iii) $\widehat{\zeta}(x, \widehat{\mathsf{w}})$ *satisfies that*

$$
\sup_{x \in \mathcal{X}} \left| \widehat{\zeta}(x, \widehat{\mathsf{w}}) - \zeta_0(x, \mathsf{w}) + \eta^{(1)}(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0)\widehat{\mathbf{b}}_{p,s}^{(1)}(x)' \bar{\mathbf{Q}}^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\eta_{i,1}\psi(y_i,\eta_i)] \right|
$$
$$
\lesssim_{\mathbb{P}} \left(\frac{J \log n}{n}\right)^{1/2} + J \left\{ \left(\frac{J \log n}{n}\right)^{3/4} \log n + J^{-\frac{p+1}{2}} \left(\frac{J \log^2 n}{n}\right)^{1/2} + J^{-p-1} + \mathfrak{r}_\gamma \right\}
$$
$$
+ \|\widehat{\mathsf{w}} - \mathsf{w}\| \left(1 + J \left(\frac{J \log n}{n}\right)^{1/2}\right).
$$

The following corollary is an immediate result of Lemma SA-3.3 and Theorem SA-3.1, and hence

its proof is omitted.

Corollary SA-3.1 (Uniform Convergence). *Suppose that the conditions of Theorem SA-3.1 hold and $\frac{J(\log n)^5}{n} \lesssim 1$. Then*

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

If, in addition, $\|\widehat{\mathsf{w}} - \mathsf{w}\| \lesssim_{\mathbb{P}} \Big(\frac{J \log n}{n}\Big)^{1/2} + J^{-p-1}$, then

$$\sup_{x \in \mathcal{X}} |\widehat{\vartheta}(x, \widehat{\mathsf{w}}) - \vartheta_0(x, \mathsf{w})| \lesssim_{\mathbb{P}} \Big(\frac{J \log n}{n}\Big)^{1/2} + J^{-p-1} \quad \text{and}$$

$$\sup_{x \in \mathcal{X}} |\widehat{\zeta}(x, \widehat{\mathsf{w}}) - \zeta_0(x, \mathsf{w})| \lesssim_{\mathbb{P}} J \Big(\Big(\frac{J \log n}{n}\Big)^{1/2} + J^{-p-1} \Big).$$

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

Theorem SA-3.2 (Variance Estimate). *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold. If $\frac{J^{\frac{\nu}{\nu-2}} (\log n)^{\frac{\nu}{\nu-2}}}{n} + \frac{J^{\frac{2\nu}{\nu-1}} (\log n)^{\frac{\nu}{\nu-1}}}{n} + \frac{J(\log n)^5}{n} + \frac{\log n}{J} = o(1)$ and $\|\widehat{\mathsf{w}} - \mathsf{w}\| \lesssim_{\mathbb{P}} \Big(\frac{J \log n}{n}\Big)^{1/2} + J^{-p-1}$, then*

$$\Big\| \widehat{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}} \Big\| \lesssim_{\mathbb{P}} J^{-p-1} + \Big(\frac{J \log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2},$$

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

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

$$\sup_{x \in \mathcal{X}} \Big| \widehat{\Omega}_{\zeta}(x) - \bar{\Omega}_{\zeta}(x) \Big| \lesssim_{\mathbb{P}} J^3 \Big(J^{-p-1} + \Big(\frac{J \log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2} \Big).$$

If, in addition, Assumption SA-RP(ii) holds, then

$$\Big\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0 \Big\| \lesssim_{\mathbb{P}} J^{-p-1} + \Big(\frac{J \log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2} + \mathfrak{r}_{\mathsf{RP}},$$

$$\sup_{x \in \mathcal{X}} \Big| \widehat{\Omega}_{\mu^{(v)}}(x) - \Omega_{\mu^{(v)}}(x) \Big| \lesssim_{\mathbb{P}} J^{1+2v} \Big(J^{-p-1} + \Big(\frac{J \log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2} + \mathfrak{r}_{\mathsf{RP}} \Big),$$

$$\sup_{x \in \mathcal{X}} \Big| \widehat{\Omega}_{\vartheta}(x) - \Omega_{\vartheta}(x) \Big| \lesssim_{\mathbb{P}} J \Big(J^{-p-1} + \Big(\frac{J \log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2} + \mathfrak{r}_{\mathsf{RP}} \Big), \quad \text{and}$$

$$\sup_{x \in \mathcal{X}} \Big| \widehat{\Omega}_{\zeta}(x) - \Omega_{\zeta}(x) \Big| \lesssim_{\mathbb{P}} J^3 \Big(J^{-p-1} + \Big(\frac{J \log n}{n^{1-\frac{2}{\nu}}}\Big)^{1/2} + \mathfrak{r}_{\mathsf{RP}} \Big).$$

17

Remark SA-3.3 (Improvements over literature)**.** Theorem SA-3.1 and Corollary SA-3.1 construct the Bahadur representation and uniform convergence of nonlinear binscatter-based M-estimators, which improve upon prior results in the literature in at least two aspects. First, our results allow for random partitioning schemes, and the key condition imposed on the partition is Assumption SA-RP(i), i.e., the conditional independence between the partition and the outcome and the quasi-uniformity of the partition. The "convergence" of the random partition (Assumption SA-RP(ii)) is not required, which implies that our results can accommodate more complex partitioning schemes other than evenly-spaced or empirical-quantile-spaced partitions.

Second, our results are established under weaker rate restrictions. Specifically, we require $J^{\frac{8}{3}}/n = o(1)$ up to $\log n$ terms when $\nu \geq 4$, thus accommodating IMSE-optimal piecewise constant binscatter estimators. In fact, for piecewise polynomials ($s = 0$), we can show that the Bahadur representation still holds under $J/n = o(1)$ up to $\log n$ terms when a subexponential moment restriction holds for the "score" $\psi(y_i, \eta_i)$, which is analogous to the result for kernel-based estimators in the literature (see, e.g., Kong et al., 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 representation and uniform convergence of the resulting estimators under $J^4/n^{1-\varepsilon} = o(1)$ for some $\varepsilon > 0$.

The results in Belloni et al. (2019) are slightly stronger than 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 representation can be extended to achieve the same level of uniformity. In general, the parameter of interest (SA-2.1) and the estimator (SA-2.2) 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 representation 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. ⌟

SA-3.3 Pointwise Inference

Starting from this section, we consider statistical inference on $\mu_0^{(v)}(x)$, $\vartheta_0(x, \mathsf{w})$ and $\zeta_0(x, \mathsf{w})$ based on the following Studentized t-statistics:

$$T_{\mu^{(v)}, p}(x) = \frac{\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}},$$

$$T_{\vartheta, p}(x) = \frac{\widehat{\vartheta}(x, \widehat{\mathsf{w}}) - \vartheta_0(x, \mathsf{w})}{\sqrt{\widehat{\Omega}_\vartheta(x)/n}} \quad \text{and}$$

$$T_{\zeta, p}(x) = \frac{\widehat{\zeta}(x, \widehat{\mathsf{w}}) - \zeta_0(x, \mathsf{w})}{\sqrt{\widehat{\Omega}_\zeta(x)/n}}.$$

The next theorem shows the pointwise asymptotic normality of the binscatter estimators.

Theorem SA-3.3 (Pointwise Asymptotic Distribution)**.** *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold, $\sup_{x \in \mathcal{X}} \mathbb{E}[|\psi(y_i, \eta_i)|^\nu | x_i = x] \lesssim 1$ for some $\nu \geq 3$, and $\frac{J^{\frac{\nu}{\nu-2}} (\log n)^{\frac{\nu}{\nu-2}}}{n} + \frac{J^{\frac{2\nu}{\nu-1}} (\log n)^{\frac{\nu}{\nu-1}}}{n} + nJ^{-2p-3} = o(1)$. Then the following conclusions hold:*

(i) For $\widehat{\mu}^{(v)}(x)$,

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}(T_{\mu^{(v)}, p}(x) \leq u) - \Phi(u) \right| = o(1), \quad \text{for each } x \in \mathcal{X}.$$

(ii) For $\widehat{\vartheta}(x, \widehat{\mathsf{w}})$, if, in addition, $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(\sqrt{J/n})$, then

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

(iii) For $\widehat{\zeta}(x, \widehat{\mathsf{w}})$, if, in addition, $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(\sqrt{J^3/n} + (\log n)^{-1/2})$, then

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

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 et al. (2019), which focuses on series-based quantile regression only. Furthermore, relative to prior work, our results allow for more general partitioning schemes,

formally take into account the potential randomness of the partition, and account for the semi-linear regression estimation structure. The key condition imposed on the partition for pointwise inference is Assumption SA-RP(i), and the "convergence" of the random partition is not required.

⌐

SA-3.4   Integrated Mean Squared Error

In this section we give a Nagar-type approximate IMSE expansion for each of the three estimators $\widehat{\mu}^{(v)}(x)$, $\widehat{\vartheta}(x,\widehat{\mathsf{w}})$ and $\widehat{\zeta}(x,\widehat{\mathsf{w}})$, with explicit characterization of the leading constants. Define

$$r_{0,v}^{\star}(x) = \frac{J^{-p-1+v}\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x - \tau_x^{\mathsf{L}}}{h_x}\Big)$$

where $\mathscr{E}_m(\cdot)$ is the mth Bernoulli polynomial for each $m \in \mathbb{Z}_+$, τ_x^{L} is the start of the interval in the non-random partition Δ_0 containing x and h_x denotes its length.

Theorem SA-3.4 (IMSE). *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP (including SA-RP(ii)) hold. Let $\omega(x)$ be a continuous weighting function over $\mathcal{X}$ bounded away from zero. Also, assume that $\frac{J^{\frac{\nu}{\nu-2}}\log n}{n} + \frac{J^{\frac{2\nu}{\nu-1}}(\log n)^{\frac{\nu}{\nu-1}}}{n} + \frac{J(\log n)^7}{n} + \frac{(\log n)^2}{J} = o(1)$.*

(i) For $\widehat{\mu}^{(v)}(x)$,

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

where

$$\mathbb{E}[\mathtt{AISE}_{\mu^{(v)}}|\mathbf{X},\mathbf{W},\widehat{\Delta}] = \frac{J^{1+2v}}{n}\mathscr{V}_n(p,s,v) + J^{-2(p+1-v)}\mathscr{B}_n(p,s,v) + o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n} + J^{-2(p+1-v)}\Big),$$

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

$$\mathscr{B}_n(p,s,v) := J^{2p+2-2v}\int_{\mathcal{X}}\Big(r_{0,v}^{\star}(x) - \mathbf{b}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i,\mathbf{w}_i)r_{0,0}^{\star}(x_i)]\Big)^2\omega(x)dx \lesssim 1.$$

(ii) For $\widehat{\vartheta}(x,\widehat{\mathsf{w}})$, if $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$, then

$$\int_{\mathcal{X}}\Big(\widehat{\vartheta}(x,\widehat{\mathsf{w}}) - \vartheta_0(x,\mathsf{w})\Big)^2\omega(x)dx = \mathtt{AISE}_{\vartheta} + o_{\mathbb{P}}\Big(\frac{J}{n} + J^{-2(p+1)}\Big)$$

20

where

$$\mathbb{E}[\texttt{AISE}_\vartheta|\mathbf{X},\mathbf{W},\widehat{\Delta}] = \frac{J}{n}\mathcal{V}_n(p,s) + J^{-2(p+1)}\mathcal{B}_n(p,s) + o_\mathbb{P}\Big(\frac{J}{n} + J^{-2(p+1)}\Big),$$

$$\mathcal{V}_n(p,s) := J^{-1}\,\mathrm{trace}\Big(\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\int_\mathcal{X}\eta_{0,1}(x,\mathsf{w})^2\mathbf{b}_{p,s}(x)\mathbf{b}_{p,s}(x)'\omega(x)dx\Big) \asymp 1,$$

$$\mathcal{B}_n(p,s) := J^{2p+2}\int_\mathcal{X}\Big[\eta_{0,1}(x,\mathsf{w})\Big(r_{0,0}^\star(x) - \mathbf{b}_{p,s}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i,\mathbf{w}_i)r_{0,0}^\star(x_i)]\Big)\Big]^2\omega(x)dx \lesssim 1.$$

(iii) For $\widehat{\zeta}(x,\widehat{\mathsf{w}})$, if $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_\mathbb{P}(\sqrt{J^3/n} + J^{-p} + (\log n)^{-1/2})$, then

$$\int_\mathcal{X}\Big(\widehat{\zeta}(x,\widehat{\mathsf{w}}) - \zeta_0(x,\mathsf{w})\Big)^2\omega(x)dx = \texttt{AISE}_\zeta + o_\mathbb{P}\Big(\frac{J^3}{n} + J^{-2p}\Big)$$

where

$$\mathbb{E}[\texttt{AISE}_\zeta|\mathbf{X},\mathbf{W},\widehat{\Delta}] = \frac{J^3}{n}\mathcal{V}_n(p,s) + J^{-2p}\mathcal{B}_n(p,s) + o_\mathbb{P}\Big(\frac{J^3}{n} + J^{-2p}\Big),$$

$$\mathcal{V}_n(p,s) := J^{-3}\,\mathrm{trace}\Big(\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\int_\mathcal{X}\eta_{0,1}(x,\mathsf{w})^2\mathbf{b}_{p,s}^{(1)}(x)\mathbf{b}_{p,s}^{(1)}(x)'\omega(x)dx\Big) \asymp 1,$$

$$\mathcal{B}_n(p,s) := J^{2p}\int_\mathcal{X}\Big[\eta_{0,1}(x,\mathsf{w})\Big(r_{0,1}^\star(x) - \mathbf{b}_{p,s}^{(1)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i,\mathbf{w}_i)r_{0,0}^\star(x_i)]\Big)\Big]^2\omega(x)dx \lesssim 1.$$

In general, $\mathcal{B}_n(p,s,v) \gtrsim 1$ (see Remark SA-3.7 in Cattaneo et al. (2024b)), and thus 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-3.4, one may design a data-driven procedure to select the IMSE-optimal number of bins for nonlinear binscatter-based M-estimators.

Remark SA-3.5 (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 nonlinear series estimators and binscatter (piecewise polynomials and splines) nonlinear series estimators in particular. Furthermore, our results allow for random partitioning schemes, formally take into account the potential randomness of the partition, and account for the semi-linear regression estimation structure. We highlight the key conditions imposed on the partition (Assumption SA-RP) for the approximate IMSE expansion. The "convergence" of the random partition (Assumption SA-RP(ii)) is needed to derive the non-random variance and bias constants $\mathcal{V}_n(p,s)$ and $\mathcal{B}_n(p,s)$. ⌟

SA-3.5   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 processes $T_{\mu^{(v)},p}(\cdot)$, $T_{\vartheta,p}(\cdot)$ and $T_{\zeta,p}(\cdot)$ can be approximated by Gaussian processes in a proper sense at certain rate.

Theorem SA-3.5 (Strong Approximation)**.** *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold,*

$$\frac{J(\log n)^2}{n^{1-\frac{2}{\nu}}} + \Big(\frac{J(\log n)^7}{n}\Big)^{1/2} + nJ^{-2p-3} + \frac{(\log n)^2}{J^{p+1}} + nJ^{-1}\mathfrak{r}_\gamma^2 = o(a_n^{-2}) \quad and \quad \frac{J^{\frac{2\nu}{\nu-1}}(\log n)^{\frac{\nu}{\nu-1}}}{n} = o(1).$$

Then the following conclusions hold:

(i) *On a properly enriched probability space, there exists some $K_{p,s}$-dimensional standard normal random vector $\mathbf{N}_{K_{p,s}}$ such that for any $\xi > 0$,*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}} |T_{\mu^{(v)},p}(x) - \bar{Z}_{\mu^{(v)},p}(x)| > \xi a_n^{-1} \Big) = o(1), \quad \bar{Z}_{\mu^{(v)},p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\widehat{\mathbf{T}}_s'\bar{\mathbf{Q}}^{-1}\bar{\mathbf{\Sigma}}^{1/2}}{\sqrt{\bar{\Omega}_{\mu^{(v)}}(x)}}\mathbf{N}_{K_{p,s}}.$$

If Assumption SA-RP(ii) also holds with $\mathfrak{r}_{\mathsf{RP}} = o(a_n^{-1}(\log n)^{-1/2})$, then

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}} |T_{\mu^{(v)},p}(x) - Z_{\mu^{(v)},p}(x)| > \xi a_n^{-1} \Big) = o(1), \quad Z_{\mu^{(v)},p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\mathbf{T}_s'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0^{1/2}}{\sqrt{\Omega_{\mu^{(v)}}(x)}}\mathbf{N}_{K_{p,s}}.$$

(ii) *If $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n})$, then on a properly enriched probability space there exists some $K_{p,s}$-dimensional standard normal random vector $\mathbf{N}_{K_{p,s}}$ such that for any $\xi > 0$,*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}} |T_{\vartheta,p}(x) - \bar{Z}_{\vartheta,p}(x)| > \xi a_n^{-1} \Big) = o(1), \quad \bar{Z}_{\vartheta,p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}(x)'\widehat{\mathbf{T}}_s'\eta_{0,1}(x,\mathsf{w})\bar{\mathbf{Q}}^{-1}}{\sqrt{\bar{\Omega}_{\vartheta}(x)}}\bar{\mathbf{\Sigma}}^{1/2}\mathbf{N}_{K_{p,s}}.$$

If Assumption SA-RP(ii) also holds with $\mathfrak{r}_{\mathsf{RP}} = o(a_n^{-1}(\log n)^{-1/2})$, then

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}} |T_{\vartheta,p}(x) - Z_{\vartheta,p}(x)| > \xi a_n^{-1} \Big) = o(1), \quad Z_{\vartheta,p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}(x)'\mathbf{T}_s'\eta_{0,1}(x,\mathsf{w})\mathbf{Q}_0^{-1}}{\sqrt{\Omega_{\vartheta}(x)}}\mathbf{\Sigma}_0^{1/2}\mathbf{N}_{K_{p,s}}.$$

(iii) *If $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(a_n^{-1}(\sqrt{J^3/n} + (\log n)^{-1/2}))$, then on a properly enriched probability space*

there exists some $K_{p,s}$-dimensional standard normal random vector $\mathbf{N}_{K_{p,s}}$ such that for any $\xi > 0$,

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |T_{\zeta,p}(x) - \bar{Z}_{\zeta,p}(x)| > \xi a_n^{-1}\right) = o(1), \quad \bar{Z}_{\zeta,p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}^{(1)}(x)'\widehat{\mathbf{T}}_s'\eta_{0,1}(x,\mathsf{w})\bar{\mathbf{Q}}^{-1}}{\sqrt{\bar{\Omega}_\zeta(x)}}\bar{\boldsymbol{\Sigma}}^{1/2}\mathbf{N}_{K_{p,s}}.$$

If Assumption SA-RP(ii) also holds with $\mathfrak{r}_{\mathsf{RP}} = o(a_n^{-1}(\log n)^{-1/2})$, then

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |T_{\zeta,p}(x) - Z_{\zeta,p}(x)| > \xi a_n^{-1}\right) = o(1), \quad Z_{\zeta,p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}^{(1)}(x)'\mathbf{T}_s'\eta_{0,1}(x,\mathsf{w})\mathbf{Q}_0^{-1}}{\sqrt{\Omega_\zeta(x)}}\boldsymbol{\Sigma}_0^{1/2}\mathbf{N}_{K_{p,s}}.$$

The approximating processes $Z_{\mu^{(v)},p}(\cdot)$, $Z_{\vartheta,p}(\cdot)$ and $Z_{\zeta,p}(\cdot)$ are Gaussian processes conditional on $\widehat{\Delta}$ by construction. In practice, one can replace all unknowns in $Z_{\mu^{(v)},p}(\cdot)$, $Z_{\vartheta,p}(\cdot)$ and $Z_{\zeta,p}(\cdot)$ by their sample analogues, and then construct the following feasible (conditional) Gaussian processes:

$$\widehat{Z}_{\mu^{(v)},p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\widehat{\mathbf{T}}_s'\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)}}\mathbf{N}_{K_{p,s}}^\star = \frac{\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)}}\mathbf{N}_{K_{p,s}}^\star,$$

$$\widehat{Z}_{\vartheta,p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}(x)'\widehat{\mathbf{T}}_s'\widehat{\eta}_{0,1}(x,\widehat{\mathsf{w}})\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}_\vartheta(x)}}\mathbf{N}_{K_{p,s}}^\star = \frac{\widehat{\mathbf{b}}_{p,s}(x)'\widehat{\eta}_{0,1}(x,\widehat{\mathsf{w}})\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}_\vartheta(x)}}\mathbf{N}_{K_{p,s}}^\star,$$

$$\widehat{Z}_{\zeta,p}(x) = \frac{\widehat{\mathbf{b}}_{p,0}^{(1)}(x)'\widehat{\mathbf{T}}_s'\widehat{\eta}_{0,1}(x,\widehat{\mathsf{w}})\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}_\zeta(x)}}\mathbf{N}_{K_{p,s}}^\star = \frac{\widehat{\mathbf{b}}_{p,s}^{(1)}(x)'\widehat{\eta}_{0,1}(x,\widehat{\mathsf{w}})\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}_\zeta(x)}}\mathbf{N}_{K_{p,s}}^\star,$$

where $\mathbf{N}_{K_{p,s}}^\star$ denotes a $K_{p,s}$-dimensional standard normal vector independent of the data $\mathbf{D}$.

For ease of presentation, from now on we will always require a fast convergence rate of $\widehat{\mathsf{w}}$: $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_\mathbb{P}(a_n^{-1}\sqrt{J/n})$. Nevertheless, it should be clear that as shown in Theorem SA-3.5, such a rate restriction on $\widehat{\mathsf{w}}$ can be different for inference of $\vartheta_0(x,\mathsf{w})$ and $\zeta_0(x,\mathsf{w})$ and are unnecessary for inference of $\mu_0^{(v)}(x)$.

Theorem SA-3.6 (Plug-in Approximation). *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold,*

$$\frac{J(\log n)^2}{n^{1-\frac{2}{\nu}}} + \left(\frac{J(\log n)^7}{n}\right)^{1/2} + nJ^{-2p-3} + \frac{(\log n)^2}{J^{p+1}} + nJ^{-1}\mathfrak{r}_\gamma^2 = o(a_n^{-2}),$$

$$\frac{J^{\frac{2\nu}{\nu-1}}(\log n)^{\frac{\nu}{\nu-1}}}{n} = o(1), \quad and \quad \|\widehat{\mathsf{w}} - \mathsf{w}\| = o_\mathbb{P}(a_n^{-1}\sqrt{J/n}).$$

23

Then on a properly enriched probability space, there exists a $K_{p,s}$-dimensional standard normal random vector $\mathbf{N}^\star_{K_{p,s}}$ independent of $\mathbf{D}$ such that for any $\xi > 0$,

(i) $\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_{\mu^{(v)},p}(x) - \bar{Z}_{\mu^{(v)},p}(x)| > \xi a_n^{-1} \Big| \mathbf{D} \Big) = o_{\mathbb{P}}(1),$

(ii) $\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_{\vartheta,p}(x) - \bar{Z}_{\vartheta,p}(x)| > \xi a_n^{-1} \Big| \mathbf{D} \Big) = o_{\mathbb{P}}(1),$

(iii) $\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_{\zeta,p}(x) - \bar{Z}_{\zeta,p}(x)| > \xi a_n^{-1} \Big| \mathbf{D} \Big) = o_{\mathbb{P}}(1).$

If Assumption SA-RP(ii) also holds with $\mathfrak{r}_{\mathtt{RP}} = o(a_n^{-1}(\log n)^{-1/2})$, then

(iv) $\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_{\mu^{(v)},p}(x) - Z_{\mu^{(v)},p}(x)| > \xi a_n^{-1} \Big| \mathbf{D} \Big) = o_{\mathbb{P}}(1),$

(v) $\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_{\vartheta,p}(x) - Z_{\vartheta,p}(x)| > \xi a_n^{-1} \Big| \mathbf{D} \Big) = o_{\mathbb{P}}(1),$

(vi) $\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_{\zeta,p}(x) - Z_{\zeta,p}(x)| > \xi a_n^{-1} \Big| \mathbf{D} \Big) = o_{\mathbb{P}}(1).$

Remark SA-3.6 (Improvements over literature)**.** Theorems SA-3.5 and SA-3.6 provide empirical researchers with powerful tools for uniform inference based on binscatter methods. Importantly, we allow for random partitioning schemes, formally take into account the potential randomness of the partition, and construct a novel strong approximation of nonlinear binscatter-based M-estimators under mild rate restrictions. For $a_n = \sqrt{\log n}$ and $\nu \geq 4$, we require $J^{\frac{8}{3}}/n = o(1)$, up to $\log n$ terms. In the literature, similar results were only available in some special cases under stringent rate restrictions. For instance, Belloni et al. (2019) considers strong approximations of more 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 $\mathcal{U} \subset (0,1)$. On the other hand, our methods allow for other loss functions (e.g., Huber regression), a large class of random partitions, and semi-linear covariate adjustment, leading to new results that were previously unavailable in the literature. $\lrcorner$

Theorems SA-3.5 and SA-3.6 offer a way to approximate the distribution of the *whole t*-statistic process based on $\widehat{\mu}^{(v)}(\cdot)$, $\widehat{\vartheta}(\cdot, \widehat{\mathsf{w}})$ or $\widehat{\zeta}(\cdot, \widehat{\mathsf{w}})$. A direct application of these results is the distributional approximations to the suprema of these *t*-statistic processes.

Theorem SA-3.7 (Supremum Approximation)**.** *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP (including SA-RP(ii)) hold,*

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

$$\frac{J^{\frac{2\nu}{\nu-1}}(\log n)^{\frac{\nu}{\nu-1}}}{n} = o(1), \quad \|\widehat{\mathsf{w}} - \mathsf{w}\| = o_\mathbb{P}\Big(\sqrt{\frac{J}{n\log J}}\Big), \quad and \quad \mathfrak{r}_{\mathtt{RP}} = o\Big(\frac{1}{\sqrt{\log n\,\log J}}\Big).$$

Then,

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|T_{\mu^{(v)},p}(x)| \leq u\Big) - \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\widehat{Z}_{\mu^{(v)},p}(x)| \leq u\Big|\mathbf{D}\Big)\Big| = o_\mathbb{P}(1),$$

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|T_{\vartheta,p}(x)| \leq u\Big) - \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\widehat{Z}_{\vartheta,p}(x)| \leq u\Big|\mathbf{D}\Big)\Big| = o_\mathbb{P}(1), \quad and$$

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|T_{\zeta,p}(x)| \leq u\Big) - \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\widehat{Z}_{\zeta,p}(x)| \leq u\Big|\mathbf{D}\Big)\Big| = o_\mathbb{P}(1).$$

SA-3.6 Confidence Bands

Let

$$\widehat{I}_{\mu^{(v)},p}(x) = \Big[\widehat{\mu}^{(v)}(x) \pm \mathfrak{c}_{\mu^{(v)}}\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}\Big],$$

$$\widehat{I}_{\vartheta,p}(x,\mathsf{w}) = \Big[\widehat{\vartheta}(x,\widehat{\mathsf{w}}) \pm \mathfrak{c}_\vartheta\sqrt{\widehat{\Omega}_\vartheta(x)/n}\Big] \quad \text{and}$$

$$\widehat{I}_{\zeta,p}(x,\mathsf{w}) = \Big[\widehat{\zeta}(x,\widehat{\mathsf{w}}) \pm \mathfrak{c}_\zeta\sqrt{\widehat{\Omega}_\zeta(x)/n}\Big]$$

be confidence bands for $\mu_0^{(v)}(\cdot)$, $\vartheta_0(\cdot,\mathsf{w})$ and $\zeta_0(\cdot,\mathsf{w})$ respectively, where $\mathfrak{c}_{\mu^{(v)}}$, $\mathfrak{c}_\vartheta$ and $\mathfrak{c}_\zeta$ are corresponding critical values to be specified. Recall that w here is taken as a fixed evaluation point for the control variables, and these bands are constructed based on a certain choice of J and the pth-order binscatter basis. Using the previous results, we have the following theorem.

Theorem SA-3.8. *Suppose that Assumptions SA-DGP, SA-SM, SA-HLE and SA-RP(i) hold,*

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

$$\frac{J^{\frac{2\nu}{\nu-1}}(\log n)^{\frac{\nu}{\nu-1}}}{n} = o(1), \quad and \quad \|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}\left(\sqrt{\frac{J}{n\log J}}\right).$$

(i) If $\mathfrak{c}_{\mu^{(v)}} = \inf\left\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}}|\widehat{Z}_{\mu^{(v)},p}(x)| \le c \,|\mathbf{D}] \ge 1-\alpha\right\}$, *then*

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

(ii) If $\mathfrak{c}_\vartheta = \inf\left\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}}|\widehat{Z}_{\vartheta,p}(x)| \le c \,|\mathbf{D}] \ge 1-\alpha\right\}$, *then*

$$\mathbb{P}\left[\vartheta_0(x,\mathsf{w}) \in \widehat{I}_{\vartheta,p}(x,\mathsf{w}), \text{ for all } x \in \mathcal{X}\right] = 1-\alpha+o(1).$$

(iii) If $\mathfrak{c}_\zeta = \inf\left\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}}|\widehat{Z}_{\zeta,p}(x)| \le c \,|\mathbf{D}] \ge 1-\alpha\right\}$, *then*

$$\mathbb{P}\left[\zeta_0(x,\mathsf{w}) \in \widehat{I}_{\zeta,p}(x,\mathsf{w}), \text{ for all } x \in \mathcal{X}\right] = 1-\alpha+o(1).$$

Remark SA-3.7. The above results construct valid uniform confidence bands for nonlinear binscatter-based M-estimators under mild rate restrictions. Specifically, when $\nu \ge 4$, we require $J^{\frac{8}{3}}/n = o(1)$, up to $\log n$ terms. In contrast, Belloni et al. (2019) considers more 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 $\varepsilon > 0$. Furthermore, our results allow for random partitioning schemes, formally taking its randomness and generic structure. The key condition imposed on the partition for the validity of confidence bands is Assumption SA-RP(i), but the "convergence" of the random partition (Assumption SA-RP(ii)) is not necessary. ⌟

SA-3.7 Parametric Specification Tests

As another application, we can test parametric specifications of $\mu_0^{(v)}(x)$, $\vartheta_0(x,\mathsf{w})$ and $\zeta_0(x,\mathsf{w})$. We introduce the following tests:

$$\dot{\mathsf{H}}_0^{\mu^{(v)}} : \quad \sup_{x\in\mathcal{X}}\left|\mu_0^{(v)}(x) - m^{(v)}(x;\boldsymbol{\theta})\right| = 0, \quad \text{for some } \boldsymbol{\theta}, \qquad vs.$$

$$\dot{\mathsf{H}}_\mathrm{A}^{\mu^{(v)}} : \quad \sup_{x\in\mathcal{X}}\left|\mu_0^{(v)}(x) - m^{(v)}(x;\boldsymbol{\theta})\right| > 0, \quad \text{for all } \boldsymbol{\theta}.$$

where $m(x; \boldsymbol{\theta})$ is some known function depending on some finite dimensional parameter $\boldsymbol{\theta}$. 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 $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\gamma}}$ as consistent estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}_0$ under $\dot{\mathsf{H}}_0^{\mu^{(v)}}$. Then we rely on the following test statistic:

$$\dot{T}_{\mu^{(v)}, p}(x) := \frac{\widehat{\mu}^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}.$$

The null hypothesis is rejected if $\sup_{x \in \mathcal{X}} |\dot{T}_{\mu^{(v)}, p}(x)| > \mathfrak{c}_{\mu^{(v)}}$ for some critical value $\mathfrak{c}_{\mu^{(v)}}$.

Similarly, to test the specification of $\vartheta_0(x, \mathsf{w})$, we introduce

$$\dot{\mathsf{H}}_0^\vartheta : \quad \sup_{x \in \mathcal{X}} \left| \vartheta_0(x, \mathsf{w}) - M(x, \mathsf{w}; \boldsymbol{\theta}, \boldsymbol{\gamma}_0) \right| = 0, \quad \text{for some } \boldsymbol{\theta}, \quad vs.$$

$$\dot{\mathsf{H}}_A^\vartheta : \quad \sup_{x \in \mathcal{X}} \left| \vartheta_0(x, \mathsf{w}) - M(x, \mathsf{w}; \boldsymbol{\theta}, \boldsymbol{\gamma}_0) \right| > 0, \quad \text{for all } \boldsymbol{\theta}.$$

where $M(x, \mathsf{w}; \boldsymbol{\theta}, \boldsymbol{\gamma}_0) = \eta(m(x; \boldsymbol{\theta}) + \mathsf{w}' \boldsymbol{\gamma}_0)$. We rely on the following test statistic:

$$\dot{T}_{\vartheta, p}(x) := \frac{\widehat{\vartheta}(x, \widehat{\mathsf{w}}) - M(x, \widehat{\mathsf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}_\vartheta(x)/n}}.$$

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

To test the specification of $\zeta_0(x, \mathsf{w})$, we introduce

$$\dot{\mathsf{H}}_0^\zeta : \quad \sup_{x \in \mathcal{X}} \left| \zeta_0(x, \mathsf{w}) - M^{(1)}(x, \mathsf{w}; \boldsymbol{\theta}, \boldsymbol{\gamma}_0) \right| = 0, \quad \text{for some } \boldsymbol{\theta}, \quad vs.$$

$$\dot{\mathsf{H}}_A^\zeta : \quad \sup_{x \in \mathcal{X}} \left| \zeta_0(x, \mathsf{w}) - M^{(1)}(x, \mathsf{w}; \boldsymbol{\theta}, \boldsymbol{\gamma}_0) \right| > 0, \quad \text{for all } \boldsymbol{\theta}.$$

where $M^{(1)}(x, \mathsf{w}; \boldsymbol{\theta}, \boldsymbol{\gamma}_0) := \eta^{(1)}(m(x; \boldsymbol{\theta}) + \mathsf{w}' \boldsymbol{\gamma}_0) m^{(1)}(x; \boldsymbol{\theta})$. We rely on the following test statistic:

$$\dot{T}_{\zeta, p}(x) := \frac{\widehat{\zeta}(x, \widehat{\mathsf{w}}) - M^{(1)}(x, \widehat{\mathsf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}_\zeta(x)/n}}.$$

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

Theorem SA-3.9 (Specification Tests)**.** *Suppose that the conditions in Theorem SA-3.8 hold.*

27

(i) Let $\mathfrak{c}_{\mu^{(v)}} = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x \in \mathcal{X}} |\widehat{Z}_{\mu^{(v)},p}(x)| \leq c|\mathbf{D}] \geq 1 - \alpha\}$.

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

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} |\dot{T}_{\mu^{(v)},p}(x)| > \mathfrak{c}_{\mu^{(v)}}\right] = \alpha.$$

Under $\dot{\mathsf{H}}_A^{\mu^{(v)}}$, if there exist some fixed $\bar{\boldsymbol{\theta}}$ such that $\sup_{x \in \mathcal{X}} |m^{(v)}(x; \widetilde{\boldsymbol{\theta}}) - m^{(v)}(x; \bar{\boldsymbol{\theta}})| = o_{\mathbb{P}}(1)$, and $J^v \left(\frac{J \log J}{n}\right)^{1/2} = o(1)$, then

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} |\dot{T}_{\mu^{(v)},p}(x)| > \mathfrak{c}_{\mu^{(v)}}\right] = 1.$$

(ii) Let $\mathfrak{c}_{\vartheta} = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x \in \mathcal{X}} |\widehat{Z}_{\vartheta,p}(x)| \leq c|\mathbf{D}] \geq 1 - \alpha\}$.

Under $\dot{\mathsf{H}}_0^{\vartheta}$, if $\sup_{x \in \mathcal{X}} |\vartheta_0(x, \mathsf{w}) - M(x, \widehat{\mathsf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})| = o_{\mathbb{P}}\left(\sqrt{\frac{J^{1+2v}}{n \log J}}\right)$, then

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

Under $\dot{\mathsf{H}}_A^{\vartheta}$, if there exist some fixed $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\gamma}}$ such that $\sup_{x \in \mathcal{X}} |M(x, \widehat{\mathsf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}}) - M(x, \mathsf{w}; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}})| = o_{\mathbb{P}}(1)$, and $J^v \left(\frac{J \log J}{n}\right)^{1/2} = o(1)$, then

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

(iii) Let $\mathfrak{c}_{\zeta} = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x \in \mathcal{X}} |\widehat{Z}_{\zeta,p}(x)| \leq c|\mathbf{D}] \geq 1 - \alpha\}$.

Under $\dot{\mathsf{H}}_0^{\zeta}$, if $\sup_{x \in \mathcal{X}} |\zeta_0(x, \mathsf{w}) - M^{(1)}(x, \widehat{\mathsf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})| = o_{\mathbb{P}}\left(\sqrt{\frac{J^{1+2v}}{n \log J}}\right)$, then

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

Under $\dot{\mathsf{H}}_A^{\zeta}$, if there exist some fixed $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\gamma}}$ such that $\sup_{x \in \mathcal{X}} |M^{(1)}(x, \widehat{\mathsf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}}) - M^{(1)}(x, \mathsf{w}; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}})| = o_{\mathbb{P}}(1)$, and $J^v \left(\frac{J \log J}{n}\right)^{1/2} = o(1)$, then

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

28

SA-3.8 Shape Restriction Tests

The third application of our results is to test certain shape restrictions on $\mu_0^{(v)}(x)$, $\vartheta_0(x, \mathsf{w})$ and $\zeta_0(x, \mathsf{w})$. To be specific, consider the following problem:

$$\ddot{\mathsf{H}}_0^{\mu^{(v)}} : \sup_{x \in \mathcal{X}} (\mu^{(v)}(x) - m^{(v)}(x; \bar{\boldsymbol{\theta}})) \leq 0 \text{ for certain } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}} \quad \text{v.s.}$$

$$\ddot{\mathsf{H}}_A^{\mu^{(v)}} : \sup_{x \in \mathcal{X}} (\mu^{(v)}(x) - m^{(v)}(x; \bar{\boldsymbol{\theta}})) > 0 \text{ for } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}}.$$

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 $\ddot{\mathsf{H}}_0^{\mu^{(v)}}$ and $\ddot{\mathsf{H}}_A^{\mu^{(v)}}$, we fix $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\gamma}}$ to be the same values in the parameter space. In such a case, we introduce $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\gamma}}$ as consistent estimators of $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\gamma}}$ under both $\ddot{\mathsf{H}}_0^{\mu^{(v)}}$ and $\ddot{\mathsf{H}}_A^{\mu^{(v)}}$. Then we will rely on the following test statistic:

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

The null hypothesis is rejected if $\sup_{x \in \mathcal{X}} \ddot{T}_{\mu^{(v)}, p}(x) > \mathfrak{c}_{\mu^{(v)}}$ for some critical value $\mathfrak{c}_{\mu^{(v)}}$.

Similarly, define the test for the shape of $\vartheta_0(x, \mathsf{w})$:

$$\ddot{\mathsf{H}}_0^{\vartheta} : \sup_{x \in \mathcal{X}} (\vartheta_0(x, \mathsf{w}) - M(x, \mathsf{w}; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}})) \leq 0 \text{ for certain } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}} \quad \text{v.s.}$$

$$\ddot{\mathsf{H}}_A^{\vartheta} : \sup_{x \in \mathcal{X}} (\vartheta_0(x, \mathsf{w}) - M(x, \mathsf{w}; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}})) > 0 \text{ for } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}}.$$

We will rely on the following test statistic:

$$\ddot{T}_{\vartheta, p}(x) := \frac{\widehat{\vartheta}(x, \widehat{\mathsf{w}}) - M(x, \widehat{\mathsf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}_{\vartheta}(x)/n}}.$$

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

Also, define the test for the shape of $\zeta_0(x, \mathsf{w})$:

$$\ddot{\mathsf{H}}_0^{\zeta} : \sup_{x \in \mathcal{X}} (\zeta_0(x, \mathsf{w}) - M^{(1)}(x, \mathsf{w}; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}})) \leq 0 \text{ for certain } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}} \quad \text{v.s.}$$

$$\ddot{\mathsf{H}}_A^{\zeta} : \sup_{x \in \mathcal{X}} (\zeta_0(x, \mathsf{w}) - M^{(1)}(x, \mathsf{w}; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}})) > 0 \text{ for } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}}.$$

We will rely on the following test statistic:

$$\ddot{T}_{\zeta,p}(x) := \frac{\widehat{\zeta}(x,\widehat{\mathsf{w}}) - M^{(1)}(x,\widehat{\mathsf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}_\zeta(x)/n}}.$$

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

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

Theorem SA-3.10 (Shape Restriction Tests)**.** *Suppose that the conditions in Theorem SA-3.8 hold.*

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

Under $\ddot{\mathsf{H}}_0^{\mu^{(v)}}$,

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} \ddot{T}_{\mu^{(v)},p}(x) > \mathfrak{c}_{\mu^{(v)}}\right] \le \alpha.$$

Under $\ddot{\mathsf{H}}_A^{\mu^{(v)}}$, *if* $J^v \left(\frac{J \log J}{n}\right)^{1/2} = o(1)$,

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in \mathcal{X}} \ddot{T}_{\mu^{(v)},p}(x) > \mathfrak{c}_{\mu^{(v)}}\right] = 1.$$

(ii) *Assume* $\sup_{x \in \mathcal{X}} |M(x,\widehat{\mathsf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}}) - M(x,\mathsf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})| = o_\mathbb{P}\left(\sqrt{\frac{J^{1+2v}}{n \log J}}\right)$. *Let* $\mathfrak{c}_\vartheta = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x \in \mathcal{X}} \widehat{Z}_{\vartheta,p}(x) \le c | \mathbf{D}] \ge 1 - \alpha\}$.

Under $\ddot{\mathsf{H}}_0^\vartheta$,

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

Under $\ddot{\mathsf{H}}_A^\vartheta$, *if* $J^v \left(\frac{J \log J}{n}\right)^{1/2} = o(1)$,

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

(iii) *Assume* $\sup_{x \in \mathcal{X}} |M^{(1)}(x,\widehat{\mathsf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}}) - M^{(1)}(x,\mathsf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})| = o_\mathbb{P}\left(\sqrt{\frac{J^{1+2v}}{n \log J}}\right)$. *Let* $\mathfrak{c}_\zeta = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x \in \mathcal{X}} \widehat{Z}_{\zeta,p}(x) \le c | \mathbf{D}] \ge 1 - \alpha\}$.

Under $\ddot{\mathsf{H}}_0^\zeta$,

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \ddot{T}_{\zeta,p}(x) > \mathfrak{c}_\zeta\Big] \leq \alpha.$$

Under $\ddot{\mathsf{H}}_A^\zeta$, *if* $J^v\left(\frac{J\log J}{n}\right)^{1/2} = o(1)$,

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

Remark SA-3.8 (Improvements over literature)**.** The results in Sections SA-3.6–SA-3.8 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 allow for a large class of random partitioning schemes, formally take into account the potential randomness of the partition, account for the generalized semi-linear structure, and consider an array of possibly nonlinear estimation and inference problems. In particular, the approach taken in Theorems SA-3.5 and SA-3.7 to establish strong approximation and related distributional approximations for nonlinear binscatter statistics may be of independent interest. The key condition imposed on the partition for uniform inference (confidence bands and hypothesis testing) is Assumption SA-RP(i), while "convergence" of the random partition (Assumption SA-RP(ii)) is not required. ⌟

SA-4 Implementation Details

SA-4.1 Standard Error Computation

In Section SA-3, we have given the variance formulas $\widehat{\Omega}_{\mu^{(v)}}(x)$, $\widehat{\Omega}_\vartheta(x)$ and $\widehat{\Omega}_\zeta(x)$ that can be used to obtain the standard errors of $\widehat{\mu}^{(v)}(x)$, $\widehat{\vartheta}(x,\widehat{\mathsf{w}})$ and $\widehat{\zeta}(x,\widehat{\mathsf{w}})$. Recall that the formula for the estimator $\widehat{\Sigma}$ of Σ_0 is

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

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 binscatter estimates $\widehat{\mu}(\cdot)$ and $\widehat{\gamma}$. Thus, $\widehat{\Sigma}$ 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 $\bar{\mathbf{Q}}$ (or $\mathbf{Q}_0$), which in general relies on an estimator $\widehat{\Psi}_1(\cdot)$ of $\Psi_1(\cdot)$ 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, $\psi(y_i, \eta_i) = -\epsilon_i$ and $\eta_{i,1} = 1$. Thus, the formula for $\widehat{\mathbf{Q}}$ given in Section SA-3 reduces to $\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)']$, 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

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

Example 4 (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, $\psi(y_i, \eta_i) = \mathbb{1}(\epsilon_i < 0) - q$, and one needs to estimate

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

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-4.2   Number of Bins Selector

We discuss the implementation details for data-driven selection of the number of bins, based on the approximate integrated mean squared error expansion in Theorem SA-3.4.

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-4.2.1   Rule-of-thumb Selector

A rule-of-thumb choice of J can be obtained based on Corollary SA-3.2 in Cattaneo et al. (2024b), which gives an explicit characterization of the variance and bias constants for least squares bin-scatter using piecewise polynomials $(s = 0)$.

Specifically, the variance constant $\mathscr{V}(p,0,v)$ is estimated by

$$\widehat{\mathscr{V}}(p,0,v) = \text{trace} \left\{ \left(\int_0^1 \boldsymbol{\varphi}(z)\boldsymbol{\varphi}(z)'dz \right)^{-1} \int_0^1 \boldsymbol{\varphi}^{(v)}(z)\boldsymbol{\varphi}^{(v)}(z)'dz \right\} \times \frac{1}{n} \sum_{i=1}^n \widehat{\sigma}^2(x_i, \mathbf{w}_i)\widehat{f}_X(x_i)^{2v}$$

where $\boldsymbol{\varphi}(z) = (1, z, \ldots, z^p)'$, $\widehat{\sigma}^2(x_i, \mathbf{w}_i)$ is some estimate of the conditional variance $\mathbb{V}[y_i|x_i, \mathbf{w}_i]$ and $\widehat{f}_X(x_i)$ is some estimate of the density $f_X(x_i)$. On the other hand, the bias constant $\mathscr{B}(p,0,v)$ is estimated by

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

where $\mathscr{B}_p(z) = (-1)^p \sum_{k=0}^p \binom{p}{k}\binom{p+k}{k}(-z)^k / \binom{2p}{p}$ for each $p \in \mathbb{Z}_+$ and $\widehat{\mu}^{(p+1)}(x_i)$ is some preliminary estimate of $\mu_0^{(p+1)}(x_i)$. The details about getting the estimates $\widehat{\sigma}^2(x_i, \mathbf{w}_i)$, $\widehat{f}_X(x_i)$ and $\widehat{\mu}^{(p+1)}(x_i)$ can be found in Section SA-4.1 in Cattaneo et al. (2024b).

This procedure still yields a choice of J with the correct rate, though the constant approximations are inconsistent for general loss.

SA-4.2.2 Direct-plug-in Selector

The direct-plug-in selector is implemented based on nonlinear 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 $\widehat{\mathbf{b}}_{p,s}(x)$ (of order p) can be constructed immediately on the preliminary partition. Implementing a nonlinear binscatter estimation using this basis and partitioning, we can obtain the variance constant estimate using the variance matrix estimators discussed in Section SA-4.1.

Regarding the bias constant, the key unknown in the expression of the leading approximation error $r_{0,v}^{\star}(x)$ in Theorem SA-3.4 is $\mu_0^{(p+1)}(x)$, which can be estimated by implementing a nonlinear binscatter estimation of order $p+1$ (with the preliminary partition unchanged). Also, an estimate of $f_X(x_i)^{-1}$ in $r_{0,v}^{\star}(x)$ is $J\hat{h}_{x_i}$ where $\hat{h}_{x_i}$ denotes the length of the interval in $\widehat{\Delta}$ containing x_i. All other quantities in the expression of $\mathscr{B}(p,s,v)$ 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 consistent constant approximations for any nonlinear binscatter with any choice of p, s and v.

SA-4.3 Fixed J and choice of polynomial order

Our main theory treats J as diverging with the sample size. This reflects the standard approach wherein a researcher selects p and s in advance (often $s = p = 0$ or $s = p = 3$) and then selects J given the data. The partition must get finer to remove the nonparametric smoothing bias in estimating the function $\mu_0(x)$ (and along with it, $\vartheta_0(x, \mathsf{w})$ or $\zeta_0(x, \mathsf{w})$). Correct recovery (either for estimation or visualization) of these functions is the primary use of binscatter. However, researchers sometimes prefer to pre-specify a fixed $J = \mathsf{J}$, and we also discuss implementation and interpretation of binscatter in this case.

Instead of modeling J as diverging and searching for the optimal choice, a researcher may desire a fixed (often small and round) number of J, which we denote by J. This is done either to make the estimate more visually appealing or because the results can be directly interpreted. In this case, instead of recovering the functions $\mu_0^{(v)}(x)$, $\vartheta_0(x, \mathsf{w})$, and $\zeta_0(x, \mathsf{w})$, the binscatter is interpreted as

estimating their coarsened versions: the distribution of y_i conditional on x_i lying in a (fixed) bin, rather than at a single point. For some J, this remains interpretable and all our inference results apply to this case. For example, in our application we can take $J = 10$ and study the distribution of uninsured rate for each decile of income. The confidence bands then become pointwise confidence intervals that are simultaneously valid. For example, this could be used to examine inequality in health care access by asking if median uninsured rates are statistically different between the top and bottom decile.

A fixed J is also interpretable, and applicable, if x_i is discrete. Then each mass point is given its own bin and the results apply to the conditional distribution of y_i given $x_i = x$. Again, our theoretical results apply directly to this case and one obtains simultaneous inference over the set of points. Cattaneo et al. (2024b) give further discussion and examples.

As a practical compromise between the visual appeal and interpretation of a small, fixed J and the demand for consistent estimation, we propose a novel, albeit ad-hoc, adjustment to the estimator aimed at addressing the smoothing bias left by fixing J by adjusting the choice of polynomial order p. The standard approach fixes p in advance and selects J based on the data, but we can invert this and search for the value of p for which the fixed J would be IMSE-optimal. That is, we solve for

$$p_{\texttt{IMSE}}(\texttt{J}, v) = \arg\min_{p \in \mathcal{P}} \left| J_{\texttt{IMSE}}(p, v) - \texttt{J} \right|, \tag{SA-4.1}$$

where in principle the set $\mathcal{P}$ is all nonnegative integers, but in practice $\mathcal{P} = \{p_{\texttt{min}}, p_{\texttt{min}} + 1, \ldots, p_{\texttt{max}} - 1, p_{\texttt{max}}\}$, for some integers $0 \leq p_{\texttt{min}} \leq p_{\texttt{max}}$. The (in)flexibility of fixed $J = \texttt{J}$ is offset by changing the polynomial approximation. This may have some practical appeal, but our theoretical results in the next section continue in the standard case of fixed p and diverging J.

To implement the data-driven choice $p_{\texttt{IMSE}}(\texttt{J}, v)$, users needs to specify the desired (often small) number of bins J, the derivative order v of interest, and a (finite) set $\mathcal{P}$ of acceptable polynomial orders. The size of $\mathcal{P}$ is usually small since in practice $p = 3$ or 4 often suffices to yield a small IMSE-optimal number of bins. Then, for each value of p in $\mathcal{P}$, we can implement the rule-of-thumb or direct plug-in procedure as described in Section SA-4.2 to obtain $J_{\texttt{IMSE}}(p, v)$. The "optimal" choice $p_{\texttt{IMSE}}(\texttt{J}, v)$ is the value of p with the resulting $J_{\texttt{IMSE}}(p, v)$ closest to J.

SA-5 Proofs

We begin with a subsection collecting some technical lemmas used in the proofs of our main results. We then collect all the proof of the results presented in this supplemental appendix, which are in several cases more general than those discussed in the main text. Some of our technical results may be of more broad independent interest in the nonlinear series estimation literature.

SA-5.1 Technical Lemmas

We first give several simple facts about $\widehat{\Delta}$ in the following lemma, which are immediate from Assumption SA-RP(ii).

Lemma SA-5.1 (Quasi-Uniformity). *Suppose that Assumption SA-RP(ii) holds. Then, (i) $J^{-1} \lesssim \min_{1 \leq j \leq J} h_j \leq \max_{1 \leq j \leq J} h_j \lesssim J^{-1}$, (ii) $\max_{1 \leq j \leq J} |\hat{\tau}_j - \tau_j| \lesssim_{\mathbb{P}} \mathfrak{r}_{\mathrm{RP}}$, and (iii) $\widehat{\Delta} \in \Pi_{3c_{\mathrm{QU}}}$ w.p.a. 1.*

Proof. By Assumption SA-RP(ii), $\mathrm{len}(\mathcal{X}) = \sum_{j=1}^J h_j \geq J \min_{1 \leq j \leq J} h_j \geq c_{\mathrm{QU}}^{-1} J \max_{1 \leq j \leq J} h_j$ where $\mathrm{len}(\mathcal{X})$ denotes the length of $\mathcal{X}$ (which is a fixed number). On the other hand, $\mathrm{len}(\mathcal{X}) \leq J \max_{1 \leq j \leq J} h_j$ $\leq c_{\mathrm{QU}} J \min_{1 \leq j \leq J} h_j$. Therefore, $c_{\mathrm{QU}}^{-1} J^{-1} \mathrm{len}(\mathcal{X}) \leq \min_{1 \leq j \leq J} h_j \leq \max_{1 \leq j \leq J} h_j \leq c_{\mathrm{QU}} J^{-1} \mathrm{len}(\mathcal{X})$.

Next, by Assumption SA-RP(ii), $\max_{1 \leq j \leq J} |\hat{\tau}_j - \tau_j| = \max_{1 \leq j \leq J} |\sum_{l=1}^j (\hat{h}_l - h_l)| \leq J \max_{1 \leq l \leq J} |\hat{h}_l - h_l| \lesssim \mathfrak{r}_{\mathrm{RP}}$. In addition, $\max_{1 \leq j \leq J} |\hat{h}_j - h_j| \leq \frac{1}{2} c_{\mathrm{QU}}^{-1} J^{-1} \mathrm{len}(\mathcal{X}) \leq \frac{1}{2} \min_{1 \leq j \leq J} h_j$ w.p.a. 1, and thus

$$\frac{\max_{1 \leq j \leq J} \hat{h}_j}{\min_{1 \leq j \leq J} \hat{h}_j} = \frac{\max_{1 \leq j \leq J} h_j + \max_{1 \leq j \leq J} |\hat{h}_j - h_j|}{\min_{1 \leq j \leq J} h_j - \max_{1 \leq j \leq J} |\hat{h}_j - h_j|} \leq 3c_{\mathrm{QU}}, \quad \text{w.p.a.1.}$$

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

The next lemma then verifies Assumption SA-RP(ii) for the special case of quantile-spaced partitions. The proof is available in the supplemental appendix of Cattaneo et al. (2024b) (see Section SA-3.1 therein) and thus omitted here.

Lemma SA-5.2 (Quasi-Uniformity of Quantile-Spaced Partitions). *Suppose that Assumption SA-DGP(i) and SA-DGP(ii) holds and $\widehat{\Delta}$ is generated by sample quantiles, i.e., $\hat{\tau}_j = \hat{F}_X^{-1}(j/J)$. If $\frac{J \log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then Assumption SA-RP(ii) holds with $\tau_j = F_X^{-1}(j/J)$ and $\mathfrak{r}_{\mathrm{RP}} = \left(\frac{J \log J}{n}\right)^{1/2}$.*

The next three lemmas SA-5.3–SA-5.5 concern the properties of binscatter basis functions. Their proofs are the same as those for quantile-based partitions that are available in the supplemental appendix of Cattaneo et al. (2024b) (see Section SA-3.1 therein) and are omitted here to conserve space.

Lemma SA-5.3 (Transformation Matrix). *Suppose that Assumption SA-RP(i) holds. Then $\widehat{\mathbf{b}}_{p,s}(x) = \widehat{\mathbf{T}}_s \widehat{\mathbf{b}}_{p,0}(x)$ with $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\| \lesssim_{\mathbb{P}} 1$. If, in addition, Assumption SA-RP(ii) holds, then $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\|_\infty \lesssim_{\mathbb{P}} \mathfrak{r}_{\mathsf{RP}}$ and $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \lesssim_{\mathbb{P}} \mathfrak{r}_{\mathsf{RP}}$.*

Lemma SA-5.4 (Local Basis). *Suppose that Assumption SA-RP(i) holds. Then $\sup_{x \in \mathcal{X}} \|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)\|_0 \leq (p+1)^2$ and $\sup_{x \in \mathcal{X}} \|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)\| \lesssim_{\mathbb{P}} J^{\frac{1}{2}+v}$.*

The following lemma provides a particular way to define $\boldsymbol{\beta}_0(\Delta)$ and $\widehat{\boldsymbol{\beta}}_0$ so that the required approximation rate is achieved. We define

$$\boldsymbol{\beta}_0^{\mathsf{LS}}(\Delta) := \underset{\boldsymbol{\beta} \in \mathbb{R}^{K_{p,s}}}{\arg\min} \; \mathbb{E}[(\mu_0(x_i) - \mathbf{b}_{p,s}(x_i; \Delta)'\boldsymbol{\beta})^2], \quad \widehat{\boldsymbol{\beta}}_0^{\mathsf{LS}} = \boldsymbol{\beta}_0^{\mathsf{LS}}(\widehat{\Delta}).$$

Lemma SA-5.5 (Approximation Error). *Suppose that Assumptions SA-DGP(i)(ii), SA-SM(v) and SA-RP(i) hold. Then*

$$\sup_{\Delta \in \Pi} \sup_{x \in \mathcal{X}} |\mathbf{b}_{p,s}^{(v)}(x; \Delta)'\boldsymbol{\beta}_0^{\mathsf{LS}}(\Delta) - \mu_0^{(v)}(x)| \lesssim J^{-p-1+v} \quad and \quad \sup_{x \in \mathcal{X}} |\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0^{\mathsf{LS}} - \mu_0^{(v)}(x)| \lesssim_{\mathbb{P}} J^{-p-1+v}.$$

Next, the following maximal inequality is useful in our analysis. Its proof is available in Cattaneo et al. (2024c) and thus omitted here.

Lemma SA-5.6 (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 joint distribution of $Z_1, \cdots, Z_n$ by $\mathbb{P}$ and the marginal 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

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

SA-5.2 Proof of Lemma SA-3.1

Proof. We write $\Psi_{i,1} := \Psi_1(x_i, \mathbf{w}_i; \eta_i)$.

(i) We first prove a convergence result for $\bar{\mathbf{Q}}$. In view of Lemma SA-5.3, it suffices to show the convergence for $s = 0$. Let $\mathcal{A}_n$ denote the event on which $\widehat{\Delta} \in \Pi$. By Assumption SA-RP(i), $\mathbb{P}(\mathcal{A}_n^c) = o(1)$. On $\mathcal{A}_n$,

$$\left\| \mathbb{E}_n[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)'\Psi_{i,1}\eta_{i,1}^2] - \mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)'\Psi_{i,1}\eta_{i,1}^2] \right\|$$

$$\leq \sup_{\Delta \in \Pi} \| \mathbb{E}_n[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)'\Psi_{i,1}\eta_i^2] - \mathbb{E}[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)'\Psi_{i,1}\eta_i^2] \|_\infty.$$

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

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

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

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

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-SM. Also, by Proposition 3.6.12 of Giné and Nickl (2016), the collection $\mathcal{G}$ is of VC type with a bounded index. Then, by Lemma SA-5.6,

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

which implies $\|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)'\Psi_{i,1}\eta_{i,1}^2] - \mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)'\Psi_{i,1}\eta_{i,1}^2]\| \lesssim_{\mathbb{P}} \sqrt{J \log J/n}$.

Then, the lower bound on the minimum eigenvalue of $\bar{\mathbf{Q}}$ follows by Theorem 4.42 of Schumaker (2007) and Assumption SA-RP(i). The upper bound immediately follows by Assumption SA-RP(i) and Lemmas SA-5.3 and SA-5.4.

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

(ii) By Assumption SA-DGP and SA-SM, $\Psi_{i,1}\eta_{i,1}^2$ is bounded and bounded away from zero uniformly over $1 \le i \le n$. Then, $\mathbb{E}[\mathbf{b}_{p,s}(x_i)\mathbf{b}_{p,s}(x_i)'] \lesssim \mathbf{Q}_0 \lesssim \mathbb{E}[\mathbf{b}_{p,s}(x_i)\mathbf{b}_{p,s}(x_i)']$. The desired bounds on the minimum and maximum eigenvalues of $\mathbf{Q}_0$ follow from Lemma SA-3.5 of Cattaneo et al. (2024b).

Next, we show the convergence of $\bar{\mathbf{Q}}$ to $\mathbf{Q}_0$. Let α_{kl} be a generic (k, l)th entry of

$$\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)'\Psi_{i,1}\eta_{i,1}^2]/J - \mathbb{E}[\mathbf{b}_{p,0}(x_i)\mathbf{b}_{p,0}(x_i)'\Psi_{i,1}\eta_{i,1}^2]/J.$$

By definition, it is either equal to zero or

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

for some $1 \le j \le J$ and $0 \le \ell \le 2p$ and $\varphi(x_i) = \mathbb{E}[\varkappa(x_i, \mathbf{w}_i)|x_i]$. By Assumptions SA-DGP and SA-SM and the argument in the proof of Lemma SA-3.5 of Cattaneo et al. (2024b),

$$\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)'\Psi_{i,1}\eta_{i,1}^2] - \mathbf{Q}_0\| \lesssim_{\mathbb{P}} \mathfrak{r}_{\mathsf{RP}}.$$

Since $\bar{\mathbf{Q}}$ and $\mathbf{Q}_0$ are banded matrices with finite band widths. Then, the bound $\|\bar{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\|_\infty$ hold by Theorem 2.2 of Demko (1977). This completes the proof. $\square$

SA-5.3 Proof of Lemma SA-3.2

Proof. Since $\mathbb{E}[\psi(y_i, \eta_i)^2|x_i = x, \mathbf{w}_i = \mathbf{w}]$ 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 $\mathbf{w} \in \mathcal{W}$, $\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'] \lesssim \bar{\boldsymbol{\Sigma}} \lesssim \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)']$. By

the same argument in the proof of Lemma SA-3.1 (we can simply drop the additional term $\Psi_{i,1}\eta_{i,1}^2$ in $\bar{\mathbf{Q}}$), the eigenvalues of $\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)']$ and thus $\bar{\boldsymbol{\Sigma}}$ are bounded and bounded away from zero. Then, the desired results follow from Lemma SA-3.1 and the fact that $\inf_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)\| \gtrsim J^{1/2+v}$ w.p.a. 1 (it was shown in the proof of Lemma SA-3.6 of Cattaneo et al. (2024b)). $\square$

SA-5.4 Proof of Lemma SA-3.3

Proof. By Lemmas SA-5.3, SA-5.4 and SA-3.1, $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)\|_1 \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\bar{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$. Recall that by Assumption SA-SM, $\psi(y_i,\eta_i) = \psi^\dagger(y_i-\eta_i)\psi^\ddagger(\eta_i) = \psi^\dagger(\epsilon_i)\psi^\ddagger(\eta_i)$. Define the following function class

$$\mathcal{G} = \left\{(x_1,\mathbf{w}_1,\epsilon_1) \mapsto b_{p,0,l}(x_1;\Delta)\eta^{(1)}(\mu_0(x_1)+\mathbf{w}_1'\boldsymbol{\gamma}_0)\psi^\dagger(\epsilon_1)\psi^\ddagger(\eta_1) : 1 \le l \le J(p+1), \Delta \in \Pi\right\}.$$

Then, $\sup_{g\in\mathcal{G}}|g| \lesssim \sqrt{J}|\psi^\dagger(\epsilon_1)|$, and hence take an envelop $\bar{G} = C\sqrt{J}|\psi^\dagger(\epsilon_1)|$ for some C large enough. Moreover, $\sup_{g\in\mathcal{G}}\mathbb{V}[g] \lesssim 1$ and $\mathcal{G}$ is of VC type with a bounded index. By Proposition 6.1 of Belloni et al. (2015),

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

and the desired result follows. $\square$

SA-5.5 Proof of Lemma SA-3.4

Proof. Let $\hat{z}_i = \widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0$ and $\mathfrak{r}(x_i,\mathbf{w}_i,y_i) := \mathfrak{r}(x_i,\mathbf{w}_i,y_i;\widehat{\Delta}) := \eta_{i,1}\psi(y_i,\eta_i)-\eta^{(1)}(\hat{z}_i)\psi(y_i,\eta(\hat{z}_i))$ $= A_1(x_i,\mathbf{w}_i,y_i) + A_2(x_i,\mathbf{w}_i,y_i)$ where

$$A_1(x_i,\mathbf{w}_i,y_i) := A_1(x_i,\mathbf{w}_i,y_i;\widehat{\Delta}) := [\eta_{i,1}\psi^\ddagger(\eta_i) - \eta^{(1)}(\hat{z}_i)\psi^\ddagger(\eta(\hat{z}_i))]\psi^\dagger(y_i,\eta_i) \text{ and}$$

$$A_2(x_i,\mathbf{w}_i,y_i) := A_2(x_i,\mathbf{w}_i,y_i;\widehat{\Delta}) := \eta^{(1)}(\hat{z}_i)\psi^\ddagger(\eta(\hat{z}_i))[\psi^\dagger(y_i,\eta_i) - \psi^\dagger(y_i,\eta(\hat{z}_i))].$$

First, by Assumption SA-SM and Lemma SA-5.5, $\sup_{x\in\mathcal{X},\mathbf{w}\in\mathcal{W}}|\eta_{i,1}\psi^\ddagger(\eta_i) - \eta^{(1)}(\hat{z}_i)\psi^\ddagger(\eta(\hat{z}_i))| \lesssim$

J^{-p-1} w.p.a. 1. Also, for every $1 \le l \le K_{p,s}$ and $\Delta \in \Pi$,

$$b_{p,s,l}(x;\Delta)\Big(\eta_{i,1}\psi^{\ddagger}(\eta_i) - \eta^{(1)}(\mathbf{b}_{p,s}(x;\Delta)'\boldsymbol{\beta}_0(\Delta) + \mathbf{w}'\boldsymbol{\gamma}_0)\psi^{\ddagger}(\mathbf{b}_{p,s}(x;\Delta)'\boldsymbol{\beta}_0(\Delta) + \mathbf{w}'\boldsymbol{\gamma}_0)\Big)$$

$$= b_{p,s,l}(x;\Delta)\eta_{i,1}\psi^{\ddagger}(\eta_i) -$$

$$b_{p,s,l}(x;\Delta)\eta^{(1)}\bigg(\sum_{k=\underline{k}_l}^{\underline{k}_l+p} b_{p,s,k}(x;\Delta)\beta_{0,k}(\Delta) + \mathbf{w}'\boldsymbol{\gamma}_0\bigg)\psi^{\ddagger}\bigg(\sum_{k=\underline{k}_l}^{\underline{k}_l+p} b_{p,s,k}(x;\Delta)\beta_{0,k}(\Delta) + \mathbf{w}'\boldsymbol{\gamma}_0\bigg)$$

for some integer $\underline{k}_l \in [1, K_{p,s}]$ where $\beta_{0,k}(\Delta)$ denotes the kth element in $\boldsymbol{\beta}_0(\Delta)$. Then, the function class $\mathcal{G} = \{(x, \mathbf{w}, y) \mapsto b_{p,s,l}(x;\Delta)A_1(x,\mathbf{w},y;\Delta) : 1 \le l \le K_{p,s}, \Delta \in \Pi\}$ is of VC type with a bounded index. By the same argument given in the proof of Lemma SA-3.3,

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

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

$$\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)A_2(x_i,\mathbf{w}_i,y_i)] = \mathbb{E}_n[\mathbb{E}[\widehat{\mathbf{b}}_{p,s}(x_i)A_2(x_i,\mathbf{w}_i,y_i)|\mathscr{F}_{XW\Delta}]] +$$

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

By Assumption SA-SM(iii) and Lemma SA-5.5,

$$\max_{1 \le i \le n} |\mathbb{E}[A_2(x_i,\mathbf{w}_i,y_i)|\mathscr{F}_{XW\Delta}]|$$

$$= \max_{1 \le i \le n} |\eta^{(1)}(\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)\Psi(x_i,\mathbf{w}_i;\eta(\hat{z}_i))| \lesssim_{\mathbb{P}} J^{-p-1}.$$

Then, $\|\mathbb{E}_n[\mathbb{E}[\widehat{\mathbf{b}}_{p,s}(x_i)A_2(x_i,\mathbf{w}_i,y_i)|\mathscr{F}_{XW\Delta}]]\|_\infty \lesssim_{\mathbb{P}} J^{-p-1-1/2}$ by the same argument in the proof of Lemma SA-3.1. On the other hand, define the following function class

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

By Assumption SA-SM, $\sup_{g \in \mathcal{G}} \|g\|_\infty \lesssim J^{1/2}$, and $\sup_{g \in \mathcal{G}} \mathbb{V}[g(x_i,\mathbf{w}_i,y_i)] \lesssim J^{-p-1}$. By a similar argument given before, this function class is of VC type with a bounded index. Then, as in the

proof of Lemma SA-3.3, by Proposition 6.1 of Belloni et al. (2019),

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

Collecting these results, we conclude that

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

The proof is complete. □

SA-5.6 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 $\boldsymbol{\alpha} \in \mathbb{R}^{K_{p,s}}$ such that $\|\boldsymbol{\alpha}\| = 1$. For notational simplicity, let $\widehat{\mathbf{b}}_i := \widehat{\mathbf{b}}_{p,s}(x_i)$. For this choice of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} \in \mathbb{R}^d$,

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

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

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

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

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

where $\xi_{i,t}$ is between $\eta(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}} + t)$ and $\eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0)$ and we use the fact that $\Psi(x, \mathbf{w}_i; \eta_i) = 0$. By Lemma SA-5.5, the fact that $\eta(\cdot)$ is strictly monotonic and $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$ and the rate condition imposed, we have $\mathbb{E}_n[\mathbb{E}[\delta_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}) | \mathscr{F}_{XW\Delta}]] \gtrsim_{\mathbb{P}} \varepsilon^2 \boldsymbol{\alpha}' \mathbb{E}_n[\widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i'] \boldsymbol{\alpha}/J \gtrsim_{\mathbb{P}} J^{-1} \varepsilon^2$.

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

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

Note that

$$\delta_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \int_0^{\varepsilon \widehat{\mathbf{b}}_i' \boldsymbol{\alpha}/\sqrt{J}} \Big(\psi(y_i, \eta(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma} + t)) - \psi(y_i, \eta_i) \Big) \eta^{(1)}(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma} + t) dt +$$
$$\int_0^{\varepsilon \widehat{\mathbf{b}}_i' \boldsymbol{\alpha}/\sqrt{J}} \psi(y_i, \eta_i) \eta^{(1)}(\widehat{\mathbf{b}}_i'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma} + t) dt.$$

By Assumption SA-SM, we have $\sup_{g \in \mathcal{G}} |g| \lesssim \varepsilon(1 + |\psi(y_i, \eta_i)|)$, $\| \max_{1 \leq i \leq n} |\psi(y_i, \eta_i)| \|_{L_2(\mathbb{P})} \lesssim n^{1/\nu}$, $\sup_{g \in \mathcal{G}} \mathbb{E}_n[\mathbb{E}[g^2 | \mathscr{F}_{XW\Delta}]] \lesssim_{\mathbb{P}} J^{-1}\varepsilon^2$, and the VC-index of $\mathcal{G}$ is bounded by $C'K_{p,s}$ for an absolute constant $C' > 0$. Therefore, by Lemma SA-5.6 and the rate restriction,

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

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. $\qquad \square$

SA-5.7 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

$$\breve{\mathbf{b}}_i := \widehat{\mathbf{b}}_{p,s}(x_i)/\sqrt{J} = (\widehat{b}_{p,s,1}(x_i), \cdots, \widehat{b}_{p,s,K_{p,s}}(x_i))'/\sqrt{J} \quad \text{and} \quad \breve{\boldsymbol{\beta}}_0 = \sqrt{J}\widehat{\boldsymbol{\beta}}_0.$$

Throughout the proof, $C, c, C_1, c_1, C_2, c_2, \cdots$ denote (strictly positive) absolute constants, $\mathscr{F}_{XW\Delta}$ denotes the σ-field generated by $\{(x_i, \mathbf{w}_i)\}_{i=1}^n$ and $\widehat{\Delta}$, and $\text{supp}(g(\cdot))$ denotes the support of a generic function $g(\cdot)$. Moreover, define

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

43

where $\varrho \in (0,1)$ is the constant given in Lemma SA-3.1, $r_{1,n} = C_1[(J\log n/n)^{1/2} + J^{-p-1}]$, $r_{2,n} = \mathfrak{z}\mathfrak{r}_{2,n}$ for $\mathfrak{z} > 0$, $\varepsilon_n = \mathfrak{z}'\mathfrak{r}_{2,n}$ for $\mathfrak{z}' > 0$, $\mathfrak{r}_{2,n} = [(\frac{J\log n}{n})^{3/4}\log n + J^{-\frac{p+1}{2}}\sqrt{\frac{J}{n}}\log n + J^{-2p-2} + \mathfrak{r}_\gamma]$, $r_{3,n} = C\mathfrak{r}_\gamma$, and $M_n = c_1 \log n$. In the last step of the proof, we will consider $\mathfrak{z} = 2^\ell$, $\ell = L, L+1, \cdots, \bar{L}$ where $\bar{L}$ is the smallest number such that $2^{\bar{L}}r_{2n} \geq c$ for some sufficiently small constant $c > 0$, and ε_n is a quantity that we can choose. By Assumption SA-HLE, $\hat{\gamma} - \gamma_0 \in \mathcal{H}_3$ with probability approaching one for C large enough, and by Lemma SA-3.5, $\sqrt{J}\hat{\beta} - \breve{\beta}_0 \leq c$ with probability approaching one.

For any $\beta_1 \in \mathcal{H}_1, \beta_2 \in \mathcal{H}_2, \boldsymbol{v} \in \mathcal{V}$ and $\gamma := \gamma_0 + \gamma_1$ with $\gamma_1 \in \mathcal{H}_3$, define

$$
\begin{aligned}
\delta_i(\beta_1, \beta_2, \boldsymbol{v}, \gamma) &= \rho\Big(y_i; \eta(\breve{\mathbf{b}}_i'(\breve{\beta}_0 + \beta_1 + \beta_2) + \mathbf{w}_i'\gamma)\Big) - \rho\Big(y_i; \eta(\breve{\mathbf{b}}_i'(\breve{\beta}_0 + \beta_1 + \beta_2 - \boldsymbol{v}) + \mathbf{w}_i'\gamma)\Big) \\
&\quad - \Big[\eta(\breve{\mathbf{b}}_i'(\breve{\beta}_0 + \beta_1 + \beta_2) + \mathbf{w}_i'\gamma) - \eta(\breve{\mathbf{b}}_i'(\breve{\beta}_0 + \beta_1 + \beta_2 - \boldsymbol{v}) + \mathbf{w}_i'\gamma)\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \times \psi(y_i, \eta(\breve{\mathbf{b}}_i'\breve{\beta}_0 + \mathbf{w}_i'\gamma_0)) \\
&= \int_{-\breve{\mathbf{b}}_i'\boldsymbol{v}}^{0} \Big[\psi\Big(y_i, \eta(\breve{\mathbf{b}}_i'(\breve{\beta}_0 + \beta_1 + \beta_2) + \mathbf{w}_i'\gamma + t)\Big) - \psi\Big(y_i, \eta(\breve{\mathbf{b}}_i'\breve{\beta}_0 + \mathbf{w}_i'\gamma_0)\Big)\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad \times \eta^{(1)}\Big(\breve{\mathbf{b}}_i'(\breve{\beta}_0 + \beta_1 + \beta_2) + \mathbf{w}_i'\gamma + t\Big)dt.
\end{aligned}
$$

Note that $\delta_i(\beta_1, \beta_2, \boldsymbol{v}, \gamma) \neq 0$ only if $\breve{\mathbf{b}}_i'\boldsymbol{v} \neq 0$. For each $\boldsymbol{v} \in \mathcal{V}$, let $\mathcal{J}_{\boldsymbol{v}} = \{j : v_j \neq 0\}$. By construction, the cardinality of $\mathcal{J}_{\boldsymbol{v}}$ is bounded by $2M_n + 1$. We have $\delta_i(\beta_1, \beta_2, \boldsymbol{v}, \gamma) \neq 0$ only if $\breve{b}_j(x_i) \neq 0$ for some $j \in \mathcal{J}_{\boldsymbol{v}}$, which happens only when $x_i \in \text{supp}(\breve{b}_j(\cdot))$ for some $j \in \mathcal{J}_{\boldsymbol{v}}$. Let $\mathcal{I}_{\boldsymbol{v}} = \cup_{j \in \mathcal{J}_{\boldsymbol{v}}} \text{supp}(\breve{b}_j(\cdot))$. Since the basis functions are locally supported, $\mathcal{I}_{\boldsymbol{v}}$ includes at most $c_2 M_n$ (connected) intervals for all $\boldsymbol{v} \in \mathcal{V}$. Moreover, at most $c_3 M_n$ basis functions in $\breve{\mathbf{b}}(\cdot)$ have supports overlapping with $\mathcal{I}_{\boldsymbol{v}}$. Denote the set of indices for such basis functions by $\bar{\mathcal{J}}_{\boldsymbol{v}}$. Let $\breve{\beta}_{0,j}$, $\beta_{1,j}$ and $\beta_{2,j}$ be the jth entries of $\breve{\beta}_0$, β_1, and β_2 respectively, and v_j be the jth entry of $\boldsymbol{v}$. Based on the above observations, we have $\delta_i(\beta_1, \beta_2, \boldsymbol{v}, \gamma) \equiv \delta_i(\beta_{1,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \beta_{2,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \boldsymbol{v}, \gamma)$ where

$$
\begin{aligned}
\delta_i(\beta_{1,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \beta_{2,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \boldsymbol{v}, \gamma) &:= \int_{-\sum_{j \in \mathcal{J}_{\boldsymbol{v}}} \breve{b}_{i,j}v_j}^{0} \Big[\psi\Big(y_i, \eta\Big(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}_{i,l}(\breve{\beta}_{0,l} + \beta_{1,l} + \beta_{2,l}) + \mathbf{w}_i'\gamma + t\Big)\Big) \\
&\quad - \psi\Big(y_i, \eta\Big(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}_{i,l}\breve{\beta}_{0,l} + \mathbf{w}_i'\gamma_0\Big)\Big)\Big] \times \eta^{(1)}\Big(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}_{i,l}(\breve{\beta}_{0,l} + \beta_{1,l} + \beta_{2,l}) + \mathbf{w}_i'\gamma + t\Big)dt \mathbb{1}_{i,\boldsymbol{v}},
\end{aligned}
$$

$\mathbb{1}_{i,\boldsymbol{v}} = \mathbb{1}(x_i \in \mathcal{I}_{\boldsymbol{v}})$, and $\beta_{1,\bar{\mathcal{J}}_{\boldsymbol{v}}}$ and $\beta_{2,\bar{\mathcal{J}}_{\boldsymbol{v}}}$ respectively denote the subvectors of β_1 and β_2 whose

indices belong to $\bar{\mathcal{J}}_{\boldsymbol{v}}$. Accordingly, define the following function class

$$\mathcal{G} = \Big\{ (x_i, \mathbf{w}_i, y_i) \mapsto \delta_i(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2, \boldsymbol{v}, \boldsymbol{\gamma}) : \boldsymbol{v} \in \mathcal{V}, \widetilde{\boldsymbol{\beta}}_1 \in \mathbb{R}^{c_3 M_n}, \widetilde{\boldsymbol{\beta}}_2 \in \mathbb{R}^{c_3 M_n},$$
$$\|\widetilde{\boldsymbol{\beta}}_1\|_\infty \le r_{1,n}, \|\widetilde{\boldsymbol{\beta}}_2\|_\infty \le r_{2,n}, \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 \in \mathcal{H}_3 \Big\}.$$

Step 1: We bound $\sup_{g \in \mathcal{G}} |\mathbb{E}_n[g(x_i, \mathbf{w}_i, y_i)] - \mathbb{E}[g(x_i, \mathbf{w}_i, y_i)|\mathscr{F}_{XW\Delta}]|$ in this step. Let $a_i(t) :=$ $\eta(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}'_{i,l} \breve{\beta}_{0,l} + \mathbf{w}'_i \boldsymbol{\gamma}_0 + t)$. Define

$$\underline{a}_i = \min \Big\{ a_i(0), a_i\Big(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}_{i,l}(\beta_{1,l} + \beta_{2,l}) + \mathbf{w}'_i \boldsymbol{\gamma}_1 \Big), a_i\Big(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}_{i,l}(\beta_{1,l} + \beta_{2,l}) + \mathbf{w}'_i \boldsymbol{\gamma}_1 + \sum_{j \in \mathcal{J}_{\boldsymbol{v}}} \breve{b}_{i,j} v_j \Big) \Big\} \text{ and}$$
$$\bar{a}_i = \max \Big\{ a_i(0), a_i\Big(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}_{i,l}(\beta_{1,l} + \beta_{2,l}) + \mathbf{w}'_i \boldsymbol{\gamma}_1 \Big), a_i\Big(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} \breve{b}_{i,l}(\beta_{1,l} + \beta_{2,l}) + \mathbf{w}'_i \boldsymbol{\gamma}_1 + \sum_{j \in \mathcal{J}_{\boldsymbol{v}}} \breve{b}_{i,j} v_j \Big) \Big\}.$$

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-SM, for all t in the interval of integration $[-\sum_{j \in \mathcal{J}_{\boldsymbol{v}}} \breve{b}_{i,j} v_j, 0]$ (or $[0, -\sum_{j \in \mathcal{J}_{\boldsymbol{v}}} \breve{b}_{i,j} v_j]$),

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

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

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

for any $(x_i, \mathbf{w}_i, y_i)$, and in this case $y_i \in (\jmath + \underline{a}_i, \jmath + \bar{a}_i)$. By Assumption SA-SM,

$$|\bar{a}_i - \underline{a}_i| \lesssim (r_{1,n} + r_{2,n} + r_{3,n} + \varepsilon_n)(|\eta_{i,1}| + r_{1,n} + r_{2,n} + r_{3,n} + \varepsilon_n).$$

By construction, for each $\boldsymbol{v} \in \mathcal{V}$, there exists some $k_{\boldsymbol{v}}$ such that $|v_\ell| \le \varrho^{|\ell - k_{\boldsymbol{v}}|} \varepsilon_n$ for $|\ell - k_{\boldsymbol{v}}| \le M_n$. Therefore, we can further write $\mathbb{1}_{i,\boldsymbol{v}} = \sum_{j: \widehat{\mathcal{B}}_j \subset \mathcal{I}_{\boldsymbol{v}}} \mathbb{1}_{i,\boldsymbol{v},j}$ where each $\mathbb{1}_{i,\boldsymbol{v},j}$ is an indicator of the

subinterval involved in $\mathcal{I}_{\boldsymbol{v}}$, and the above facts imply that for any $x_i \in \widehat{\mathcal{B}}_l$ for some $\widehat{\mathcal{B}}_l \subset \mathcal{I}_{\boldsymbol{v}}$,

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

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

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

This inequality holds for any event in $\mathscr{F}_{XW\Delta}$. 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-3.1, $\mathbb{P}(\mathcal{A}_1^c) \to 0$. On $\mathcal{A}_1$,

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

On the other hand,

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

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

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

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,

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

where the supremum is taken over all finite discrete probability measures $\mathbb{Q}$. Then, by Lemma SA-5.6,

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

where $\bar{\bar{G}} = \max_{1 \leq i \leq n} \bar{G}(x_i, \mathbf{w}_i, y_i)$. Note that $(\mathbb{E}[\bar{\bar{G}}^2])^{1/2} \lesssim \varepsilon_n(1 + n^{1/\nu}(r_{1,n} + r_{2,n} + r_{3,n} + \varepsilon_n))$.

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

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

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

Step 2: For $\widetilde{\mathbf{Q}} := \mathbb{E}_n[\breve{\mathbf{b}}_i \breve{\mathbf{b}}_i' \Psi_1(x_i, \mathbf{w}_i; \eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0))(\eta^{(1)}(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0))^2]$, by Assumption SA-SM 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, $\|\bar{\boldsymbol{\beta}}\|_\infty \leq r_{1,n}$ with probability approaching one for C_1 large enough, where

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

Step 3: By Taylor expansion, we have

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

$$= \mathbb{E}_n\Bigg[\int_{-\breve{\mathbf{b}}'_i \boldsymbol{v}}^0 \Big\{ \Psi_1(x_i, \mathbf{w}_i; \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0)) \Big(\eta^{(1)}(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0)(\breve{\mathbf{b}}'_i(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma}_1 + t)$$

$$+ \frac{1}{2}\eta^{(2)}(\xi_{i,t})(\breve{\mathbf{b}}'_i(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma}_1 + t)^2 \Big)$$

$$+ \frac{1}{2}\Psi_2(x_i, \mathbf{w}_i; \tilde{\xi}_{i,t}) \Big(\eta(\breve{\mathbf{b}}'_i(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma} + t) - \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0) \Big)^2 \Big\}$$

$$\times \Big(\eta^{(1)}(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0) + \eta^{(2)}(\check{\xi}_{i,t})(\breve{\mathbf{b}}'_i(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma}_1 + t) \Big) dt \Bigg]$$

$$= \boldsymbol{v}' \widetilde{\mathbf{Q}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \boldsymbol{v}' \mathbb{E}_n[\mathbf{b}_i \widetilde{\varkappa}_i \mathbf{w}'_i] \boldsymbol{\gamma}_1 - \frac{1}{2}\boldsymbol{v}\widetilde{\mathbf{Q}}\boldsymbol{v} + \mathrm{I} + \mathrm{II} + \mathrm{III},$$

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

$$\mathrm{I} = \mathbb{E}_n\Bigg[\int_{-\breve{\mathbf{b}}'_i \boldsymbol{v}}^0 \Psi_1(x_i; \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0))\eta^{(1)}(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0)$$

$$\times \eta^{(2)}(\check{\xi}_{i,t})(\breve{\mathbf{b}}'_i(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma}_1 + t)^2 dt \mathbb{1}_{i,\boldsymbol{v}} \Bigg] \lesssim \varepsilon_n J^{-1}(r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n})^2,$$

$$\mathrm{II} = \mathbb{E}_n\Bigg[\int_{-\breve{\mathbf{b}}'_i \boldsymbol{v}}^0 \Psi_1(x_i; \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0)) \times \frac{1}{2}\eta^{(2)}(\xi_{i,t})(\breve{\mathbf{b}}'_i(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma}_1 + t)^2$$

$$\times \eta^{(1)}\Big(\breve{\mathbf{b}}'_i(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma} + t\Big) dt \mathbb{1}_{i,\boldsymbol{v}} \Bigg] \lesssim \varepsilon_n J^{-1}(r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n})^2,$$

$$\mathrm{III} = \mathbb{E}_n\Bigg[\int_{-\breve{\mathbf{b}}'_i \boldsymbol{v}}^0 \frac{1}{2}\Psi_2(\tilde{\xi}_{i,t})\Big(\eta(\breve{\mathbf{b}}'_i(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma} + t) - \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0)\Big)^2$$

$$\times \eta^{(1)}\Big(\breve{\mathbf{b}}'_i(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma} + t\Big) dt \mathbb{1}_{i,\boldsymbol{v}} \Bigg] \lesssim \varepsilon_n J^{-1}(r_{1,n} + r_{2,n} + \varepsilon_n + r_{3,n})^2.$$

These bounds hold uniformly for $\boldsymbol{v} \in \mathcal{V}$, $\boldsymbol{\beta}_1 \in \mathcal{H}_1$, $\boldsymbol{\beta}_2 \in \mathcal{H}_2$ 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 $\mathcal{A}_2 = \{\lambda_{\max}(\widetilde{\mathbf{Q}}) \leq c_4 J^{-1}\}$ 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-SM and Taylor's expansion,

$$\mathrm{IV} = \mathbb{E}_n\Bigg[\Big(\eta(\breve{\mathbf{b}}'_i(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \mathbf{w}'_i \boldsymbol{\gamma}) - \eta(\breve{\mathbf{b}}'_i(\breve{\boldsymbol{\beta}}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}) + \mathbf{w}'_i \boldsymbol{\gamma}) \Big) \psi(y_i, \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0)) \Bigg]$$

$$- \mathbb{E}_n\Big[\boldsymbol{v}' \breve{\mathbf{b}}_i \psi(y_i, \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0))\eta^{(1)}(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0) \Big]$$

$$= \mathbb{E}_n\Big[\boldsymbol{v}' \breve{\mathbf{b}}_i \psi(y_i, \eta(\breve{\mathbf{b}}'_i \breve{\boldsymbol{\beta}}_0 + \mathbf{w}'_i \boldsymbol{\gamma}_0))\Big(\eta^{(2)}(\xi_i)(\breve{\mathbf{b}}'_i(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}) + \mathbf{w}'_i \boldsymbol{\gamma}_1) + \frac{1}{2}\eta^{(2)}(\tilde{\xi}_i)\boldsymbol{v}' \breve{\mathbf{b}}_i \Big) \Big]$$

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

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

$$\mathcal{A}_3=\Bigg\{\sup\ \Bigg(\Big\|\mathbb{E}_n\Big[\breve{\mathbf{b}}_i\breve{\mathbf{b}}_i'\psi(y_i,\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\eta^{(2)}(\varpi_i)\Big]\Big\|_\infty+$$
$$\Big\|\mathbb{E}_n\Big[\breve{\mathbf{b}}_i\psi(y_i,\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\boldsymbol{\gamma}_0))\eta^{(2)}(\varpi_i)\mathbf{w}_i\Big]\Big\|_\infty\Bigg)\lesssim J^{-1}\Big(\Big(\frac{J\log n}{n}\Big)^{1/2}+J^{-p-1}\Big)\Bigg\},$$

where the supremum is taken over $\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V},\boldsymbol{\gamma}_1\in\mathcal{H}_3$ and ϖ_i within the range of ξ_i or $\tilde{\xi}_i$. Note that $\mathbb{E}[\psi(y_i,\eta_i)|\mathscr{F}_{XW\Delta}]=0$ 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 to obtain $\mathbb{P}(\mathcal{A}_3)\to 1$ by choosing $C_3>0$ sufficiently large.

Step 5: Let $\bar{\boldsymbol{v}}=c_5\varepsilon_n J^{-1}[\bar{\mathbf{Q}}^{-1}]_k$ for some k such that $|\beta_{2,k}|=\|\boldsymbol{\beta}_2\|_\infty$ for some $c_5>0$ where $[\bar{\mathbf{Q}}^{-1}]_k$ denotes the kth row of $\bar{\mathbf{Q}}^{-1}$. Note that $\boldsymbol{v}'\bar{\mathbf{Q}}\boldsymbol{\beta}_2=\beta_{2,k}$. Take $\boldsymbol{v}=(v_1,\cdots,v_{K_{p,s}})$ where $v_j=\bar{v}_j$ for $|j-k|\le M_n$ and zero otherwise. Clearly, $\boldsymbol{v}\in\mathcal{V}$ on an event $\mathcal{A}_4$ with $\mathbb{P}(\mathcal{A}_4)\to 1$. On $\mathcal{A}_2\cap\mathcal{A}_4$,

$$|(\boldsymbol{v}-\bar{\boldsymbol{v}})'\bar{\mathbf{Q}}\boldsymbol{\beta}_2|\lesssim\varepsilon_n J^{-1}r_{2,n}n^{-c_6}$$

for some large $c_6>0$ 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 $\mathcal{O}_\ell=\{\boldsymbol{\beta}\in\mathbb{R}^{K_{p,s}}:2^{\ell-1}\mathfrak{r}_{2,n}\le\|\boldsymbol{\beta}-\breve{\boldsymbol{\beta}}_0-\bar{\boldsymbol{\beta}}\|_\infty\le 2^\ell\mathfrak{r}_{2,n}\}$ for the smallest $\bar{L}$ such that $2^{\bar{L}}r_{2,n}\ge c$, 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 $L\le\bar{L}$, we have by Lemma SA-3.5 and the results given in the previous steps,

$$\mathbb{P}(\|\breve{\boldsymbol{\beta}}-\breve{\boldsymbol{\beta}}_0-\bar{\boldsymbol{\beta}}\|_\infty\ge 2^L r_{2,n}|\mathscr{F}_{XW\Delta})$$

$$\le\mathbb{P}\Big(\bigcup_{\ell=L}^{\bar{L}}\Big\{\inf_{\boldsymbol{\beta}\in\mathcal{O}_\ell}\sup_{\boldsymbol{v}\in\mathcal{V}}\mathbb{E}_n[\rho(y_i;\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))-\rho(y_i;\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{v})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))]<0\Big\}\Big|\mathscr{F}_{XW\Delta}\Big)+o_{\mathbb{P}}(1)$$

$$=\mathbb{P}\Big(\bigcup_{\ell=L}^{\bar{L}}\Big\{\inf_{\boldsymbol{\beta}\in\mathcal{O}_\ell}\sup_{\boldsymbol{v}\in\mathcal{V}}\Big\{\mathbb{E}\Big[\rho(y_i;\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))-\rho(y_i;\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{v})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))$$

$$-[\eta(\breve{\mathbf{b}}_i'\boldsymbol{\beta}+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})-\eta(\breve{\mathbf{b}}_i'(\boldsymbol{\beta}-\boldsymbol{v})+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})]\psi(y_i,\eta(\breve{\mathbf{b}}_i'\breve{\boldsymbol{\beta}}_0+\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))|\mathscr{F}_{XW\Delta}\Big]+$$

$$\mathbb{E}_n\Big[(\eta(\check{\mathbf{b}}_i'\boldsymbol{\beta} + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}) - \eta(\check{\mathbf{b}}_i'(\boldsymbol{\beta} - \boldsymbol{v}) + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))\psi(y_i, \eta(\check{\mathbf{b}}_i'\check{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}))\Big] +$$

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

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

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

$$C_4 J^{-1} 2^\ell r_{2,n}\varepsilon_n\Big\} \cap \mathcal{A}\Big|\mathscr{F}_{XW\Delta}\Big) + o_{\mathbb{P}}(1)$$

$$\leq \sum_{\ell=L}^{\bar{L}}(C_6 J^{-1} 2^\ell \mathfrak{r}_{2,n}\varepsilon_n)^{-1}\mathbb{1}(\mathcal{A}_1)\mathbb{E}\Big[\sup_{\boldsymbol{\beta}_1 \in \mathcal{H}_1}\sup_{\boldsymbol{\beta}_2 \in \mathcal{H}_{2,\ell}}\sup_{\boldsymbol{\gamma}_1 \in \mathcal{H}_3}\sup_{\boldsymbol{v} \in \mathcal{V}}\frac{1}{\sqrt{n}}\mathbb{G}_n[\delta_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v}, \boldsymbol{\gamma})]\Big|\mathscr{F}_{XW\Delta}\Big] + o_{\mathbb{P}}(1),$$

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 $\varepsilon_n = 2^L r_{2,n}$, 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 for part (i) is complete.

Step 7: To show part (ii) and part (iii), by Taylor expansion and the result in part (i),

$$\eta(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}) - \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)$$

$$= \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\Big(\widehat{\mathbf{b}}_{p,s}(x)'\widehat{\boldsymbol{\beta}} - \mu_0(x)\Big)$$

$$+ O_{\mathbb{P}}\Big(\|\widehat{\mathbf{w}} - \mathbf{w}\| + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| + \frac{J\log n}{n} + J^{-2p-2} + \mathfrak{r}_{2,n}^2\Big)$$

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

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

and

$$\eta^{(1)}(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}})\widehat{\mu}^{(1)}(x) - \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\mu_0^{(1)}(x)$$

$$= \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\Big(\widehat{\mu}^{(1)}(x) - \mu_0^{(1)}(x)\Big)$$

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

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

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

$$+ \|\widehat{\mathsf{w}} - \mathsf{w}\| \Big(1 + \Big(\frac{J^3 \log n}{n}\Big)^{1/2}\Big)\Big).$$

In the above derivation the probability bound holds uniformly over $x \in \mathcal{X}$ as well. Then the proof is complete. $\qquad\square$

SA-5.8   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

$$\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\eta}_{i,1}^2\psi^\ddagger(\widehat{\eta}_i)^2\psi^\dagger(\widehat{\epsilon}_i)^2] - \mathbb{E}[\mathbf{b}_{p,s}(x_i)\mathbf{b}_{p,s}(x_i)'\eta_{i,1}^2\sigma^2(x_i, \mathbf{w}_i)]$$

$$= \mathbb{E}_n\Big[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\eta}_{i,1}^2\psi^\ddagger(\widehat{\eta}_i)^2\Big(\psi^\dagger(\epsilon_i + u_i)^2 - \psi^\dagger(\epsilon_i)^2\Big)\Big]$$

$$\quad + \mathbb{E}_n\Big[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\Big(\widehat{\eta}_{i,1}^2\psi^\ddagger(\widehat{\eta}_i)^2 - \eta_{i,1}^2\psi^\ddagger(\eta_i)^2\Big)\psi^\dagger(\epsilon_i)^2\Big]$$

$$\quad + \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2(\psi(y_i, \eta_i)^2 - \sigma^2(x_i, \mathbf{w}_i))]$$

$$\quad + \Big(\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2\sigma^2(x_i, \mathbf{w}_i)] - \mathbb{E}[\mathbf{b}_{p,s}(x_i)\mathbf{b}_{p,s}(x_i)'\eta_{i,1}^2\sigma^2(x_i, \mathbf{w}_i)]\Big)$$

$$=: \mathbf{V}_1 + \mathbf{V}_2 + \mathbf{V}_3 + \mathbf{V}_4.$$

We bound each term in the following. The first part of the theorem only concerns $\mathbf{V}_1 + \mathbf{V}_2 + \mathbf{V}_3$, and the second part needs a bound on $\mathbf{V}_4$ as well where the additional Assumption SA-RP(ii) is used.

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

$$\mathbf{V}_{11} := \mathbb{E}_n\Big[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2\psi^\ddagger(\eta_i)^2\Big(\psi^\dagger(\epsilon_i + u_i)^2 - \psi^\dagger(\epsilon_i)^2\Big)\Big],$$

$$\mathbf{V}_{12} := \mathbb{E}_n\Big[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\Big(\widehat{\eta}_{i,1}^2\psi^\ddagger(\widehat{\eta}_i)^2 - \eta_{i,1}^2\psi^\ddagger(\eta_i)^2\Big)\Big(\psi^\dagger(\epsilon_i + u_i)^2 - \psi^\dagger(\epsilon_i)^2\Big)\Big].$$

Let $r_{1,n} = C_1(J \log n/n)^{1/2} + J^{-p-1}$ for a constant $C_1 > 0$. By Assumption SA-SM and Corollary SA-3.1, $\max_{1 \le i \le n} |u_i| \le r_{1,n}$ with arbitrarily large probability for C_1 sufficiently large. For $\mathbf{V}_{11}$, let $\mathcal{J}$ be the set of all 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 $\mathcal{D} := \{a : |a - \jmath| \le r_{1,n} \text{ for some } \jmath \in \mathcal{J}\}$. Define

$$\mathbf{V}_{111} := \mathbb{E}_n\Big[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2\psi^\ddagger(\eta_i)^2\Big(\psi^\dagger(\epsilon_i + u_i)^2 - \psi^\dagger(\epsilon_i)^2\Big)\mathbb{1}_{i,\mathcal{D}}\Big],$$

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

By definition of $\mathcal{D}$ and Assumption SA-SM,

$$\|\mathbf{V}_{111}\| \lesssim \|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\mathbb{E}[\mathbb{1}_{i,\mathcal{D}}|\mathscr{F}_{XW\Delta}]]\| + \|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'(\mathbb{1}_{i,\mathcal{D}}-\mathbb{E}[\mathbb{1}_{i,\mathcal{D}}|\mathscr{F}_{XW\Delta}])]\|.$$

By Assumption SA-SM and Lemma SA-3.5 of Cattaneo et al. (2024b), the first term on the right hand side is $O_{\mathbb{P}}(r_{1,n})$. For the second term, conditional on $\mathscr{F}_{XW\Delta}$, 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 $O_{\mathbb{P}}(\sqrt{r_{1,n}J\log J/n}+J\log J/n)$. In this case, the indicator $\mathbb{1}_{i,\mathcal{D}}$ is trivially bounded uniformly.

On the other hand, by Assumption SA-SM,

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

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

$$\mathbb{E}_n\left[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2\psi^{\ddagger}(\eta_i)^2|\psi^{\dagger}(\epsilon_i)|\right] \leq \frac{1}{2}\mathbb{E}_n\left[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2\psi^{\ddagger}(\eta_i)^2(1+\psi^{\dagger}(\epsilon_i)^2)\right] \lesssim_{\mathbb{P}} 1,$$

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

$$\mathbb{E}_n\left[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2\psi^{\ddagger}(\eta_i)^2|\psi^{\dagger}(\epsilon_i+u_i)|\right]$$
$$= \mathbb{E}_n\left[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\eta_{i,1}^2\psi^{\ddagger}(\eta_i)^2|\psi^{\dagger}(\epsilon_i)+(\psi^{\dagger}(\epsilon_i+u_i)-\psi^{\dagger}(\epsilon_i))|\right].$$

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-SM, Corollary SA-3.1 and the argument given later in Step 3, we have

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

Step 3: For $\mathbf{V}_3$, in view of Lemmas SA-5.2 and SA-5.3, it suffices to show that

$$\sup_{\Delta \in \Pi} \left\| \mathbb{E}_n[\mathbf{b}_{p,0}(x_i; \Delta)\mathbf{b}_{p,0}(x_i; \Delta)' \eta_{i,1}^2 (\psi(y_i, \eta_i)^2 - \sigma^2(x_i, \mathbf{w}_i))] \right\| \lesssim_{\mathbb{P}} \left(\frac{J \log J}{n^{\frac{\nu-2}{\nu}}} \right)^{1/2}.$$

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

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

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

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

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

$$\lesssim JM \sup_{x \in \mathcal{X}, \mathbf{w} \in \mathcal{W}} \mathbb{E}\left[|\varphi_1| \Big| x_i = x \right] \lesssim JM.$$

The VC condition holds by the same argument given in the proof of Lemma SA-3.1. Then, by Lemma SA-5.6,

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

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

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} \left| \mathbb{E}_n[g(x_i, \mathbf{w}_i, \varphi_i^+)] \right| \right] \lesssim \frac{1}{\sqrt{n}} J \mathbb{E}\left[\sqrt{\mathbb{E}_n[|\varphi_i^+|^2]} \right]$$

$$\leq \frac{1}{\sqrt{n}} J (\mathbb{E}[\max_{1 \leq i \leq n} |\varphi_i^+|])^{1/2} (\mathbb{E}[\mathbb{E}_n[|\varphi_i^+|]])^{1/2}$$

$$\lesssim \frac{J}{\sqrt{n}} \cdot \frac{n^{\frac{1}{\nu}}}{M^{(\nu-2)/4}},$$

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

$$\mathbb{E}[\max_{1 \leq i \leq n} |\varphi_i^+|] \lesssim \mathbb{E}[\max_{1 \leq i \leq n} \psi(y_i, \eta_i)^2] \lesssim n^{2/\nu} \quad \text{and} \quad \mathbb{E}[\mathbb{E}_n[|\varphi_i^+|]] \leq \mathbb{E}[|\varphi_1|^+|] \lesssim \frac{\mathbb{E}[|\psi(y_1, \eta_1)|^\nu]}{M^{(\nu-2)/2}}.$$

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

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

$$\left\| \mathbb{E}_{\widehat{\Delta}} \left[\widehat{\mathbf{b}}_{p,s}(x_i) \widehat{\mathbf{b}}_{p,s}(x_i)' \eta_{i,1}^2 \psi(y_i, \eta_i)^2 \right] - \mathbb{E} \left[\mathbf{b}_{p,s}(x_i) \mathbf{b}_{p,s}(x_i)' \eta_{i,1}^2 \psi(y_i, \eta_i)^2 \right] \right\| \lesssim_{\mathbb{P}} \sqrt{J \log J / n} + \mathfrak{r}_{\mathrm{RP}}.$$

The proof for the first conclusion is complete.

Step 5: The results about $\widehat{\Omega}_{\mu^{(v)}}(x)$, $\widehat{\Omega}_{\vartheta}(x)$ and $\widehat{\Omega}_{\zeta}(x)$ follow by Assumptions SA-SM and SA-HLE, Lemmas SA-5.4 and SA-3.1, and Corollary SA-3.1. The proof is complete.

$\square$

SA-5.9   Proof of Theorem SA-3.3

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

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

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

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

By Lemmas SA-5.4, SA-3.1 and SA-3.2,

$$\frac{1}{n^{3/2}} \sum_{i=1}^n \mathbb{E} \left[|a_i \psi(y_i, \eta_i)|^3 \Big| \mathscr{F}_{XW\Delta} \right]$$

$$\lesssim \bar{\Omega}_{\mu^{(v)}}(x)^{-3/2} \frac{1}{n^{3/2}} \sum_{i=1}^n \mathbb{E} \left[|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)' \bar{\mathbf{Q}}^{-1} \widehat{\mathbf{b}}_{p,s}(x_i) \eta_{i,1} \psi(y_i, \eta_i)|^3 \Big| \mathscr{F}_{XW\Delta} \right]$$

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

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

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

since $J/n = o(1)$. By Theorem SA-3.2, the above weak convergence still holds if $\bar{\Omega}_{\mu^{(v)}}(x)$ is replaced by $\widehat{\Omega}_{\mu^{(v)}}(x)$. Then, the desired results follow by Theorem SA-3.1. $\qquad\square$

SA-5.10 Proof of Theorem SA-3.4

Proof. We let $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{r}_{0,v}$ be defined as in Lemma SA-5.5. By Lemmas SA-5.5 and SA-3.1, Theorem SA-3.1 and the results given in the proof of Lemma SA-3.4, we have

$$\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x) = \widehat{\mathbf{b}}_{p,s}(x_i)'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0) - \widehat{r}_{0,v}(x)$$

$$= -\widehat{\mathbf{b}}_{p,s}^{(v)}(x)' \mathbf{Q}_0^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\eta_{i,1}\psi(y_i, \eta_i)] - \widehat{\mathbf{b}}_{p,s}^{(v)}(x)' \mathbf{Q}_0^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\eta_{i,1}\Psi(x_i, \mathbf{w}_i; \check{\eta}_i)]$$

$$- \widehat{r}_{0,v}(x) + O_{\mathbb{P}}\Big(J^v \Big\{ \Big(\frac{J \log n}{n} \Big)^{3/4} \sqrt{\log n} + J^{-\frac{p+1}{2}} \Big(\frac{J \log^2 n}{n} \Big)^{1/2} + \mathfrak{r}_\gamma \Big\} \Big),$$

where $\check{\eta}_i = \eta(\widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}}_0 + \mathbf{w}_i'\boldsymbol{\gamma}_0)$. Recall that the $O_{\mathbb{P}}(\cdot)$ in the last line 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 imposed. Then,

$$\mathtt{AISE}_{\mu^{(v)}} = \int_{\mathcal{X}} \Big(\widehat{\mathbf{b}}_{p,s}^{(v)}(x)' \mathbf{Q}_0^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\eta_{i,1}\psi(y_i, \eta_i)]$$

$$+ \widehat{\mathbf{b}}_{p,s}^{(v)}(x)' \mathbf{Q}_0^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\eta_{i,1}\Psi(x_i, \mathbf{w}_i; \check{\eta}_i)] + \widehat{r}_{0,v}(x) \Big)^2 \omega(x) dx.$$

Next, taking conditional expectation given $\mathbf{X}$, $\mathbf{W}$ and $\widehat{\Delta}$ and using the argument in the proof of Lemma SA-3.1 again, we have

$$\mathbb{E}[\mathtt{AISE}_{\mu^{(v)}} | \mathbf{X}, \mathbf{W}, \widehat{\Delta}] = \frac{1}{n} \operatorname{trace} \Big(\mathbf{Q}_0^{-1} \boldsymbol{\Sigma}_0 \mathbf{Q}_0^{-1} \int_{\mathcal{X}} \mathbf{b}_{p,s}^{(v)}(x) \mathbf{b}_{p,s}^{(v)}(x)' \omega(x) dx \Big) + o_{\mathbb{P}}(J^{2v+1}/n)$$

$$+ \int_{\mathcal{X}} \Big(\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0 - \mu_0^{(v)}(x) \Big)^2 \omega(x) dx$$

$$+ \int_{\mathcal{X}} \Big(\widehat{\mathbf{b}}_{p,s}^{(v)}(x)' \mathbf{Q}_0^{-1} \mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\eta_{i,1}\Psi(x_i, \mathbf{w}_i; \check{\eta}_i)] \Big)^2 \omega(x) dx$$

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

By Assumption SA-SM, $\Psi(x_i, \mathbf{w}_i; \check{\eta}_i) = -\Psi_1(x_i, \mathbf{w}_i; \eta_{i,0})\eta_{i,1}\widehat{r}_0(x_i) + O_{\mathbb{P}}(J^{-2p-2})$ where $O_{\mathbb{P}}(\cdot)$ holds uniformly over i. The terms in the last three lines correspond to the integrated squared bias. Also, using the same argument in the proof of Lemma SA-3.1, $\mathbb{E}_n[\cdot]$ in the last two lines can be safely replaced by $\mathbb{E}_{\widehat{\Delta}}[\cdot]$, which only introduces some additional approximation error of order $o_{\mathbb{P}}(J^{-2p-2+2v})$.

The proof of Theorem SA-3.4 in Cattaneo et al. (2024b) shows that

$$
\begin{aligned}
\widehat{r}_{0,v}(x) &= \mu_0^{(v)}(x) - \widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\boldsymbol{\beta}}_0 \\
&= \frac{J^{-p-1+v}\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x - \hat{\tau}_x^{\mathrm{L}}}{\hat{h}_x}\Big) \\
&\quad - J^{-p-1}\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_{p,0}(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i - \hat{\tau}_{x_i}^{\mathrm{L}}}{\hat{h}_{x_i}}\Big)\Big] + o_{\mathbb{P}}(J^{-p-1+v}),
\end{aligned}
$$

where $\hat{\tau}_x^{\mathrm{L}}$ is the start of the (random) interval in $\widehat{\Delta}$ containing x and $\hat{h}_x$ denotes its length. Then, using the same argument as in the proof of Theorem SA-3.4 in Cattaneo et al. (2024b), we can approximate the integrated squared bias by the analogue based on the non-random partition Δ_0, i.e., $\int_{\mathcal{X}}(r_{0,v}^{\dagger}(x) - \mathbf{b}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i, \mathbf{w}_i)r_{0,0}^{\dagger}(x_i)])^2\omega(x)dx$ where

$$
\begin{aligned}
r_{0,v}^{\dagger}(x) &= \frac{J^{-p-1+v}\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x - \tau_x^{\mathrm{L}}}{h_x}\Big) \\
&\quad - J^{-p-1}\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}\Big[\mathbf{b}_{p,0}(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i - \tau_{x_i}^{\mathrm{L}}}{h_{x_i}}\Big)\Big].
\end{aligned}
$$

The expression of the bias term can be further simplified. For both $R_v(x) = r_{0,v}^{\dagger}(x)$ and $R_v(x) = r_{0,v}^{\star}(x)$, there exists some vector $\boldsymbol{\beta}$ such that $\sup_{x \in \mathcal{X}}|\mu_0(x) - \mathbf{b}_{p,s}(x_i)'\boldsymbol{\beta} - R_v(x)| = o(J^{-p-1+v})$ (see Lemma SA-5.5 and Lemma SA-6.1 of Cattaneo et al. (2020)). Define

$$
r_{0,v}^{\mathrm{P}}(x) = \mu_0^{(v)}(x) - \mathbf{b}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i, \mathbf{w}_i)\mu_0(x_i)].
$$

Then, it follows that $r_{0,v}^{\mathrm{P}}(x) = R_v(x) - \mathbf{b}_{p,s}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i, \mathbf{w}_i)R_0(x_i)] + o(J^{-p-1+v})$. Thus,

$$
\begin{aligned}
&\{r_{0,v}^{\dagger}(x) - \mathbf{b}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i, \mathbf{w}_i)r_{0,0}^{\dagger}(x_i)]\} \\
&- \{[r_{0,v}^{\star}(x) - \mathbf{b}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbb{E}[\mathbf{b}_{p,s}(x_i)\varkappa(x_i, \mathbf{w}_i)r_{0,0}^{\star}(x_i)]]\} = o(J^{-p-1+v}).
\end{aligned}
$$

Therefore, the expression of $\mathscr{B}_n(p, s, v)$ given in the theorem holds.

Finally, the desired results in part (ii) and part (iii) follow by Theorem SA-3.1, the rate condition imposed and the same argument for part (i). $\qquad\square$

SA-5.11 Proof of Theorem SA-3.5

Proof. The proof is divided into several steps.

Step 1: Note that

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

where the last step uses Lemma SA-3.2 and Corollary SA-3.1. Then, in view of Lemmas SA-5.5, SA-3.4, Theorems SA-3.1, SA-3.2 and the rate restriction given in the lemma, we have

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

Step 2: Let us write $\mathscr{K}(x, x_i) = \Omega_{\mu^{(v)}}(x)^{-1/2}\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_{p,s}(x_i)$ (the dependence of $\widehat{\mathbf{b}}_{p,s}^{(v)}(x)$, $\bar{\mathbf{Q}}$ and $\bar{\Omega}_{\mu^{(v)}}(x)$ on $\mathbf{X}$, $\mathbf{W}$ and $\widehat{\Delta}$ is omitted for simplicity), and $\tilde{\sigma}^2(x_i, \mathbf{w}_i) = \mathbb{E}[\psi^\dagger(\epsilon_i)^2 | x_i, \mathbf{w}_i]$. 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 $\{\tilde{\sigma}^2(x_i, \mathbf{w}_i)\}_{i=1}^n$ are ordered as concomitants $\{\epsilon_{[i]}\}_{i=1}^n$, $\{\mathbf{w}_{[i]}\}$ and $\{\tilde{\sigma}_{[i]}^2\}_{i=1}^n$ where $\tilde{\sigma}_{[i]}^2 = \tilde{\sigma}^2(x_{(i)}, \mathbf{w}_{[i]})$. Clearly, conditional on $\mathscr{F}_{XW\Delta}$ (the σ-field generated by $\{(x_i, \mathbf{w}_i)\}$ and $\widehat{\Delta}$), $\{\psi^\dagger(\epsilon_{[i]})\}_{i=1}^n$ is still an independent mean-zero sequence. Then by Assumptions SA-DGP, SA-SM 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

$$
\max_{1 \leq \ell \leq n} |S_\ell| := \max_{1 \leq \ell \leq n} \left| \sum_{i=1}^{\ell} \eta^{(1)}(\mu_0(x_{(i)}) + \mathbf{w}_{[i]}'\boldsymbol{\gamma}_0)\psi^\ddagger(\eta(\mu_0(x_{(i)}) + \mathbf{w}_{[i]}'\boldsymbol{\gamma}_0))\psi^\dagger(\epsilon_{[i]})
$$

$$-\sum_{i=1}^{\ell}\eta^{(1)}(\mu_0(x_{(i)})+\mathbf{w}'_{[i]}\boldsymbol{\gamma}_0)\psi^{\ddagger}(\eta(\mu_0(x_{(i)})+\mathbf{w}'_{[i]}\boldsymbol{\gamma}_0))\tilde{\sigma}_{[i]}\zeta_{[i]}\Big|\lesssim_{\mathbb{P}} n^{\frac{1}{\nu}}.$$

Then, using summation by parts,

$$\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^{n}\mathcal{K}(x,x_{(i)})\eta^{(1)}(\mu_0(x_{(i)})+\mathbf{w}'_{[i]}\boldsymbol{\gamma}_0)\psi^{\ddagger}(\eta(\mu_0(x_{(i)})+\mathbf{w}'_{[i]}\boldsymbol{\gamma}_0))(\psi^{\dagger}(\epsilon_{[i]})-\tilde{\sigma}_{[i]}\zeta_{[i]})\right|$$

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

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

$$\leq\sup_{x\in\mathcal{X}}\max_{1\leq i\leq n}|\mathcal{K}(x,x_i)||S_n|+\sup_{x\in\mathcal{X}}\left\|\frac{\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}^{(v)}_{p,s}(x)}{\sqrt{\bar{\Omega}_{\mu^{(v)}}(x)}}\right\|_1\left\|\sum_{i=1}^{n-1}S_i\left(\widehat{\mathbf{b}}_{p,s}(x_{(i+1)})-\widehat{\mathbf{b}}_{p,s}(x_{(i)})\right)\right\|_{\infty}.$$

By Lemmas SA-5.4, SA-3.1 and SA-3.2, $\sup_{x\in\mathcal{X}}\sup_{x_i\in\mathcal{X}}|\mathcal{K}(x,x_i)|\lesssim_{\mathbb{P}}\sqrt{J}$, and

$$\sup_{x\in\mathcal{X}}\left\|\frac{\bar{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}^{(v)}_{p,s}(x)}{\sqrt{\bar{\Omega}_{\mu^{(v)}}(x)}}\right\|_1\lesssim_{\mathbb{P}}1.$$

Then, notice that

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

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

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

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

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

Step 3: Suppose that Assumption SA-RP(ii) also holds. Note that

$$\sup_{x \in \mathcal{X}} |\bar{Z}_{\mu^{(v)},p}(x) - Z_{\mu^{(v)},p}(x)|$$

$$\leq \sup_{x \in \mathcal{X}} \left| \frac{\widehat{\mathbf{b}}^{(v)}(x)'(\bar{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1})}{\sqrt{\Omega_{\mu^{(v)}}(x)}} \bar{\mathbf{\Sigma}}^{1/2} \mathbf{N}_{K_{p,s}} \right| + \sup_{x \in \mathcal{X}} \left| \frac{\widehat{\mathbf{b}}^{(v)}(x)'\mathbf{Q}_0^{-1}}{\sqrt{\Omega_{\mu^{(v)}}(x)}} \left(\bar{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}_0^{1/2} \right) \mathbf{N}_{K_{p,s}} \right| +$$

$$\sup_{x \in \mathcal{X}} \left| \frac{\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'(\widehat{\mathbf{T}}_s - \mathbf{T}_s)\mathbf{Q}_0^{-1}}{\sqrt{\Omega_{\mu^{(v)}}(x)}} \mathbf{\Sigma}_0^{1/2} \mathbf{N}_{K_{p,s}} \right| + \sup_{x \in \mathcal{X}} \left| \left(\frac{1}{\sqrt{\bar{\Omega}_{\mu^{(v)}}(x)}} - \frac{1}{\sqrt{\Omega_{\mu^{(v)}}(x)}} \right) \widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\widehat{\mathbf{T}}_s \bar{\mathbf{Q}}^{-1} \bar{\mathbf{\Sigma}}^{1/2} \mathbf{N}_{K_{p,s}} \right|$$

$$= I + II + III + IV,$$

where each term on the right-hand side is a mean-zero Gaussian process conditional on $\mathscr{F}_{XW\Delta}$. By Theorem SA-3.2 (see Step 4 of its proof), $\sup_{x \in \mathcal{X}} |\bar{\Omega}_{\mu^{(v)}}(x) - \Omega_{\mu^{(v)}}(x)| \lesssim_{\mathbb{P}} J^{1+2v}(\sqrt{J \log n/n} + \mathfrak{r}_{\mathsf{RP}})$. By a similar calculation given in Step 1 and the rate condition imposed, the last term is $o_{\mathbb{P}}(a_n^{-1})$. By Lemmas SA-5.3 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-5.4 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),

$$\mathbb{E}\Big[I + II + III \Big| \mathscr{F}_{XW\Delta} \Big] \lesssim_{\mathbb{P}} \sqrt{\log J} \Big(\|\bar{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}_0^{1/2}\| + \|\bar{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\| + \|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \Big) = o_{\mathbb{P}}(a_n^{-1})$$

where the last line follows from the imposed rate restriction. Then the proof for part (i) is complete.

The results in parts (ii) and (iii) immediately follow by Theorem SA-3.1 and the fact that the leading variance term in the Bahadur representation for $\widehat{\vartheta}(x, \widehat{\mathbf{w}})$ or $\widehat{\zeta}(x, \widehat{\mathbf{w}})$ differs from that for $\widehat{\mu}(x)$ or $\widehat{\mu}^{(1)}(x)$ up to a sign only. □

SA-5.12 Proof of Theorem SA-3.6

Proof. This conclusion follows from Lemmas SA-5.4, SA-3.1, Theorem SA-3.2 and Gaussian Maximal Inequality as applied in Step 3 in the proof of Theorem SA-3.5. □

SA-5.13   Proof of Theorem SA-3.7

Proof. We first show that

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|T_{\mu^{(v)},p}(x)|\le u\Big)-\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\le u\Big)\Big|=o(1).$$

By Theorem SA-3.5, there exists a sequence of constants ξ_n such that $\xi_n=o(1)$ and

$$\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|T_{\mu^{(v)},p}(x)|-\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\Big|>\xi_n/a_n\Big)=o(1).$$

Then,

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|T_{\mu^{(v)},p}(x)|\le u\Big)\le\mathbb{P}\Big(\Big\{\sup_{x\in\mathcal{X}}|T_{\mu^{(v)},p}(x)|\le u\Big\}\cap\Big\{\Big|\sup_{x\in\mathcal{X}}|T_{\mu^{(v)},p}(x)|-\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\Big|\le\xi_n/a_n\Big\}\Big)+o(1)$$

$$\le\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\le u+\xi_n/a_n\Big)+o(1)$$

$$\le\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\le u\Big)+\sup_{u\in\mathbb{R}}\mathbb{E}\Big[\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|-u\Big|\le\xi_n/a_n\Big|\widehat{\Delta}\Big)\Big]$$

$$\le\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\le u\Big)+\mathbb{E}\Big[\sup_{u\in\mathbb{R}}\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|-u\Big|\le\xi_n/a_n\Big|\widehat{\Delta}\Big)\Big]+o(1).$$

Apply the Anti-Concentration Inequality conditional on $\widehat{\Delta}$ (Chernozhukov et al., 2014) to the second term:

$$\sup_{u\in\mathbb{R}}\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|-u\Big|\le\xi_n/a_n\Big|\widehat{\Delta}\Big)\le4\xi_n a_n^{-1}\mathbb{E}\Big[\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\Big|\widehat{\Delta}\Big]+o(1)$$

$$\lesssim_{\mathbb{P}}\xi_n a_n^{-1}\sqrt{\log J}+o(1)\to0$$

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

$$\mathbb{E}\Big[\sup_{u\in\mathbb{R}}\mathbb{P}\Big(\Big|\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|-u\Big|\le\xi_n/a_n\Big|\widehat{\Delta}\Big)\Big]=o(1).$$

The other side of the inequality follows similarly.

By similar argument, using Theorem SA-3.6, we have

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{x\in\mathcal{X}}|\widehat{Z}_{\mu^{(v)},p}(x)|\leq u\Big|\mathbf{D}\right)-\mathbb{P}\left(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\leq u\Big|\widehat{\Delta}\right)\right|=o_{\mathbb{P}}(1).$$

Then, it remains to show that

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\leq u\right)-\mathbb{P}\left(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\leq u|\widehat{\Delta}\right)\right|=o_{\mathbb{P}}(1). \tag{SA-5.1}$$

We can write

$$Z_{\mu^{(v)},p}(x)=\frac{\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'}{\sqrt{\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\mathbf{V}_0\widehat{\mathbf{b}}_{p,0}^{(v)}(x)}}\breve{\mathbf{N}}_{K_{p,0}}$$

where $\mathbf{V}_0=\mathbf{T}_s'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0\mathbf{Q}_0^{-1}\mathbf{T}_s$ and $\breve{\mathbf{N}}_{K_{p,0}}:=\mathbf{T}_s'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0^{1/2}\mathbf{N}_{K_{p,s}}$ is a $K_{p,0}$-dimensional Gaussian random vector. Importantly, by this construction, $\breve{\mathbf{N}}_{K_{p,0}}$ and $\mathbf{V}_0$ do not depend on $\widehat{\Delta}$ and x, and they are only determined by the deterministic partition Δ_0.

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

$$\mathbf{b}_{p,0}^{(0)}(x;\Delta_1)=\mathbf{b}_{p,0}^{(0)}(\breve{x};\Delta_2),$$

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,

$$\mathbb{E}\left[\mathbb{P}\left(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\leq u\Big|\widehat{\Delta}\right)\right]=\mathbb{P}\left(\sup_{x\in\mathcal{X}}|Z_{\mu^{(v)},p}(x)|\leq u\Big|\widehat{\Delta}\right)+o_{\mathbb{P}}(1).$$

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

For $v>0$, simply notice that $\widehat{\mathbf{b}}_{p,0}^{(v)}(x)=\widehat{\mathfrak{T}}_v\widehat{\mathbf{b}}_{p,0}(x)$ 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 Lemma SA-5.2, it can be shown that $\|\widehat{\mathfrak{T}}_v-\mathfrak{T}_v\|\lesssim J^v\sqrt{J\log J/n}$ where $\mathfrak{T}_v$ is the population analogue ($\hat{h}_j$ replaced by h_j). Repeating the argument given in the proof of Theorems SA-3.5 and SA-3.6, we can replace $\widehat{\mathfrak{T}}_v$ in $Z_{\mu^{(v)},p}(x)$ by $\mathfrak{T}_v$ without affecting the approximation rate. Then the desired result for $T_{\mu^{(v)},p}(x)$ follows by repeating the argument given for $v=0$ above.

Finally, the result for $T_{\vartheta,p}(x)$ ($T_{\zeta,p}(x)$) follows by the fact that $Z_{\vartheta,p}(x)$ and $\widehat{Z}_{\vartheta,p}(x)$ ($Z_{\zeta,p}(x)$ and $\widehat{Z}_{\zeta,p}(x)$) differ from $Z_{\mu^{(v)},p}(x)$ and $\widehat{Z}_{\mu^{(v)},p}(x)$ up to a sign only. $\qquad\square$

SA-5.14   Proof of Theorem SA-3.8

Proof. We only consider $\widehat{I}_{\mu^{(v)},p}(x)$. The results in part (ii) and part (iii) follow similarly.

Let $\xi_{1,n} = o(1)$, $\xi_{2,n} = o(1)$ and $\xi_{3,n} = o(1)$. Then,

$$
\begin{aligned}
\mathbb{P}\left[\sup_{x\in\mathcal{X}} |T_{\mu^{(v)},p}(x)| \le \mathfrak{c}_{\mu^{(v)}}\right] &\le \mathbb{P}\left[\sup_{x\in\mathcal{X}} |\bar{Z}_{\mu^{(v)},p}(x)| \le \mathfrak{c}_{\mu^{(v)},p} + \xi_{1,n}/a_n\right] + o(1) \\
&\le \mathbb{P}\left[\sup_{x\in\mathcal{X}} |\bar{Z}_{\mu^{(v)},p}(x)| \le c^0(1 - \alpha + \xi_{3,n}) + (\xi_{1,n} + \xi_{2,n})/a_n\right] + o(1) \\
&\le \mathbb{P}\left[\sup_{x\in\mathcal{X}} |\bar{Z}_{\mu^{(v)},p}(x)| \le c^0(1 - \alpha + \xi_{3,n})\right] + o(1) \to 1 - \alpha,
\end{aligned}
$$

where $c^0(1 - \alpha + \xi_{3,n})$ denotes the $(1 - \alpha + \xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}} |\bar{Z}_{\mu^{(v)},p}(x)|$ conditional on $\mathscr{F}_{XW\Delta}$ (the σ-field generated by $\mathbf{X}$, $\mathbf{W}$ and the partition $\widehat{\Delta}$), the first inequality holds by Theorem SA-3.5, the second by Lemma A.1 of Belloni et al. (2015), and the third by Anti-Concentration Inequality in Chernozhukov et al. (2014). The other side of the bound follows similarly. $\qquad\square$

SA-5.15   Proof of Theorem SA-3.9

Proof. We only consider the proof for part (i). The results in part (ii) and part (iii) follow similarly.

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 $\sqrt{\log \bar{J}} A_n \lesssim \sqrt{\frac{n}{J^{1+2v}}}$. Note that

$$
\sup_{x\in\mathcal{X}} |\dot{T}_{\mu^{(v)},p}(x)| \le \sup_{x\in\mathcal{X}} \left|\frac{\widehat{\mu}(x) - \mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right| + \sup_{x\in\mathcal{X}} \left|\frac{\mu_0^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right|.
$$

Therefore, under $\dot{\mathsf{H}}_0^{\mu^{(v)}}$,

$$
\begin{aligned}
\mathbb{P}\left[\sup_{x\in\mathcal{X}} |\dot{T}_{\mu^{(v)},p}(x)| > \mathfrak{c}_{\mu^{(v)}}\right] &\le \mathbb{P}\left[\sup_{x\in\mathcal{X}} |T_{\mu^{(v)},p}(x)| > \mathfrak{c}_{\mu^{(v)}} - \sup_{x\in\mathcal{X}} \left|\frac{\mu_0^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right|\right] \\
&\le \mathbb{P}\left[\sup_{x\in\mathcal{X}} |\bar{Z}_{\mu^{(v)},p}(x)| > \mathfrak{c}_{\mu^{(v)}} - \xi_{1,n}/a_n - \sup_{x\in\mathcal{X}} \left|\frac{\mu_0^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right|\right] + o(1)
\end{aligned}
$$

$$\leq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|\bar{Z}_{\mu^{(v)},p}(x)| > c^0(1-\alpha-\xi_{3,n}) - (\xi_{1,n}+\xi_{2,n})/a_n - \right.$$
$$\left. \sup_{x\in\mathcal{X}}\left|\frac{\mu_0^{(v)}(x) - m^{(v)}(x;\widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right|\right] + o(1)$$
$$\leq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|\bar{Z}_{\mu^{(v)},p}(x)| > c^0(1-\alpha-\xi_{3,n})\right] + o(1)$$
$$= \alpha + o(1)$$

where $c^0(1-\alpha-\xi_{3,n})$ denotes the $(1-\alpha-\xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}}|\bar{Z}_{\mu^{(v)},p}(x)|$ conditional on $\mathscr{F}_{XW\Delta}$ (the σ-field generated by $\mathbf{X}$, $\mathbf{W}$ and $\widehat{\Delta}$), the second inequality holds by Theorem SA-3.5, the third by Lemma A.1 of Belloni et al. (2015), the fourth by the fact that $\sup_{x\in\mathcal{X}}\left|\frac{\mu_0^{(v)}(x)-m^{(v)}(x;\widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right| = o_{\mathbb{P}}(\frac{1}{\sqrt{\log J}})$ and Anti-Concentration Inequality in Chernozhukov et al. (2014). The other side of the bound follows similarly.

On the other hand, under $\dot{\mathsf{H}}_A^{\mu^{(v)}}$,

$$\mathbb{P}\left[\sup_{x\in\mathcal{X}}|\dot{T}_{\mu^{(v)},p}(x)| > \mathfrak{c}_{\mu^{(v)}}\right]$$
$$= \mathbb{P}\left[\sup_{x\in\mathcal{X}}\left|T_{\mu^{(v)},p}(x) + \frac{\mu_0^{(v)}(x) - m^{(v)}(x;\bar{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}} + \frac{m^{(v)}(x;\bar{\boldsymbol{\theta}}) - m^{(v)}(x;\widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right| > \mathfrak{c}_{\mu^{(v)}}\right]$$
$$\geq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|T_{\mu^{(v)},p}(x)| < \sup_{x\in\mathcal{X}}\left|\frac{\mu_0^{(v)}(x) - m^{(v)}(x;\bar{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}} + \frac{m^{(v)}(x;\bar{\boldsymbol{\theta}}) - m^{(v)}(x;\widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}\right| - \mathfrak{c}_{\mu^{(v)}}\right]$$
$$\geq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|\bar{Z}_{\mu^{(v)},p}(x)| \leq \sqrt{\log J}A_n - \xi_{1,n}/a_n\right] - o(1)$$
$$\geq 1 - o(1).$$

where the fourth line holds by Lemma SA-3.2, Theorem SA-3.2, Theorem SA-3.5, the condition that $J^v\sqrt{J\log J/n} = o(1)$ and the definition of A_n, and the last by the Talagrand-Samorodnitsky Concentration Inequality (van der Vaart and Wellner, 1996, Proposition A.2.7). $\qquad\square$

SA-5.16 Proof of Theorem SA-3.10

Proof. We only consider the proof for part (i). The results in part (ii) and part (iii) follow similarly.

Throughout this 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-3.9. Under $\ddot{\mathsf{H}}_0^{\mu^{(v)}}$,

$$\sup_{x \in \mathcal{X}} \ddot{T}_{\mu^{(v)},p}(x) \leq \sup_{x \in \mathcal{X}} T_{\mu^{(v)},p}(x) + \sup_{x \in \mathcal{X}} \frac{|m^{(v)}(x; \bar{\boldsymbol{\theta}}) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})|}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}}.$$

Then,

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

$$\leq \mathbb{P}\Big[\sup_{x \in \mathcal{X}} \bar{Z}_{\mu^{(v)},p}(x) > \mathfrak{c}_{\mu^{(v)}} - \xi_{1,n}/a_n \Big] + o(1)$$

$$\leq \mathbb{P}\Big[\sup_{x \in \mathcal{X}} \bar{Z}_{\mu^{(v)},p}(x) > c^0(1 - \alpha - \xi_{3,n}) - (\xi_{1,n} + \xi_{2,n})/a_n \Big] + o(1)$$

$$\leq \mathbb{P}\Big[\sup_{x \in \mathcal{X}} \bar{Z}_{\mu^{(v)},p}(x) > c^0(1 - \alpha - \xi_{3,n}) \Big] + o(1)$$

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

where $c^0(1 - \alpha - \xi_{3,n})$ denotes the $(1 - \alpha - \xi_{3,n})$-quantile of $\sup_{x \in \mathcal{X}} \bar{Z}_{\mu^{(v)},p}(x)$ conditional on $\mathscr{F}_{XW\Delta}$ (the σ-field generated by $\mathbf{X}$, $\mathbf{W}$ and $\widehat{\Delta}$), the second line holds by Theorem SA-3.5, the third by Lemma A.1 of Belloni et al. (2015), the fourth by Anti-Concentration Inequality in Chernozhukov et al. (2014).

On the other hand, under $\ddot{\mathsf{H}}_A^{\mu^{(v)}}$,

$$\mathbb{P}\Big[\sup_{x \in \mathcal{X}} \ddot{T}_{\mu^{(v)},p}(x) > \mathfrak{c}_{\mu^{(v)}} \Big] = \mathbb{P}\Big[\sup_{x \in \mathcal{X}} \Big(T_{\mu^{(v)},p}(x) + \frac{\mu_0^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}} - \mathfrak{c}_{\mu^{(v)}} \Big) > 0 \Big]$$

$$\geq \mathbb{P}\Big[\sup_{x \in \mathcal{X}} |T_{\mu^{(v)},p}(x)| < \sup_{x \in \mathcal{X}} \frac{\mu_0^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}} - \mathfrak{c}_{\mu^{(v)}},$$

$$\sup_{x \in \mathcal{X}} \frac{\mu_0^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}} > \mathfrak{c}_{\mu^{(v)}} \Big]$$

$$\geq \mathbb{P}\Big[\sup_{x \in \mathcal{X}} |T_{\mu^{(v)},p}(x)| < \sup_{x \in \mathcal{X}} \frac{\mu_0^{(v)}(x) - m^{(v)}(x; \widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu^{(v)}}(x)/n}} - \mathfrak{c}_{\mu^{(v)}} \Big] - o(1)$$

$$\geq \mathbb{P}\Big[\sup_{x \in \mathcal{X}} |T_{\mu^{(v)},p}(x)| < \sqrt{\log J} A_n \Big] - o(1)$$

$$\geq \mathbb{P}\Big[\sup_{x \in \mathcal{X}} |\bar{Z}_{\mu^{(v)},p}(x)| < \sqrt{\log J} A_n - \xi_{1,n}/a_n \Big] - o(1)$$

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

where the third line holds by Lemma SA-3.2, Theorem SA-3.2, Lemma A.1 of Belloni et al. (2015), the assumption that $\sup_{x \in \mathcal{X}} |m^{(v)}(x; \widetilde{\boldsymbol{\theta}}) - m^{(v)}(x; \bar{\boldsymbol{\theta}})| = o_{\mathbb{P}}(1)$ and $J^v \sqrt{J \log J/n} = o(1)$, the fourth by definition of A_n, and the fifth by Theorem SA-3.5, and the last by Proposition A.2.7 in van der Vaart and Wellner (1996).

$\square$

References

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

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186, 345–366.

Bhatia, R. (2013), *Matrix Analysis*, Springer.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024a), "Binscatter Regressions," *Stata Journal, revised and resubmit.*

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024b), "On Binscatter," *American Economic Review*, 114, 1488–1514.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020), "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, 48, 1718–1741.

Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2024c), "Uniform Inference for Kernel Density Estimators with Dyadic Data," *Journal of the American Statistical Association*, forthcoming.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014), "Anti-Concentration and Honest Adaptive Confidence Bands," *Annals of Statistics*, 42, 1787–1818.

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

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

Devroye, L., Györfi, L., and Lugosi, G. (2013), *A Probabilistic Theory of Pattern Recognition*, Vol. 31, Springer Science & Business Media.

Giné, E., and Nickl, R. (2016), *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Vol. 40, Cambridge University Press.

Koenker, R. (2005), *Quantile Regression*, Vol. 38, Cambridge University Press.

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

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

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

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

Zhang, H., and Singer, B. H. (2010), *Recursive Partitioning and Applications*, Springer Science & Business Media.