Online Appendix
On Binscatter[*]

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

March 15, 2024

Abstract

This supplement presents additional methodological results, general theoretical results encompassing those reported in the paper, and all technical proofs. Our new theoretical results for least squares partitioning-based semi-linear series estimation and inference are of independent interest. Companion general-purpose software and replication files are available at https://nppackages.github.io/binsreg/.

[*] Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805, SES-2019432, and SES-2241575. Feng gratefully acknowledges financial support from the National Natural Science Foundation of China (NSFC) through grants 72203122, 72133002, and 72250064. The views expressed in this supplemental appendix are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

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

[‡] Macrofinance Studies, Federal Reserve Bank of New York.

[§] Department of Economics, UC Santa Barbara.

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

Contents

SA-1 Additional Methodological Results

We discuss two important issues related to the results in the main text. First, building on Section I.A, we provide two simple and stylized analytical examples which explicitly characterize the effect of using the incorrect covariate adjustment for binscatter. Second, building on Section I.B, we discuss the role of the choice of the evaluation point $\mathbf{w}$ for visualization, estimation, and inference for $\Upsilon_0(x, \mathbf{w}) = \mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$.

SA-1.1 Bias of Residualized Binscatter

We present two examples to showcase the potential problems resulting from the incorrect residualization method. In the following we use $m!!$ to denote the double factorial of a number m, $\mathsf{U}(a, b)$ to denote the uniform distribution on $[a, b]$ and $\mathsf{Bernoulli}(p)$ to denote the Bernoulli distribution with mean equal to p.

SA-1.1.1 Example 1: Gaussian Polynomial Regression Model

Suppose that for some integer $m > 1$,

$$y_i = x_i^m + w_i\gamma_0 + \epsilon_i, \qquad \gamma_0 = 0, \qquad \begin{bmatrix} x_i \\ w_i \\ \epsilon_i \end{bmatrix} \sim \mathsf{Normal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x & 0 \\ \rho\sigma_x & 1 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \right).$$

Thus, using the notation in the paper, $\mu_0(x_i) = x_i^m$ and $\mathbf{w}_i$ is scalar ($d = 1$).

Residualizing the covariate x_i with respect to the control w_i in this Gaussian model yields

$$x_i - \mathrm{L}(x_i|w_i) = x_i - \rho\sigma_x w_i,$$

The residualized covariate $x_i - \rho\sigma_x w_i$ is still supported on the whole real line, but its variance shrinks to $(1 - \rho^2)\sigma_x^2$. In addition, residualizing the outcome y_i with respect to w_i yields

$$y_i - \mathrm{L}(y_i|w_i) = y_i - (1 \quad w)(\mathbb{E}[x_i^m] \quad \mathbb{E}[x_i^m w_i])' = y_i - \alpha_0 - \alpha_1 w_i$$

where

$$\alpha_0 = \begin{cases} 0 & \text{if } m \text{ is odd} \\ \sigma_x^m (m-1)!! & \text{if } m \text{ is even} \end{cases} \quad \text{and} \quad \alpha_1 = \begin{cases} m\rho\sigma_x^m (m-2)!! & \text{if } m \text{ is odd} \\ 0 & \text{if } m \text{ is even} \end{cases}.$$

Then, letting $z_i = x_i - \rho\sigma_x w_i$, we have

$$\mathbb{E}[y_i - \mathrm{L}(y_i|w_i)|x_i - \mathrm{L}(x_i|w_i)] = \mathbb{E}[x_i^m - \alpha_0 - \alpha_1 w_i | z_i] = \mathbb{E}[x_i^m|z_i] - \alpha_0.$$

Note that $x_i|z_i \sim \mathsf{N}(z_i, \rho^2\sigma_x^2)$. Then, we can concisely write

$$\mathbb{E}[x_i^m|z_i] = \sum_{\substack{0\leq l\leq m \\ m-l \text{ is even}}} \binom{m}{l} z_i^l |\rho\sigma_x|^{m-l}(m-l-1)!!.$$

For instance, if the true underlying model is a quadratic regression model ($m = 2$) we obtain

$$\mathbb{E}[y_i - \mathrm{L}(y_i|w_i)|z_i] = (\rho^2 - 1)\sigma_x^2 + z_i^2 \qquad \text{(for } m = 2\text{)},$$

while for a cubic regression model ($m = 3$) we obtain

$$\mathbb{E}[y_i - \mathrm{L}(y_i|w_i)|z_i] = 3\rho^2\sigma_x^2 z_i + z_i^3 \qquad \text{(for } m = 3\text{)}.$$

Clearly, for $m = 2$, the residualization leads to a vertical shift of the true function (quadratic monomial). For $m = 3$, however, the problem is more severe: residualization adds a linear function of the covariate to the true function (cubic monomial), and when $|\rho\sigma_x|$ is large, the linear component $3\rho^2\sigma_x^2 z_i$ will visually dominate in a binscatter plot, leading to an incorrect "linear" specification of the model. Moreover, in any sample, this effect is likely to be amplified because z_i is more concentrated around its mean than x_i is.

Using the above results, we can even obtain the functional form of the residualized binscatter when μ_0 is any polynomial function and all variables are multivariate normal. Generally, the residualized binscatter yields a polynomial relationship between the residualized y_i and the residualized x_i that may be different from the original polynomial μ_0.

SA-1.1.2 Example 2: Semiparametric Bernoulli Model

Suppose that

$$y_i = \mu_0(x_i) + w_i\gamma_0 + \epsilon_i, \qquad \gamma_0 = 0,$$

where

$$w_i \sim \mathsf{Bernoulli}(0.5), \qquad x_i|w_i = 0 \sim \mathsf{U}(0,1), \qquad x_i|w_i = 1 \sim \mathsf{U}(1,2), \qquad \epsilon_i \perp\!\!\!\perp (x_i, w_i).$$

It follows that $x_i \sim \mathsf{U}(0,2)$. Residualizing the covariate x_i with respect to w_i yields

$$x_i - \mathrm{L}(x_i|w_i) = x_i - 0.5 - w_i \in [-0.5, 0.5]\,.$$

The support of this residualized covariate is different from that of the original one, not only in the location but also in the length.

In addition, residualizing the outcome y_i with respect to w_i yields

$$y_i - \mathrm{L}(y_i|w_i) = y_i - \alpha_0 - \delta_0 w_i$$

where $\alpha_0 = \mathbb{E}[\mu_0(x_i)|w_i = 0]$, and $\delta_0 = \mathbb{E}[\mu_0(x_i)|w_i = 1] - \mathbb{E}[\mu_0(x_i)|w_i = 0]$. Then, letting $z_i = x_i - 0.5 - w_i$, we have

$$\begin{aligned}
&\mathbb{E}[y_i - \mathrm{L}(y_i|w_i)|x_i - \mathrm{L}(x_i|w_i)] \\
&\quad= \mathbb{E}[y_i - \alpha_0 - \delta_0 w_i|z_i] \\
&\quad= (\mu_0(z_i + 0.5) - \alpha_0) \times \mathbb{P}(w_i = 0|z_i) + (\mu_0(z_i + 1.5) - \alpha_0 - \delta_0) \times \mathbb{P}(w_i = 1|z_i) \\
&\quad= \frac{1}{2}\mu_0(z_i + 0.5) + \frac{1}{2}\mu_0(z_i + 1.5) - \alpha_0 - \frac{1}{2}\delta_0.
\end{aligned}$$

Ignoring the constants, the residualized binscatter in this example characterizes a linear combination of two "horizontally shifted" versions of the true function $\mu_0(\cdot)$, which in general can be very different from the original $\mu_0(\cdot)$. For instance, consider

$$\mu_0(x) = x^2\mathbb{1}(x \in [0,1)) + (2 - (x-2)^2)\mathbb{1}(x \in [1,2]),$$

which is continuously differentiable. This specification actually implies that y_i and x_i have a quadratic relationship which is heterogeneous across the two groups with $w_i = 0$ and $w_i = 1$. However, the residualized binscatter yields

$$\mathbb{E}[y_i - \mathrm{L}(y_i|w_i)|x_i - \mathrm{L}(x_i|w_i)] = z_i + 1 - \alpha_0 - \frac{1}{2}\delta_0$$

which becomes a linear function in z_i, thereby giving a (visually and theoretically) wrong functional form for the true underlying conditional expectation.

SA-1.2 Impact of Evaluation Point $\mathbf{w}$

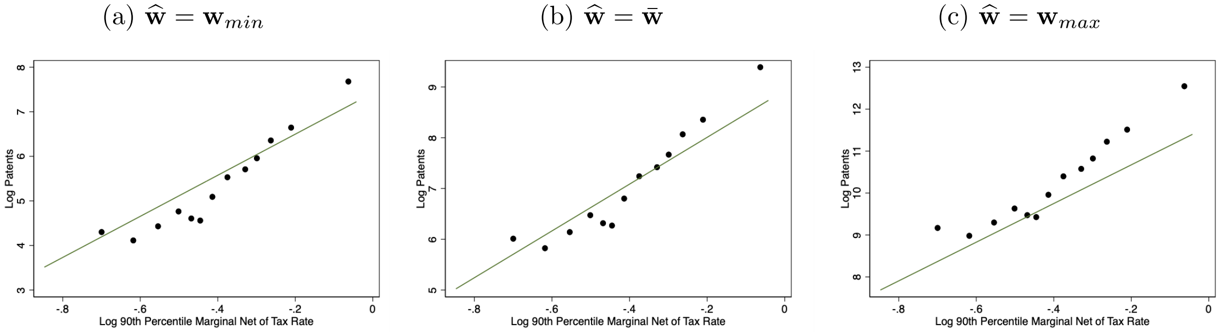
This supplemental appendix will focus on estimation and inference for the conditional expectation $\Upsilon_0(x, \mathbf{w}) = \mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$ and its derivatives with respect to x, where $\mathbf{w}$ is a user-specified value of control variables at which $\Upsilon_0(x, \cdot)$ is evaluated, such as $\mathbf{w} = \mathbf{0}$, $\mathbb{E}[\mathbf{w}_i]$, or median$(\mathbf{w}_i)$, where $\mathbf{0}$ denotes a vector of zeros and median$(\mathbf{w}_i)$ denotes the population median of each component in $\mathbf{w}_i$. In the paper attention is restricted to $\Upsilon_0(x) = \Upsilon_0(x, \mathbb{E}[\mathbf{w}_i])$. In this section we provide a detailed discussion regarding the role of the evaluation point $\mathbf{w}$, which may be important for interpretation and for numerical results, and even for the visualization itself.

One might expect that since the additional controls are modeled as additively linear, the evaluation point $\mathbf{w}$ (and the coefficient $\boldsymbol{\gamma}_0$) should not impact conclusions about the nonparametric relationship between y and x. But this intuition overlooks the fact that the function $\mu_0(x)$ is only defined relative to how $\mathbf{w}_i$ is coded. We will show that the results of parametric specification tests and confidence bands for the mean function $\Upsilon_0(x, \mathbf{w})$ might be sensitive to the choice of $\mathbf{w}$, and how this issue may be circumvented by focusing instead on the derivative of the mean function, highlighting the importance of our theoretical contributions which can accommodate the estimation of derivatives.

Let us first consider the hypothesis testing procedure behind the informal practice of checking if the "dots" are roughly linear, and then running ordinary least squares regression of y_i on x_i and $\mathbf{w}_i$. This idea motivates the standard practice of plotting the fitted regression line along with the binned scatter plot, as in Figures 1 and 2 in the paper. In this case, the null hypothesis is *not* merely that $\mu_0(x) = \theta_0 + \theta_1 x$, i.e., a linear function, but rather that the full model is linear, so that $\Upsilon_0(x,\mathbf{w}) = \theta_0 + x\theta_1 + \mathbf{w}'\boldsymbol{\gamma}_0$. Under the partially linear assumption of the model (SA-2.2) below, these would seem identical, because in either case $\mathbf{w}$ enters linearly. But this is not so in practice for two reasons: the estimates of the coefficients $\boldsymbol{\gamma}_0$ will differ in general, as will the implied intercepts, and the chosen $\mathbf{w}$ will impact the uncertainty about the estimate of θ_0.

In a standard binscatter plot such as Figure 1 in the paper, the "dots" show the semiparametric estimate $\widehat{\Upsilon}(x,\widehat{\mathbf{w}}) = \widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}$, defined in (SA-2.3) below, while the plotted line is the parametric fit $\widetilde{\theta}_0 + x\widetilde{\theta}_1 + \widehat{\mathbf{w}}'\widetilde{\boldsymbol{\gamma}}$, obtained from least squares regression. Thus, while we are only interested in assessing the linearity of $\mu_0(x)$, we are *actually* testing these two functional forms for $\Upsilon_0(x,\mathbf{w})$, and the fact that $\widehat{\boldsymbol{\gamma}} \neq \widetilde{\boldsymbol{\gamma}}$ becomes important. Moreover, because $\widetilde{\theta}_0 + x\widetilde{\theta}_1 + \widehat{\mathbf{w}}'\widetilde{\boldsymbol{\gamma}}$ is a global parametric fit while $\widehat{\Upsilon}(x,\widehat{\mathbf{w}}) = \widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}$ is local and nonparametric, the implied intercept when plotted depends on the chosen $\widehat{\mathbf{w}}$, and this can shift the line away from the dots. Figure SA-1 demonstrates this by example: everything is identical between the three plots except for choice of $\widehat{\mathbf{w}}$. Notice the shift in absolute position (note the y axis) and the change in the relative position of the line and the binscatter. This phenomenon is unavoidable in this setting, and the user must select $\widehat{\mathbf{w}}$ appropriately. (Note that this does not occur when using the incorrect residualization because the covariates are mishandled.)

Figure SA-1: **Role of the Evaluation Point.** This figure demonstrates that the choice of $\widehat{\mathbf{w}}$ shifts both the absolute position (note the y axis) of the visualization and estimator, but also affects the comparison to parametric fits. The data is the same as in Figure 2 in the paper except that state and year fixed effects are omitted for simplicity.


(a) $\widehat{\mathbf{w}} = \mathbf{w}_{min}$ (b) $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$ (c) $\widehat{\mathbf{w}} = \mathbf{w}_{max}$

Beyond the visual inspection of a plot like Figure SA-1, we can also consider a formal test for the hypothesis $\Upsilon_0(x,\mathbf{w}) = M(x,\mathbf{w};\boldsymbol{\theta},\boldsymbol{\gamma}_0) = m(x;\boldsymbol{\theta}) + \mathbf{w}'\boldsymbol{\gamma}_0$. (In the case of linearity, $\boldsymbol{\theta} = (\theta_0,\theta_1)'$ and $m(x;\boldsymbol{\theta}) = \theta_0 + x\theta_1$.) This is a special case of the specification tests discussed in Section SA-3.7:

$$\dot{\mathsf{H}}_0: \quad \sup_{x\in\mathcal{X}} \Big|\Upsilon_0(x,\mathbf{w}) - M(x,\mathbf{w};\boldsymbol{\theta},\boldsymbol{\gamma}_0)\Big| = 0, \qquad \text{for some } \boldsymbol{\theta}, \qquad vs.$$

$$\dot{\mathsf{H}}_{\mathrm{A}}: \quad \sup_{x \in \mathcal{X}} \Big| \Upsilon_0(x, \mathbf{w}) - M(x, \mathbf{w}; \boldsymbol{\theta}, \boldsymbol{\gamma}_0) \Big| > 0, \qquad \text{for all } \boldsymbol{\theta}.$$

One rejects $\dot{\mathsf{H}}_0$ if and only if $\sup_{x \in \mathcal{X}} |\dot{T}_p(x)| \geq \mathfrak{c}$ for some critical value $\mathfrak{c}$ where $\dot{T}_p(x) = \frac{\widehat{\Upsilon}(x, \widehat{\mathbf{w}}) - M(x, \widehat{\mathbf{w}}; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}}$.

This testing procedure formalizes the idea of visually examining a binned scatter plot compared to a parametric specification; a common step before regression analysis. But it also formalizes the problematic dependency on the evaluation point $\mathbf{w}$ and the difference between $\widehat{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$. Despite the fact that $\mathbf{w}'\boldsymbol{\gamma}_0$ cancels out in both the null and alternative statements, the numerator of the t-statistic depends on $\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}})$, because in finite samples $\boldsymbol{\gamma}_0$ is unknown. Therefore our uncertainty about how x enters the model depends on the controls $\mathbf{w}_i$. As mentioned above, this comes about because $\mu_0(x)$ is only defined relative to $\mathbf{w}_i$.

Consider the case where $\mathbf{w}_i$ is an indicator (or fixed effect). Then setting $\widehat{\mathbf{w}} = \mathbf{0}$ would seem to remove the problem, because the numerator of $\dot{T}_p(x)$ depends only on $\widehat{\mu}(x)$ and $m(x; \widetilde{\boldsymbol{\theta}})$, while setting $\widehat{\mathbf{w}} = \mathbf{1}$ maximizes it. This is correct, but is then sensitive to how the researcher has coded $\mathbf{w}_i$, i.e., which category is considered the baseline. Thus we can get a different answer to the test depending on which category of $\mathbf{w}$ we consider, even though the hypothesis applies to both. This is intuitively the same as the fact that in a linear model with dummy variables the standard error of the intercept changes depending on how $\mathbf{w}$ is coded. The case of a continuous $\mathbf{w}_i$ (especially with large support, such as annual income) is perhaps worse: if $\widehat{\boldsymbol{\gamma}} \neq \widetilde{\boldsymbol{\gamma}}$, then there is *always* some value $\widehat{\mathbf{w}}$ for which we reject the null. Thus, using the procedure described above to test parametric specifications is potentially confusing at best, and at worst is vulnerable to p-hacking. It is worth noting that in most papers studying the partially linear model, the parameter of interest is $\boldsymbol{\gamma}_0$, and so these concerns have gone largely unnoticed. (And are masked by construction when using the incorrect residualization approach.)

To avoid these issues, and motivated by the fact that the central point of binscatter is to study how y_i relates to x_i, controlling for $\mathbf{w}_i$, we advocate reformulating the hypothesis as pertaining to the *derivative* of $\mu_0(x)$, instead of the level. Under the partially linear model maintained throughout, any derivative of $\mathbb{E}[y_i | x_i = x, \mathbf{w}_i = \mathbf{w}]$ is exactly $\mu_0^{(v)}(x)$, and is by definition $\Upsilon_0^{(v)}(x, \mathbf{w})$. Therefore, instead of testing the null $\Upsilon_0(x, \mathbf{w}) = m(x; \boldsymbol{\theta}) + \mathbf{w}'\boldsymbol{\gamma}_0$, we test the equivalent hypothesis that $\Upsilon_0^{(v)}(x, \mathbf{w}) = m^{(v)}(x; \boldsymbol{\theta})$ for some $v \geq 1$. For example, instead of testing that $\mu_0(x)$ is linear, we test that it has constant first derivative. To test if $\mu_0(x)$ itself is constant, the null would be that $\mu_0^{(1)}(x) = m^{(1)}(x;\boldsymbol{\theta}) = 0$.

Such (more robust) tests are still special cases of the specification tests discussed in Section SA-3.7: for some $v \geq 1$,

$$\begin{aligned}
\dot{\mathsf{H}}_0: &\quad \sup_{x\in\mathcal{X}} \left|\Upsilon_0^{(v)}(x,\mathbf{w}) - m^{(v)}(x;\boldsymbol{\theta})\right| = 0, \quad \text{for some } \boldsymbol{\theta}, \qquad vs. \\
\dot{\mathsf{H}}_\mathrm{A}: &\quad \sup_{x\in\mathcal{X}} \left|\Upsilon_0^{(v)}(x,\mathbf{w}) - m^{(v)}(x;\boldsymbol{\theta})\right| > 0, \quad \text{for all } \boldsymbol{\theta}.
\end{aligned}$$

One rejects $\dot{\mathsf{H}}_0$ if and only if $\sup_{x\in\mathcal{X}} |\dot{T}_p(x)| \geq \mathfrak{c}$ for some critical value $\mathfrak{c}$ where $\dot{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - m^{(v)}(x;\widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}$.

Finally, notice that the visual appearance of the confidence band for the mean function $\Upsilon_0(x,\mathbf{w}) = \mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$ will also be impacted by the evaluation point $\mathbf{w}$ (or its feasible version $\widehat{\mathbf{w}}$). This is important to keep in mind when evaluating binscatter plots. By definition, each binscatter plot shows only one choice of $\mathbf{w}$, and therefore while the shape of $\widehat{\Upsilon}(x,\widehat{\mathbf{w}})$ is unchanged, a level shift will occur and the size of the band can change. For an intuitive example, again consider the case where $\mathbf{w}$ is categorical, and some categories have much larger or smaller sample sizes. These different sample sizes will naturally be reflected in the uncertainty for $\Upsilon_0(x,\mathbf{w})$.

For this reason, we must be careful when using confidence bands as visual aids in parametric specification testing. If we plot $\widehat{\Upsilon}(x,\widehat{\mathbf{w}})$ and its associated confidence band, it is tempting to say that if this band does not contain a line (or quadratic function), then we say that at level α we reject the null hypothesis that $\mu_0(x)$ is linear (or quadratic). Although this is formally justified, we must interpret such analyses with caution because of the role of the evaluation point.

SA-2 General Setup and Notation

To present all our complete theoretical results we first review and generalize the notation introduced in the main text. 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. Define the following least squares estimand:

$$(\mu_0(\cdot), \boldsymbol{\gamma}_0) = \underset{\mu\in\mathcal{M},\boldsymbol{\gamma}\in\mathbb{R}^d}{\arg\min}\ \mathbb{E}\big[\big(y_i - \mu(x_i) - \mathbf{w}_i'\boldsymbol{\gamma}\big)^2\big], \qquad \text{(SA-2.1)}$$

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

We study binscatter estimators in the partially linear regression model:

$$y_i = \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0 + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|x_i, \mathbf{w}_i] = 0. \tag{SA-2.2}$$

The parameter of interest is

$$\Upsilon_0^{(v)}(x, \mathbf{w}) = \frac{\partial^v}{\partial x^v}\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}], \qquad v \in \mathbb{N}_0,$$

for some evaluation points x and $\mathbf{w}$. Given the assumption $\mathbb{E}[\epsilon_i|x_i, \mathbf{w}_i] = 0$ in (SA-2.2):

$$\Upsilon_0(x, \mathbf{w}) = \Upsilon_0^{(0)}(x, \mathbf{w}) = \mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0 \qquad \text{and} \qquad \Upsilon_0^{(v)}(x, \mathbf{w}) = \mu_0^{(v)}(x) \text{ for } v > 0.$$

In the paper, we focused on $\Upsilon_0^{(v)}(x) = \Upsilon_0^{(v)}(x, \mathbb{E}[\mathbf{w}_i])$, one special case of $\Upsilon_0^{(v)}(x, \mathbf{w})$ defined above for some evaluation point $\mathbf{w}$.

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

Assumption SA-DGP (Data Generating Process)**.** *$\{(y_i, x_i, \mathbf{w}_i') : 1 \leq i \leq n\}$ is i.i.d. satisfying (SA-2.1) with $\mathcal{X}$ a compact interval; x_i has a distribution function $F_X(x)$ with a Lipschitz continuous (Lebesgue) density $f_X(x)$ bounded away from zero on $\mathcal{X}$; and $\mu_0(x)$ is ς_μ-times continuously differentiable for some $\varsigma_\mu \geq p + 1$.*

We next impose a condition that is specific to the least squares binscatter. Binscatters in more general models are studied in Cattaneo et al. (2023). Section SA-2.1 defines standard notation.

Assumption SA-LS (Least Squares)**.**

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

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

Part (i) imposes some moment conditions on the error term which are commonly used in the nonparametric series estimation literature. Part (ii) includes a set of conditions similar to those used in Cattaneo, Jansson and Newey (2018a,b) to analyze the semiparametric partially linear regression model. They ensure the negligibility of the estimation error of $\widehat{\boldsymbol{\gamma}}$. To reduce notation, we use the same constant $\nu > 2$ in the conditional moment bounds for ϵ_i and $\mathbf{w}_i$.

Binscatter estimators are typically constructed based on quantile-spaced partitions, and a major innovation herein is accounting for this additional randomness. Our results allow for other options as well, including evenly spaced partitioning. Specifically, the relevant support of x_i is partitioned into J disjoint intervals employing the empirical quantiles, leading to the partitioning scheme $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \widehat{\mathcal{B}}_2, \ldots, \widehat{\mathcal{B}}_J\}$, where

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

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

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

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

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

$$\sqrt{J} \times \mathbb{1}_{\widehat{\mathcal{B}}_{\bar{j}}}(x) \times \Big(\frac{x - x_{(\lfloor (\bar{j}-1)n/J \rfloor)}}{\hat{h}_{\bar{j}}}\Big)^{j-1-(\bar{j}-1)(p+1)}, \quad j = 1, \cdots, (p+1)J,$$

where $\bar{j} = \lceil j/(p+1) \rceil$, $\lceil \cdot \rceil$ is the ceiling operator, and $\hat{h}_{\bar{j}} = x_{(\lfloor \bar{j}n/J \rfloor)} - x_{(\lfloor (\bar{j}-1)n/J \rfloor)}$. Thus, each local polynomial is centered at the start of each bin and scaled by the length of the bin. $\sqrt{J}$ is an additional scaling factor which 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 $1 \leq s \leq p$, we introduce a new basis

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

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 to a smoothed binscatter basis. When $s = 0$, we let $\widehat{\mathbf{T}}_0 = \mathbf{I}_{(p+1)J}$, the identity matrix of dimension $(p+1)J$. Thus $\widehat{\mathbf{b}}_{p,0}(x)$ is the discontinuous basis without any constraints defined previously. 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. We require $s \leq p$, since if $s = p+1$, $\widehat{\mathbf{b}}_{p,s}(x)$ reduces to a global polynomial basis of degree p.

A key feature of the transformation matrix $\widehat{\mathbf{T}}_s$ is that on every row it has *at most* $(p+1)^2$ nonzeros, and on every column it has *at most* $p+1$ nonzeros. The expression of these elements is cumbersome. The proof of Lemma SA-3.2 describes the structure of $\widehat{\mathbf{T}}_s$ in more detail and provides an explicit representation for $\widehat{\mathbf{T}}_s$.

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

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} \Big(y_i - \widehat{\mathbf{b}}_{p,s}(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma}\Big)^2, \tag{SA-2.3}$$

where $\widehat{\mathbf{b}}^{(v)}_{p,s}(x) = \frac{d^v}{dx^v}\widehat{\mathbf{b}}_{p,s}(x)$ for some $v \in \mathbb{Z}_+$ such that $v \leq p$. It is well known that this estimator admits the following "backfitting" expression, which will be convenient for later theoretical analysis:

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

where $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\mathbf{B} = (\widehat{\mathbf{b}}_{p,s}(x_1), \ldots, \widehat{\mathbf{b}}_{p,s}(x_n))'$, $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_n)'$ and $\mathbf{M}_{\mathbf{B}} = \mathbf{I}_n - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ with $\mathbf{I}_n$ denoting the identify matrix of size n.

Given an estimator $\widehat{\mathbf{w}}$ of the evaluation point $\mathbf{w}$, we have the following estimator of $\Upsilon_0^{(v)}(x,\mathbf{w})$:

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

Throughout the supplement (and the paper), we always assume that the estimator $\widehat{\mathbf{w}}$ is either nonrandom (e.g., a fixed value) or generated based on $\mathbf{W}$.

Remark SA-2.1 (Smoothness and Bias Correction)**.** We remind readers that 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 paper only uses the binscatter basis with $s = p$, where $p = 0$ for binscatter estimation and $p = 1$ for inference. In addition, in the paper we let J be the IMSE-optimal choice corresponding to $p = \mathtt{p}$ for a fixed number $\mathtt{p}$ (see Theorem SA-3.4), and inference is conducted based on the binscatter basis of degree $p = \mathtt{p} + 1$. In particular, we set $\mathtt{p} = 0$ to construct confidence bands in Section III. This can be viewed as a bias correction strategy (Calonico, Cattaneo and Farrell, 2018, 2022) which guarantees the smoothing bias of the binscatter estimator is negligible in inference under mild conditions. ⌟

SA-2.1 Notation

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

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 $\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$ 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 $g^{(v)}(z) = d^v g(z)/dz^v$ 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. $l_n \asymp m_n$ 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 $N(\mathcal{G}, L_2(\mathbb{Q}), \varepsilon\|\bar{G}\|_{\mathbb{Q},2})$.

Partitions. Given the random partition $\widehat{\Delta}$, we use the notation $\mathbb{E}_{\widehat{\Delta}}[\cdot]$ to denote that the expectation is taken with the partition $\widehat{\Delta}$ understood as fixed. To further simplify notation, we let $\{\hat{\tau}_0 \leq \hat{\tau}_1 \leq \cdots \leq \hat{\tau}_J\}$ denote the empirical quantile sequence employed by $\widehat{\Delta}$ and $\hat{h}_j = \hat{\tau}_j - \hat{\tau}_{j-1}$ be the width of the j-th bin $\widehat{\mathcal{B}}_j$. Accordingly, let $\{\tau_0 \leq \cdots \leq \tau_J\}$ be the population quantile sequence, i.e., $\tau_j = F_X^{-1}(j/J)$ for $0 \leq j \leq J$. Then $\Delta_0 = \{\mathcal{B}_1, \ldots, \mathcal{B}_J\}$ denotes the partition based on population quantiles, i.e.,

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

Let $h_j = F_X^{-1}(j/J) - F_X^{-1}((j-1)/J)$ 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 $\mathbf{b}_{p,s}(x) = \mathbf{b}_{p,s}(x; \Delta_0)$.

For any given partition Δ, the *population* least squares projection of $\mu_0(\cdot)$ is given by $\mathbf{b}_{p,s}(\cdot;\Delta)'\boldsymbol{\beta}_0(\Delta)$

with

$$\boldsymbol{\beta}_0(\Delta) := \underset{\boldsymbol{\beta}\in\mathbb{R}^{K_{p,s}}}{\arg\min}\ \mathbb{E}[(\mu_0(x_i) - \mathbf{b}_{p,s}(x_i;\Delta)'\boldsymbol{\beta})^2]. \tag{SA-2.4}$$

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

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

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

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

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

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

Other. Let $\mathbf{X} = [x_1,\ldots,x_n]'$, $\mathbf{W} = [\mathbf{w}_1,\cdots,\mathbf{w}_n]'$, and $\mathbf{D} = [(y_i, x_i, \mathbf{w}_i')' : i = 1,2,\ldots,n]$. $\lceil z\rceil$ 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-3 Theoretical Results

Our main theoretical results are presented in this section. We will focus on the estimator $\widehat{\Upsilon}^{(v)}(x,\widehat{\mathbf{w}})$ of $\Upsilon_0^{(v)}(x,\mathbf{w})$. The estimator $\widehat{\Upsilon}(x)$ of $\Upsilon_0^{(v)}(x) = \Upsilon_0^{(v)}(x,\mathbb{E}[\mathbf{w}_i])$ discussed in the paper is covered as a special case.

SA-3.1 Properties of Quantile-Based Partition and Binscatter Basis

In this section we first give some preliminary lemmas concerning the basic properties of the quantile-based partition and the binscatter basis, which are necessary for our main analysis and may be of independent interest.

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

$\hat{h}_j = h_j(\widehat{\Delta})$ and $h_j = h_j(\Delta_0)$. Then, we introduce the family of partitions:

$$\Pi = \Big\{\Delta : \frac{\max_{1\leq j\leq J} h_j(\Delta)}{\min_{1\leq j\leq J} h_j(\Delta)} \leq \frac{3\bar{f}_X}{\underline{f}_X}\Big\}. \tag{SA-3.1}$$

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

Lemma SA-3.1 (Quasi-Uniformity of Quantile-Spaced Partitions)**.** *Suppose that Assumption SA-DGP holds. If $\frac{J\log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then (i) $\max_{1\leq j\leq J}|\hat{h}_j - h_j| \lesssim_{\mathbb{P}} J^{-1}\Big(\frac{J\log J}{n}\Big)^{1/2}$, and (ii) $\widehat{\Delta} \in \Pi$ w.p.a. 1.*

As discussed previously, $\widehat{\mathbf{T}}_s$ links the more complex spline basis with a simple piecewise polynomial basis. Recall that $\widehat{\mathbf{T}}_s = \widehat{\mathbf{T}}_s(\widehat{\Delta})$ depends on the empirical-quantile-based partition $\widehat{\Delta}$. The next lemma describes its key features. We let $\mathbf{T}_s := \mathbf{T}_s(\Delta_0)$ be the transformation matrix corresponding to the nonrandom basis $\mathbf{b}_{p,s}(x)$, i.e., $\mathbf{b}_{p,s}(x) = \mathbf{T}_s\mathbf{b}_{p,0}(x)$.

Lemma SA-3.2 (Transformation Matrix)**.** *Suppose that Assumption SA-DGP holds. If $\frac{J\log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then $\widehat{\mathbf{b}}_{p,s}(x) = \widehat{\mathbf{T}}_s\widehat{\mathbf{b}}_{p,0}(x)$ with $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$, $\|\widehat{\mathbf{T}}_s\| \lesssim_{\mathbb{P}} 1$, $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\|_\infty \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n}\Big)^{1/2}$, and $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n}\Big)^{1/2}$.*

The following lemma provides some simple bounds on the basis.

Lemma SA-3.3 (Local Basis)**.** *Suppose that Assumption SA-DGP holds. Then, $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}^{(v)}_{p,s}(x)\|_0 \leq (p+1)^2$. If, in addition, $\frac{J\log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}^{(v)}_{p,s}(x)\| \lesssim_{\mathbb{P}} J^{\frac{1}{2}+v}$.*

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

Lemma SA-3.4 (Approximation Error)**.** *Suppose that Assumption SA-DGP holds. If $\frac{J\log J}{n} = o(1)$ and $\frac{\log n}{J} = o(1)$, then*

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

Remark SA-3.1 (Improvements over literature)**.** Lemmas SA-3.1–SA-3.4 show some basic characteristics of the binscatter basis, which are used in the subsequent main analysis. Compared with other studies of splines (see, e.g., Shen, Wolfe and Zhou, 1998; Huang, 2003; Schumaker, 2007), we formally take into account the randomness of the partition formed by empirical quantiles. $\lrcorner$

SA-3.2 Preliminary Technical Lemmas

This section collects a set of technical lemmas, which are key ingredients of our main theorems.

We first introduce the following quantities that will be frequently used:

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

where $\widehat{\epsilon}_i = y_i - \widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}}$. All quantities with $\widehat{\ }$ or $\bar{\ }$ depend on the random partition $\widehat{\Delta}$, and those without any accents are nonrandom with the only exception of $\Omega(x)$, where the basis $\widehat{\mathbf{b}}^{(v)}_{p,s}(x)$ still depends on $\widehat{\Delta}$. The dependence on p, s and v is often omitted for simplicity.

The following lemma characterizes the properties of the Gram matrix of the binscatter basis.

Lemma SA-3.5 (Gram)**.** *Suppose that Assumption SA-DGP holds. Then,* $1 \lesssim \lambda_{\min}(\mathbf{Q}_0) \leq \lambda_{\max}(\mathbf{Q}_0) \lesssim 1$. *If, in addition,* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

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

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

Lemma SA-3.6 (Asymptotic Variance)**.** *Suppose that Assumptions SA-DGP and SA-LS(i) hold. If* $\frac{J\log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then w.p.a. 1,*

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

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

Lemma SA-3.7 (Uniform Convergence: Variance)**.** *Suppose that Assumptions SA-DGP and SA-LS(i) hold. If* $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n}=o(1)$ *and* $\frac{\log n}{J}=o(1)$*, then*

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

As explained before, $\widehat{r}_0(x)$ is understood as the L_2 approximation error of least squares estimators for $\mu_0(x)$. The next lemma establishes the bound on the projection of $\widehat{r}_0(x)$ onto the space spanned by $\widehat{\mathbf{b}}_{p,s}(x)$ in terms of sup-norm.

Lemma SA-3.8 (Projection of Approximation Error)**.** *Under Assumption SA-DGP, if* $\frac{J\log J}{n}=o(1)$ *and* $\frac{\log n}{J}=o(1)$*, then*

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

The last lemma in this subsection characterizes the convergence of the parametric component in the expression of $\widehat{\boldsymbol{\beta}}$.

Lemma SA-3.9 (Covariate Adjustment)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. If* $\frac{J\log J}{n}=o(1)$ *and* $\frac{\log n}{J}=o(1)$*, then*

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

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

Let $(a_n:n\geq 1)$ be a sequence of non-vanishing constants, which will be used later to characterize the strong approximation rate. Lemma SA-3.9 implies that if $\frac{a_n}{\sqrt{J}}=o(1)$ and $a_n\sqrt{n}J^{-p-(\varsigma_w\wedge(p+1))-\frac{3}{2}}=o(1)$, then we have

$$\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|=o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n}).$$

This result suffices to make the estimation error of $\widehat{\boldsymbol{\gamma}}$ negligible in the large sample inference on $\mu_0^{(v)}(\cdot)$ or $\Upsilon_0(\cdot,\mathbf{w})$.

Remark SA-3.2 (Improvements over literature)**.** The results in this subsection give novel rates of approximations for semi-linear partitioning-based estimators with random partitions. Compared to standard semi-linear regression results, our results provide sharper approximation rates due to the specific binscatter basis, and also formally take into account the randomness of the partition formed by empirical quantiles. See Cattaneo, Jansson and Newey (2018a,b), and reference therein, for related literature. ⌟

SA-3.3 Stochastic Linear Approximation and Point Estimation

Theorem SA-3.1 (Stochastic Linear Approximation)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. If* $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} \lesssim 1$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\sup_{x\in\mathcal{X}}\Big|\widehat{\Upsilon}^{(v)}(x,\widehat{\mathbf{w}}) - \Upsilon_0^{(v)}(x,\mathbf{w}) - \widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\epsilon_i]\Big|$$
$$\lesssim_{\mathbb{P}} J^v\Big(\frac{1}{\sqrt{n}} + J^{-p-1-(\varsigma_w\wedge(p+1))} + J^{-p-1}\Big) + \|\widehat{\mathbf{w}}-\mathbf{w}\|\mathbb{1}(v=0).$$

An immediate corollary of Theorem SA-3.1 is the uniform convergence of $\widehat{\Upsilon}^{(v)}(\cdot,\widehat{\mathbf{w}})$.

Corollary SA-3.1 (Uniform Convergence)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. If* $\sqrt{n}J^{-p-(\varsigma_w\wedge(p+1))-\frac{3}{2}} = o(1)$ *and* $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} \lesssim 1$*, then*

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

If, in addition, $\|\widehat{\mathbf{w}}-\mathbf{w}\| \lesssim_{\mathbb{P}} \sqrt{\frac{J\log J}{n}} + J^{-p-1}$*, then*

$$\sup_{x\in\mathcal{X}}\Big|\widehat{\Upsilon}^{(0)}(x,\widehat{\mathbf{w}}) - \Upsilon^{(0)}(x,\mathbf{w})\Big| \lesssim_{\mathbb{P}} \Big(\frac{J\log J}{n}\Big)^{1/2} + J^{-p-1}.$$

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

Theorem SA-3.2 (Variance Estimate)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. If* $\frac{J^{\frac{\nu}{\nu-2}}(\log J)^{\frac{\nu}{\nu-2}}}{n} = o(1)$ *and* $\sqrt{n}J^{-p-(\varsigma_w\wedge(p+1))-\frac{3}{2}} = o(1)$*, then*

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

Remark SA-3.3 (Improvements over literature)**.** The results in this subsection improve on the linear series estimation literature (Belloni, Chernozhukov, Chetverikov and Kato, 2015; Cattaneo, Farrell and Feng, 2020) by formally taking into account the randomness of the partition formed by empirical quantiles, and by accounting for the semi-linear regression estimation structure. The final approximation rate in the Bahadur-type (linear) approximation is sharp for the binscatter basis (with or without random binning). $\lrcorner$

SA-3.4 Pointwise Distributional Approximation and Inference

In this subsection we focus on the pointwise inference on the unknown parameter $\Upsilon_0^{(v)}(x, \mathbf{w}) = \frac{\partial^v}{\partial x^v}\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$ and construct the t-statistic based on $\widehat{\Upsilon}^{(v)}(x, \widehat{\mathbf{w}})$:

$$T_p(x) = \frac{\widehat{\Upsilon}^{(v)}(x, \widehat{\mathbf{w}}) - \Upsilon_0^{(v)}(x, \mathbf{w})}{\sqrt{\widehat{\Omega}(x)/n}}.$$

Recall in our semi-linear model $\widehat{\Upsilon}^{(v)}(x, \widehat{\mathbf{w}})$ differs from $\widehat{\mu}^{(v)}(x)$ only when $v = 0$ and $\widehat{\mathbf{w}} \neq 0$. Therefore, the condition that $\widehat{\mathbf{w}}$ converges to $\mathbf{w}$ at a fast rate imposed below is needed only when $v = 0$.

Let $\Phi(\cdot)$ be the cumulative distribution function of a standard normal random variable. The following theorem constructs the pointwise inference for $\Upsilon_0^{(v)}(x, \mathbf{w})$.

Theorem SA-3.3 (Pointwise Asymptotic Distribution)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. If* $\sup_{x\in\mathcal{X}} \mathbb{E}[|\epsilon_i|^\nu|x_i = x] \lesssim 1$ *for some* $\nu \geq 3$, $\frac{J^{\frac{\nu}{\nu-2}}(\log J)^{\frac{\nu}{\nu-2}}}{n} = o(1)$, $nJ^{-2p-3} = o(1)$ *and* $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o(\sqrt{J/n})$, *then*

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

and accordingly,

$$\mathbb{P}\Big[\Upsilon_0^{(v)}(x, \mathbf{w}) \in \widehat{I}_p(x)\Big] = 1 - \alpha + o(1), \quad \textit{for each } x \in \mathcal{X},$$

where $\widehat{I}_p(x) = [\widehat{\Upsilon}^{(v)}(x, \widehat{\mathbf{w}}) \pm \mathfrak{c}\sqrt{\widehat{\Omega}(x)/n}]$ *and* $\mathfrak{c} = \Phi^{-1}(1 - \alpha/2)$.

Remark SA-3.4 (Robust Bias Correction)**.** In practice, we suggest employing the robust bias correction method (Calonico, Cattaneo and Farrell, 2018, 2022) to construct valid confidence intervals. Specifically, for a given p, let J be the corresponding IMSE-optimal choice $J_{\texttt{IMSE}}$ (see Section SA-4 for implementation details). By Theorem SA-3.4 and Remark SA-3.7 below, $J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$ in general. Then, construct the confidence intervals $\widehat{I}_{p+q}(x)$ (i.e., use $(p+q)$th-order binscatter estimator). This particular choice of $J = J_{\texttt{IMSE}}$ satisfies $nJ^{-2p-2q-3} = o(1)$ and $\frac{J^2 \log^2 J}{n} = o(1)$. Then, the conclusion of Theorem SA-3.3 immediately applies to $\widehat{I}_{p+q}(x)$ if $\nu = 4$ and $\varsigma_\mu = \varsigma_w = p+q+1$. $\lrcorner$

Remark SA-3.5 (Improvements over literature)**.** The results in this subsection improve upon Cattaneo, Farrell and Feng (2020, Section 5), the best results available for partitioning-based estimation, by formally taking into account the randomness of the partition formed by empirical quantiles, and by accounting for the semi-linear regression estimation structure. $\lrcorner$

SA-3.5 Integrated Mean Squared Error

Theorem SA-3.4 (IMSE)**.** *Suppose that Assumptions SA-DGP and SA-LS hold. Let $\omega(x)$ be a continuous weighting function over $\mathcal{X}$ bounded away from zero. If $\sqrt{n}J^{-p-(\varsigma_w \wedge (p+1))-\frac{3}{2}} = o(1)$, $\frac{J\log J}{n} = o(1)$ and $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$, then*

$$\int_{\mathcal{X}} \mathbb{E}\Big[\Big(\widehat{\Upsilon}^{(v)}(x,\widehat{\mathbf{w}}) - \Upsilon_0^{(v)}(x,\mathbf{w})\Big)^2 \Big|\mathbf{X},\mathbf{W}\Big]\omega(x)dx$$
$$= \frac{J^{1+2v}}{n}\mathscr{V}_n(p,s,v) + J^{-2(p+1-v)}\mathscr{B}_n(p,s,v) + o_{\mathbb{P}}\Big(\frac{J^{1+2v}}{n} + J^{-2(p+1-v)}\Big),$$

where

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

Remark SA-3.6 (Proof of Theorem 1)**.** Theorem 1 stated in the paper is a special case of Theorem SA-3.4. In Theorem 1 we let $s = p$ and $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$ and take $\omega(x)$ in Theorem SA-3.4 to be $f_X(x)$; Assumption 1 implies that Assumption SA-DGP holds with $\varsigma_\mu = p+2$, and Assumption SA-LS holds with $\nu = 4$ and $\varsigma_w = p+2$; and the rate condition $\sqrt{n}J^{-p-(\varsigma_w \wedge (p+1))-\frac{3}{2}} = o(1)$ in Theorem SA-3.4 is equivalent to $nJ^{-4p-5} = o(1)$. $\lrcorner$

As a consequence, the IMSE-optimal choice of J is $J_{\texttt{IMSE}} = J_{\texttt{IMSE}}(p,s,v) \asymp n^{\frac{1}{2p+3}}$ whenever $\mathscr{B}_n(p,s,v) \gtrsim 1$. See Remark SA-3.7 below for discussion of the lower bound on $\mathscr{B}_n(p,s,v)$. More precisely, if $\mathscr{B}_n(p,s,v) = \mathscr{B}(p,s,v) + o(1)$ and $\mathscr{V}_n(p,s,v) = \mathscr{V}(p,s,v) + o(1)$ for some constants $\mathscr{B}(p,s,v)$ and $\mathscr{V}(p,s,v)$, then we can take

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

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

Corollary SA-3.2. *Under the assumptions in Theorem SA-3.4,* $\mathscr{V}_n(p,0,v) = \mathscr{V}(p,0,v) + o(1)$ *and* $\mathscr{B}_n(p,0,v) = \mathscr{B}(p,0,v) + o(1)$ *where*

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

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

Remark SA-3.7. The above corollary implies that the bias constant $\mathscr{B}(p,0,v)$ is nonzero unless $\mu_0^{(p+1)}(x)$ is zero almost everywhere on $\mathcal{X}$. For other $s > 0$, notice that $\mathbf{b}_{p,s}^{(v)}(x)'\boldsymbol{\beta}_0$ can be viewed as an approximation of $\mu_0^{(v)}(x)$ in the space spanned by piecewise polynomials of order $(p-v)$. The best $L_2(x)$ approximation error in this space, according to the above corollary, is bounded away from zero if rescaled by J^{p+1-v}. $\mathbf{b}_{p,s}^{(v)}(x)'\boldsymbol{\beta}_0$, as a non-optimal L_2 approximation in such a space, must have a larger L_2 error than the best one (in terms of L_2-norm). Since $\omega(x)$ and $f_X(x)$ are both bounded and bounded away from zero, the above fact implies that except for the quite special case mentioned previously, $\mathscr{B}(p,s,v) \asymp 1$, a slightly stronger result than that in Theorem SA-3.4. We exclude this special case by assuming that the leading bias is non-degenerate, and thus $J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$. $\lrcorner$

Remark SA-3.8 (Improvements over literature)**.** The results in this subsection improve upon Cattaneo, Farrell and Feng (2020, Section 4), the best results available for partitioning-based estimation, by formally taking into account the randomness of the partition formed by empirical quantiles, and by accounting for the semi-linear regression estimation structure. ⌟

SA-3.6 Uniform Distributional Approximation

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

Theorem SA-3.5 (Strong Approximation)**.** *Suppose that Assumptions SA-DGP and SA-LS hold and $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n})$. If*

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

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

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

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

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

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

Theorem SA-3.6 (Plug-in Approximation)**.** *Suppose that the conditions in Theorem SA-3.5 hold. 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* $\xi > 0$*,*

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

Remark SA-3.9 (Proof of Theorem 2)**.** Theorem 2 in the paper is a special case of Theorems SA-3.5 and SA-3.6. In Theorem 2 we let $s = p$ and $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$; Assumption 1 imposed in the paper implies that Assumption SA-DGP holds with $\varsigma_\mu = p+2$ and Assumption SA-LS holds with $\varsigma_w = p+2$ and $\nu = 4$. Therefore, the desired strong approximation for $\widehat{\Upsilon}^{(v)}(x, \widehat{\mathbf{w}})$ follows from Theorem SA-3.5 and Theorem SA-3.6. For ease of presentation, Theorem 2 in the paper defines

$$Z_p(x) = \frac{\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0^{1/2}}{\sqrt{\Omega(x)}}\mathbf{N}_{K_{p,s}} = \frac{\widehat{\mathbf{b}}^{(v)}_{p,0}(x)'\widehat{\mathbf{T}}_s\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0^{1/2}}{\sqrt{\Omega(x)}}\mathbf{N}_{K_{p,s}}.$$

That is, we replace $\mathbf{T}_s$ in Theorem SA-3.5 with $\widehat{\mathbf{T}}_s$. As shown in the proof of Theorem SA-3.5 (see Step 3 therein), this does not affect the strong approximation result. ⌟

Remark SA-3.10 (Improvements over literature)**.** Theorems SA-3.5 and SA-3.6 offer a new easy-to-implement approach to conduct binscatter-based uniform distributional approximation and inference. We formally take into account the randomness of the empirical-quantile-based partition and approximate the *whole* t-statistic process by a (conditional) Gaussian process under seemingly minimal rate conditions. In fact, it can be shown that when $a_n = \sqrt{\log n}$ and a subexponential moment restriction holds for the error term, it suffices that $J/n = o(1)$, up to $\log n$ terms. In contrast, a strong approximation of the t-statistic process for general series estimators was obtained based on Yurinskii coupling in Belloni, Chernozhukov, Chetverikov and Kato (2015), which requires $J^5/n = o(1)$, up to $\log n$ terms. Alternatively, a strong approximation of the *supremum* of the t-statistic process can be obtained under weaker rate restrictions. For instance, Chernozhukov, Chetverikov and Kato (2014a) requires $J/n^{1-2/\nu} = o(1)$, up to $\log n$ terms, a result that applies exclusively to the suprema of the stochastic process. ⌟

Theorems SA-3.5 and SA-3.6 offer a way to approximate the distribution of the *whole* t-statistic process based on $\widehat{\Upsilon}^{(v)}(\cdot, \widehat{\mathbf{w}})$. One direct application of these results is to approximate the supremum of the t-statistic process. The following theorem shows that our strong approximation results can be used to obtain the convergence of the Kolmogorov distance between the distributions of $\sup_{x\in\mathcal{X}}|T_p(x)|$ and its (conditionally) Gaussian analogue $\sup_{x\in\mathcal{X}}|\widehat{Z}_p(x)|$.

Theorem SA-3.7 (Supremum Approximation)**.** *Let $a_n=\sqrt{\log J}$. Suppose that the conditions of Theorem SA-3.5 hold. Then,*

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

SA-3.7 Uniform Inference

One important application of the strong approximation results in Theorems SA-3.5 and SA-3.6 is to construct uniform confidence bands. Let $\widehat{I}_p(x)=[\widehat{\Upsilon}^{(v)}(x,\widehat{\mathbf{w}})\pm\mathfrak{c}\sqrt{\widehat{\Omega}(x)/n}]$ for some critical value $\mathfrak{c}$ to be specified, which is constructed based on a certain choice of J and the pth-order binscatter basis.

Theorem SA-3.8. *Let $a_n=\sqrt{\log J}$. Suppose that the conditions in Theorem SA-3.5 hold. If $\mathfrak{c}=\inf\Big\{c\in\mathbb{R}_+:\mathbb{P}[\sup_{x\in\mathcal{X}}|\widehat{Z}_p(x)|\leq c\,|\mathbf{D}]\geq 1-\alpha\Big\}$, then*

$$\mathbb{P}\Big[\Upsilon_0^{(v)}(x,\mathbf{w})\in\widehat{I}_p(x),\ \textit{for all } x\in\mathcal{X}\Big]=1-\alpha+o(1).$$

Remark SA-3.11 (Robust Bias Correction)**.** In practice, we suggest employing the robust bias correction method to construct valid confidence bands. Specifically, for a given p, let J be the corresponding IMSE-optimal choice $J_{\texttt{IMSE}}$ (see Section SA-4 for implementation details). By Theorem SA-3.4 and Remark SA-3.7, $J_{\texttt{IMSE}}\asymp n^{\frac{1}{2p+3}}$ in general. Then, construct the confidence band $\widehat{I}_{p+q}(x)$ (i.e., use $(p+q)$th-order binscatter estimator). This particular choice of $J=J_{\texttt{IMSE}}$ satisfies

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

Then, the conclusion of Theorem SA-3.8 immediately applies to $\widehat{I}_{p+q}(x)$ if $\nu=4$ and $\varsigma_\mu=\varsigma_w=p+q+1$.

In the paper we considered one special case of such robust bias-corrected confidence band: let J be the IMSE-optimal choice corresponding to $p=s=v=0$, and construct the confidence band $\widehat{I}_1(x)$ (i.e., let $q=1$ in the above construction). ⌟

Remark SA-3.12. The above results construct valid uniform confidence bands for least squares binscatter estimators under mild rate restrictions. Specifically, when $\nu \geq 4$, we require $J^2/n = o(1)$, up to $\log n$ terms. By contrast, Belloni, Chernozhukov, Chetverikov and Kato (2015) considers general series-based least squares estimators, and Theorem 5.6 therein can construct confidence bands under similar rate restrictions, which relies on the strong approximation technique for the suprema of the stochastic process developed in Chernozhukov, Chetverikov and Kato (2014a). $\lrcorner$

Using our main theoretical results, we can also test parametric specifications of the unknown function $\Upsilon_0^{(v)}(x,\mathbf{w})$. Consider the following testing problem:

$$\begin{aligned} \dot{\mathsf{H}}_0: &\quad \sup_{x\in\mathcal{X}}\left|\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\mathbf{w};\boldsymbol{\theta},\boldsymbol{\gamma}_0)\right| = 0, \quad \text{for some } \boldsymbol{\theta}, \qquad \mathit{vs.} \\ \dot{\mathsf{H}}_\text{A}: &\quad \sup_{x\in\mathcal{X}}\left|\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\mathbf{w};\boldsymbol{\theta},\boldsymbol{\gamma}_0)\right| > 0, \quad \text{for all } \boldsymbol{\theta}. \end{aligned}$$

where $M(x,\mathbf{w};\boldsymbol{\theta},\boldsymbol{\gamma}_0) = m(x;\boldsymbol{\theta}) + \mathbf{w}'\boldsymbol{\gamma}_0$. This testing problem can be viewed as a two-sided test where the equality between two functions holds *uniformly* over $x \in \mathcal{X}$. We introduce $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\gamma}}$ as consistent estimators of $\boldsymbol{\theta}$ and and $\boldsymbol{\gamma}_0$ under $\dot{\mathsf{H}}_0$, and then consider the following test statistic:

$$\dot{T}_p(x) := \frac{\widehat{\Upsilon}^{(v)}(x,\widehat{\mathbf{w}}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}}.$$

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

Theorem SA-3.9 (Parametric Specification Tests)**.** *Let $a_n = \sqrt{\log J}$. Suppose that the conditions in Theorem SA-3.5 hold. Let $\mathfrak{c} = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}}|\widehat{Z}_p(x)| \leq c|\mathbf{D}] \geq 1-\alpha\}$.*

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

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

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

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

Remark SA-3.13 (Robust Bias Correction)**.** In practice, we suggest employing the robust bias correction method to conduct specification tests. Specifically, for a given p, let J be the corresponding IMSE-optimal choice $J_{\texttt{IMSE}}$ (see Section SA-4 for implementation details). By Theorem SA-3.4 and Remark SA-3.7, $J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$ in general. Then, construct the t-statistic $\dot{T}_{p+q}(x)$, (i.e., use $(p+q)$th-order binscatter estimator). This particular choice of $J = J_{\texttt{IMSE}}$ satisfies

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

Also, $\frac{J^{1+2v}(\log J)}{n} \asymp n^{-\frac{2p-2v+2}{2p+3}} \log n = o(1)$ since we always require $p \geq v$. Then, the conclusion of Theorem SA-3.9 immediately applies to the test based on $\dot{T}_{p+q}(x)$ if $\nu = 4$ and $\varsigma_\mu = \varsigma_w = p+q+1$. ⌟

Another application of our theoretical results is to test certain shape restrictions on the unknown $\Upsilon_0^{(v)}(x, \mathbf{w})$. To be specific, consider the following testing problem:

$$\ddot{\mathsf{H}}_0: \sup_{x\in\mathcal{X}} (\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})) \leq 0 \text{ for certain } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}} \quad \text{v.s.}$$
$$\ddot{\mathsf{H}}_\text{A}: \sup_{x\in\mathcal{X}} (\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})) > 0 \text{ for } \bar{\boldsymbol{\theta}} \text{ and } \bar{\boldsymbol{\gamma}},$$

which 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 $\ddot{\mathsf{H}}_0$ and $\ddot{\mathsf{H}}_\text{A}$, we fix $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\gamma}}$ to be the same values in the parameter space. We introduce $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\gamma}}$ as consistent estimators of $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\gamma}}$ under both $\ddot{\mathsf{H}}_0$ and $\ddot{\mathsf{H}}_\text{A}$, and then rely on the following test statistic:

$$\ddot{T}_p(x) := \frac{\widehat{\Upsilon}^{(v)}(x,\widehat{\mathbf{w}}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}}.$$

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

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

Theorem SA-3.10 (Shape Restriction Tests)**.** *Let $a_n = \sqrt{\log J}$. Suppose that the conditions in Theorem SA-3.5 hold. In addition,* $\sup_{x\in\mathcal{X}} |M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}}) - M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})| = o_{\mathbb{P}}\Big(\sqrt{\frac{J^{1+2v}}{n\log J}}\Big)$. *Let* $\mathfrak{c} = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[\sup_{x\in\mathcal{X}} \widehat{Z}_p(x) \leq c|\mathbf{D}] \geq 1-\alpha\}$.

Under $\ddot{\mathsf{H}}_0$*,*

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

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

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

Remark SA-3.14 (Robust Bias Correction)**.** In practice, we suggest employing the robust bias correction method to conduct shape restriction tests. Specifically, for a given p, let J be the corresponding IMSE-optimal choice $J_{\texttt{IMSE}}$ (see Section SA-4 for implementation details). By Theorem SA-3.4 and Remark SA-3.7, $J_{\texttt{IMSE}}\asymp n^{\frac{1}{2p+3}}$ in general. Then, construct the t-statistic $\ddot{T}_{p+q}(x)$, (i.e., use $(p+q)$th-order binscatter estimator). This particular choice of $J=J_{\texttt{IMSE}}$ satisfies

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

Also, $\frac{J^{1+2v}(\log J)}{n}\asymp n^{-\frac{2p-2v+2}{2p+3}}\log n=o(1)$ since we always require $p\geq v$. Then, the conclusion of Theorem SA-3.10 immediately applies to the test based on $\ddot{T}_{p+q}(x)$ if $\nu=4$ and $\varsigma_\mu=\varsigma_w=p+q+1$.

⌟

Remark SA-3.15 (Improvements over literature)**.** The results presented in this section improve on 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), thereby offering faster approximation rates under weaker side restrictions (c.f., Belloni, Chernozhukov, Chetverikov and Kato, 2015; Cattaneo, Farrell and Feng, 2020). Furthermore, relative to prior work, our results formally take into account the randomness of the partition formed by empirical quantiles, account for the semi-linear regression estimation structure, and consider an array of inference problems. In particular, the underlying approach to establish strong approximation and related distributional approximations for binscatter statistics may be of independent interest.

⌟

SA-4 Feasible Number of Bins Selector

We discuss the implementation details for data-driven selection of the number of bins, based on the integrated mean squared error expansion for least squares binscatter estimators (see Theorem SA-3.4 and Corollary SA-3.2). Thus, the selectors given below can provide a choice of J that is optimal in the IMSE sense.

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

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

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

SA-4.1 Rule-of-thumb Selector

A rule-of-thumb choice of J is obtained based on Corollary SA-3.2, in which case $s = 0$.

Regarding the variance constants $\mathscr{V}(p,0,v)$, the unknowns are the density function $f_X(x)$ and the conditional variance $\sigma^2(x)$. A Gaussian reference model is employed to get the estimate $\widehat{f}_X$ of $f_X(x)$. For the conditional variance, recall $\sigma^2(x_i, \mathbf{w}_i) = \mathbb{E}[y_i^2|x_i,\mathbf{w}_i] - (\mathbb{E}[y_i|x_i,\mathbf{w}_i])^2$, where the two conditional expectations can be approximated by global polynomial regressions of degree $p+1$. Let $\widehat{\sigma}^2(x_i,\mathbf{w}_i)$ denote the resulting estimate. Then, the variance constant is estimated by

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

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

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

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

SA-4.2 Direct-plug-in Selector

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

More generally, suppose that a preliminary choice $J_{\texttt{pre}}$ is given, and then a binscatter basis $\widehat{\mathbf{b}}_{p,s}(x)$ (of order p) can be constructed immediately on the preliminary partition. Implementing a binscatter regression using this basis and partitioning, we can obtain the variance constant estimate using a standard variance estimator, such as the one in Theorem SA-3.2.

Regarding the bias constant, we employ the uniform approximation (SA-5.6) in the proof of Theorem SA-3.4. The key idea of the bias representation is to "orthogonalize" the leading error of the uniform approximation based on splines with simple knots (i.e., p smoothness constraints are imposed) with respect to the preliminary binscatter basis $\widehat{\mathbf{b}}_{p,s}(x)$. Specifically, the key unknown in the expression of the leading error is $\mu_0^{(p+1)}(x)$, which can be estimated by implementing a binscatter regression of order $p+1$ (with the preliminary partition unchanged). Plug it in (SA-5.7), and all other quantities in that equation can be replaced by their sample analogues. Then, a bias constant estimate is available.

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

SA-5 Proofs

SA-5.1 Proof of Lemma SA-3.1

Proof. The first result follows by Lemma SA2 of Calonico, Cattaneo and Titiunik (2015). To show the second result, first consider the deterministic partition sequence Δ_0 based on the population quantiles. By the mean value theorem,

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

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

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

Then,

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

and the desired result follows. □

SA-5.2 Proof of Lemma SA-3.2

Proof. For $s=0$, the result is trivial. For $0<s\leq p$, $\widehat{\mathbf{b}}_{p,s}(x)$ is formally known as B-spline basis of order $p+1$ with knots $\{\hat{\tau}_1,\ldots,\hat{\tau}_{J-1}\}$ of multiplicities $(p-s+1,\ldots,p-s+1)$. See Schumaker (2007, Definition 4.1). Without loss of generality, suppose $\mathcal{X}=[0,1]$. Specifically, such a basis is constructed on an extended knot sequence $\{\xi_j\}_{j=1}^{2(p+1)+(p-s+1)(J-1)}$:

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

and

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

By the well-known Recursive Relation of Splines, a typical function $\widehat{b}_{p,s,\ell}(x)$ in $\widehat{\mathbf{b}}_{p,s}(x)$ supported on $(\xi_\ell,\xi_{\ell+p+1})$ is expressed as

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

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

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

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

We want to linearly represent $b_{p,s,\ell}(x)$ in terms of $\mathbf{b}_{p,0}(x)$ with typical element

$$\varphi_{j,\alpha}(x) = \sqrt{J} \cdot \mathbb{1}_{\widehat{\mathcal{B}}_j}(x)\Big(\frac{x - \hat{\tau}_{j-1}}{\hat{h}_j}\Big)^{\alpha}, \quad 0 \leq \alpha \leq p, \quad 1 \leq j \leq J. \qquad \text{(SA-5.1)}$$

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

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

The quantities within the summation only depend on distance between knots, which is no greater than $(p+1)\max_j \hat{h}_j$ since the support covers at most $(p+1)$ bins. Both denominator and numerator are products of p such distances, and hence by Lemma SA-3.1, $\sup_{j,\alpha} |c_{j,\alpha}| \lesssim_{\mathbb{P}} 1$. Then, $b_{p,s,\ell}(x)$ can be written as

$$b_{p,s,\ell}(x) = \sum_{j:\mathcal{B}_j \subset [\xi_\ell, \xi_{\ell+p+1}]} \sum_{\alpha=0}^{p} c_{j,\alpha}\psi_{j,\alpha}(x).$$

The above expression gives the elements of the ℓth row of $\widehat{\mathbf{T}}_s$.

Since each row and each column of $\widehat{\mathbf{T}}_s$ only contain a finite number of nonzeros, $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\| \lesssim_{\mathbb{P}} 1$. Using the fact $\max_{1\leq j\leq J} |\hat{h}_j - h_j| \lesssim_{\mathbb{P}} J^{-1}\sqrt{J \log J/n}$ given in the proof of Lemma SA-3.1, and noticing the form of $c_{j,\alpha}$, $\max_{k,l} |(\widehat{\mathbf{T}}_s - \mathbf{T}_s)_{k,l}| \lesssim \sqrt{J \log J/n}$ where $(\widehat{\mathbf{T}}_s - \mathbf{T}_s)_{k,l}$ is (k,l)th element of $\widehat{\mathbf{T}}_s - \mathbf{T}_s$. Since $(\widehat{\mathbf{T}}_s - \mathbf{T}_s)$ only has a finite number of nonzeros on every row and column, $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\|_\infty \lesssim_{\mathbb{P}} \sqrt{J \log J/n}$ and $\|\widehat{\mathbf{T}}_s - \mathbf{T}_s\| \lesssim_{\mathbb{P}} \sqrt{J \log J/n}$.

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

$$\widehat{b}_{p,p,\ell}(x) = (\hat{\tau}_{\ell+p+1} - \hat{\tau}_\ell) \sum_{j=\ell}^{\ell+p+1} \frac{(x-\hat{\tau}_j)_+^p}{\prod_{\substack{k=\ell \\ k\neq j}}^{\ell+p+1}(\hat{\tau}_k - \hat{\tau}_j)}, \tag{SA-5.2}$$

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

$$\widehat{b}_{p,p,\ell}(x) = \sum_{\alpha=0}^{p} \sum_{j=\ell+1}^{\ell+p+1} c_{j,\alpha}\varphi_{j,\alpha}(x), \quad \text{for some } c_{j,\alpha}. \tag{SA-5.3}$$

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

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

Thus, the contribution of $(x-\hat{\tau}_k)_+^p$ to the coefficients of $\varphi_{j,\alpha}(x)$ in Equation (SA-5.3), combined with its coefficient in Equation (SA-5.2), is

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

Collecting all such coefficients contributed by $(x-\hat{\tau}_k)_+^p$, $k = \ell, \dots, j-1$, we obtain

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

□

SA-5.3 Proof of Lemma SA-3.3

Proof. The sparsity of the basis follows by construction. To show the bound on $\|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)\|$, notice that when $s=0$, for any $x \in \mathcal{X}$ and any $j = 1, \dots, J(p+1)$, $0 \leq \widehat{b}_{p,0,j}(x) \leq \sqrt{J}$. Define $\varphi_{j,\alpha}(x)$ as in Equation (SA-5.1). Since

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

the bound on $\|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)\|$ simply follows from Lemma SA-3.1 and Lemma SA-3.2. □

SA-5.4 Proof of Lemma SA-3.4

Proof. By Lemma SA-3.1, it suffices to establish the approximation power of $\mathbf{b}_{p,s}(x;\Delta)$ for all $\Delta\in\Pi$. For $v=0$, by Theorem 6.27 of Schumaker (2007), $\max_{\Delta\in\Pi}\min_{\boldsymbol{\beta}\in\mathbb{R}^{K_{p,s}}}\sup_{x\in\mathcal{X}}|\mu_0(x)-\mathbf{b}_{p,s}(x;\Delta)'\boldsymbol{\beta}|\lesssim J^{-p-1}$. By Huang (2003) and Assumption SA-DGP, the Lebesgue factor of spline bases is bounded. Then, the bound on uniform approximation error coincides with that for L_2 projection error up to some universal constant.

For $v>0$, again, we only need to consider the case where Δ belongs to Π. For any $\Delta\in\Pi$, we can take the best L_∞-approximation: for some $\boldsymbol{\beta}_\infty(\Delta)\in\mathbb{R}^{K_{p,s}}$, $\|\mu_0(\cdot)-\mathbf{b}_{p,s}(\cdot;\Delta)'\boldsymbol{\beta}_\infty(\Delta)\|_\infty\lesssim J^{-p-1}$, and $\|\mu_0^{(v)}(\cdot)-\mathbf{b}_{p,s}^{(v)}(\cdot;\Delta)'\boldsymbol{\beta}_\infty(\Delta)\|_\infty\lesssim J^{-p-1+v}$. Such a construction exists by Lemma SA-6.1 of Cattaneo, Farrell and Feng (2020). Then, $\|\mu_0^{(v)}(\cdot)-\mathbf{b}_{p,s}^{(v)}(\cdot;\Delta)'\boldsymbol{\beta}_0(\Delta)\|_\infty\lesssim\|\mu_0^{(v)}(\cdot)-\mathbf{b}_{p,s}^{(v)}(\cdot;\Delta)'\boldsymbol{\beta}_\infty(\Delta)\|_\infty+\|\mathbf{b}_{p,s}^{(v)}(\cdot;\Delta)'(\boldsymbol{\beta}_\infty(\Delta)-\boldsymbol{\beta}_0(\Delta))\|_\infty\lesssim J^{-p-1+v}+\|\mathbf{b}_{p,s}^{(v)}(\cdot;\Delta)'(\boldsymbol{\beta}_\infty(\Delta)-\boldsymbol{\beta}_0(\Delta))\|_\infty$. By definition of $\boldsymbol{\beta}_0(\Delta)$,

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

where $r_\infty(x_i;\Delta)=\mu_0(x_i)-\mathbf{b}_{p,s}(x_i;\Delta)'\boldsymbol{\beta}_\infty(\Delta)$. By the argument given later in the proof of Lemma SA-3.5 in Section SA-3, we have $\|\mathbb{E}[\mathbf{b}_{p,s}(x_i;\Delta)\mathbf{b}_{p,s}(x_i;\Delta)']^{-1}\|_\infty\lesssim 1$ uniformly over $\Delta\in\Pi$. Since $\mathbf{b}_{p,s}(x_i;\Delta)$ is supported on a finite number of bins, $\|\mathbb{E}[\mathbf{b}_{p,s}(x_i;\Delta)r_\infty(x_i;\Delta)]\|_\infty\lesssim J^{-p-1-1/2}$. Then the desired result follows. □

SA-5.5 Proof of Lemma SA-3.5

Proof. The upper bound on the maximum eigenvalue of $\mathbf{Q}_0$ follows from Lemma SA-3.2 and the quasi-uniformity property of population quantiles shown in the proof of Lemma SA-3.1. Also, in view of Lemma SA-3.1, the lower bound on the minimum eigenvalue of $\mathbf{Q}_0$ follows from Theorem

4.41 of Schumaker (2007), by which the minimum eigenvalue of $\mathbf{Q}_0/J$ (the scaling factor dropped) is bounded by $\min_{1\leq j\leq J} h_j$ up to some universal constant.

Now, we prove the convergence of $\widehat{\mathbf{Q}}$. In view of Lemma SA-3.2, it suffices to show the convergence of $\widehat{\mathbf{Q}}$ when $s=0$, i.e., $\|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']-\mathbb{E}[\mathbf{b}_{p,0}(x_i)\mathbf{b}_{p,0}(x_i)']\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$. By Lemma SA-3.1, with probability approaching one, $\widehat{\Delta}\in\Pi$. Let $\mathcal{A}_n$ denote the event on which $\widehat{\Delta}\in\Pi$. Thus, $\mathbb{P}(\mathcal{A}_n^c)=o(1)$. On $\mathcal{A}_n$,

$$\begin{aligned}&\left\|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']-\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']\right\|\\ \leq&\sup_{\Delta\in\Pi}\left\|\mathbb{E}_n[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)']-\mathbb{E}[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)']\right\|.\end{aligned}$$

By the relation between matrix norms, the right-hand-side of the above inequality is further bounded by $\sup_{\Delta\in\Pi}\|\mathbb{E}_n[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)']-\mathbb{E}[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)']\|_\infty$. Let a_{kl} be a generic (k,l)th entry of the matrix inside $\|\cdot\|_\infty$. Then,

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

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

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

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

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

for ε small enough where we take $\bar{G}=CJ$ for some constant $C>0$ large enough. Theorem 6.1 of

Belloni, Chernozhukov, Chetverikov and Kato (2015),

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

implying that

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

Since any row or column of the matrix $n^{-1/2}\cdot\mathbb{G}_n[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)']$ only contains a finite number of nonzero entries, only depending on p, the above result suffices to show that

$$\Big\|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']-\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']\Big\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}.$$

Next, let α_{kl} be a generic (k,l)th entry of $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']/J-\mathbb{E}[\mathbf{b}_{p,0}(x_i)\mathbf{b}_{p,0}(x_i)']/J$, where by dividing the matrix by J, we drop the normalizing constant for notation simplicity. By definition, it is either equal to zero or can be rewritten as

$$\begin{aligned}\alpha_{kl}=&\int_{\widehat{\mathcal{B}}_j}\Big(\frac{x-\hat{\tau}_j}{\hat{h}_j}\Big)^{\ell}f_X(x)dx-\int_{\widehat{\mathcal{B}}_j}\Big(\frac{x-\tau_j}{h_j}\Big)^{\ell}f_X(x)dx\\=&\hat{h}_j\int_0^1 z^{\ell}f_X(z\hat{h}_j+\hat{\tau}_j)dz-h_j\int_0^1 z^{\ell}f_X(zh_j+\tau_j)dz\\=&(\hat{h}_j-h_j)\int_0^1 z^{\ell}f_X(z\hat{h}_j+\hat{\tau}_j)dz+h_j\int_0^1 z^{\ell}\Big(f_X(z\hat{h}_j+\hat{\tau}_j)-f_X(zh_j+\tau_j)\Big)dz\end{aligned}\qquad\text{(SA-5.4)}$$

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

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

Since $f_X(\cdot)$ is uniformly continuous on $\mathcal{X}$, the second term in (SA-5.4) is also $O_{\mathbb{P}}(J^{-1}\sqrt{J\log J/n})$. Again, using the sparsity structure of the matrix $\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']/J-\mathbb{E}[\mathbf{b}_{p,0}(x_i)\mathbf{b}_{p,0}(x_i)']/J$, the above result suffices to show that $\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)']-\mathbf{Q}_0\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$.

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

SA-5.6 Proof of Lemma SA-3.6

Proof. Since $\mathbb{E}[\epsilon_i^2|x_i = x]$ is bounded and bounded away from zero uniformly over $x \in \mathcal{X}$, we have $\widehat{\mathbf{Q}} \lesssim \bar{\mathbf{\Sigma}} \lesssim \widehat{\mathbf{Q}}$. Then, by Lemma SA-3.5, $1 \lesssim_{\mathbb{P}} \lambda_{\min}(\bar{\mathbf{\Sigma}}) \lesssim \lambda_{\max}(\bar{\mathbf{\Sigma}}) \lesssim_{\mathbb{P}} 1$. The upper bound on $\bar{\Omega}(x)$ immediately follows by Lemmas SA-3.3 and SA-3.5.

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

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

For any $x_0 \in \mathcal{X}$, there are only a finite number of basis functions in $\breve{\mathbf{b}}_{p,s}(x)$ supported on $\mathcal{B}_{x_0}$. By Theorem 4.41 of Schumaker (2007), for each $\breve{b}_{p,s,j}(x)$, $j = 1, \cdots, K_{p,s}$, we have $|\varphi_j \varrho| \lesssim \|\varrho\|_{L_\infty[\mathcal{I}_j]}$ where $\mathcal{I}_j$ denotes the support of $\breve{b}_{p,s,j}(x)$ and $\|\cdot\|_{L_\infty[\mathcal{I}_j]}$ denotes the sup-norm on $\mathcal{I}_j$. All points within such $\mathcal{I}_j$ should be no greater than $(p+1)\max_{1\leq j\leq J} h_j(\Delta)$ away from x_0 where $h_j(\Delta)$ denotes the length of the jth bin in Δ. Hence, $\|\varrho\|_{L_\infty[\mathcal{I}_j]} \lesssim 1$. The desired lower bound follows. The bound on $\Omega(x)$ can be established similarly. $\square$

SA-5.7 Proof of Lemma SA-3.7

Proof. By Lemmas SA-3.2, SA-3.3 and SA-3.5, $\sup_{x\in\mathcal{X}} \|\widehat{\mathbf{b}}_{p,s}^{(v)}(x)\|_1 \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$. Define a function class

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

Then, $\sup_{g\in\mathcal{G}}|g| \lesssim \sqrt{J}|\epsilon_1|$, and hence take an envelop $\bar{G} = C\sqrt{J}|\epsilon_1|$ for some C large enough. Moreover, $\sup_{g\in\mathcal{G}}\mathbb{V}[g] \lesssim 1$ and, as in the proof of Lemma SA-3.5, $\mathcal{G}$ is of VC-type. By Proposition 6.1 of Belloni, Chernozhukov, Chetverikov and Kato (2015),

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

and the desired result follows. □

SA-5.8 Proof of Lemma SA-3.8

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

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

By Lemma SA-3.4, $\sup_{\Delta\in\Pi}|r_0(x;\Delta)|_\infty \lesssim J^{-p-1}$. Then, $\sup_{g\in\mathcal{G}}|g|_\infty \lesssim J^{-p-1+1/2}$, and $\sup_{g\in\mathcal{G}}\mathbb{V}[g] \lesssim J^{-2(p+1)}$. In addition, any function $g\in\mathcal{G}$ can be rewritten as

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

for some $1 \leq l, \underline{k} \leq K_{p,s}$ where $\beta_{0,k}(\Delta)$ denotes the k-th element of $\boldsymbol{\beta}_0(\Delta)$. Here we use the sparsity property of the partitioning basis: the summand in the second term is nonzero only if $b_{p,s,l}(x;\Delta)$ and $b_{p,s,k}(x;\Delta)$ have overlapping supports. For each l, there are at most $(p+1)$ such basis functions $b_{p,s,k}(x;\Delta)$s. Also, the first term and every summand in the second term are bounded by $\sqrt{J}$ up to some constant. Then, using the same argument given in the proof of Lemma SA-3.5,

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

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

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

and, by Lemma SA-3.5, $\|\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}_0^{-1}\|_\infty \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$. Then, using the bound on the basis given in Lemma SA-3.3,

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

These results complete the proof. □

SA-5.9 Proof of Lemma SA-3.9

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

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

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

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

We have

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

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

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

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

On the other hand, $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \mathbf{w}_i M_{ij}\epsilon_j$ has mean zero with variance of order $O(1/n)$ by Lemma SA-2 of Cattaneo, Jansson and Newey (2018b). In addition, as in Lemma 2 of Cattaneo, Jansson and Newey (2018a), let $\mathbf{G} = (\mu_0(x_1), \dots, \mu_0(x_n))'$ and note that

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

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

To show the second result, by Lemmas SA-3.2, SA-3.3 and SA-3.5, $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}^{(v)}_{p,s}(x)\|_1 \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$. $\mathbb{E}_n[\widehat{\mathbf{b}}_{p,0}(x_i)\mathbf{w}_i']$ is a $J(p+1)\times d$ matrix and can be decomposed as follows:

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

By the argument in the proof of Lemma SA-3.5 and the conditions that $\sup_{x\in\mathcal{X}}\|\mathbb{E}[\mathbf{w}_i|x_i = x]\| \lesssim 1$ and $\frac{J\log J}{n} = o(1)$, $\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\mathbb{E}[\mathbf{w}_i'|x_i]]\|_\infty \lesssim_{\mathbb{P}} J^{-1/2}$. Regarding the second term, note that it is a mean zero sequence, and for the lth covariate in $\mathbf{w}$, $l = 1, \dots, d$,

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

Thus the second result follows by Markov's inequality.

Now suppose $\frac{J^{\frac{\nu}{\nu-2}}\log J}{n} \lesssim 1$ also holds. Using the argument given in Lemma SA-3.7 and the assumption that $\sup_{x\in\mathcal{X}}\mathbb{E}[|w_{i,l}|^\nu|x_i = x] \lesssim 1$ for all l, we have $\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)(w_{i,l} - \mathbb{E}[w_{i,l}|x_i])]\|_\infty \lesssim_{\mathbb{P}} \sqrt{\log J/n}$. Thus, the last result follows. □

SA-5.10 Proof of Theorem SA-3.1

Proof. The result follows by Lemmas SA-3.4, SA-3.8 and SA-3.9. □

SA-5.11 Proof of Corollary SA-3.1

Proof. The result follows by Theorem SA-3.1 and Lemma SA-3.7. □

SA-5.12 Proof of Theorem SA-3.2

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

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

Now, we bound each term in the following.

Step 1: For $\mathbf{V}_1$, we further write $u_i = (\mu_0(x_i) - \widehat{\mathbf{b}}_{p,s}(x_i)'\widehat{\boldsymbol{\beta}}) - \mathbf{w}_i'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) =: u_{i1} - u_{i2}$. Then

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

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

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

where the last inequality holds by Lemma SA-3.5 and Corollary SA-3.1. On the other hand, let $\widehat{\gamma}_\ell$ and $\gamma_{0,\ell}$ denote the ℓth entry of $\widehat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}_0$. We have

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

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

by CR-inequality. By Lemma SA-3.9, $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|^2=o_{\mathbb{P}}(J/n)$. Then it suffices to show that for every $\ell=1,\dots,d$, $\|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'w_{i,\ell}^2]\|\lesssim_{\mathbb{P}}1$. Under the conditions given in the theorem, this bound can be established using the argument that will be given in Step 3 and 4 and that in Lemma SA-3.5.

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

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

where the last step follows by Lemma SA-3.5 and the result given in Step 3. In addition,

$$\|\mathbf{V}_{22}\|=\Big\|2\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\epsilon_i\sum_{\ell=1}^{d}w_{i,\ell}(\widehat{\gamma}_\ell-\gamma_{0,\ell})]\Big\|\lesssim_{\mathbb{P}}\frac{1}{\sqrt{n}}+J^{-p-1-(\varsigma_w\wedge(p+1))}.$$

Since $\|2\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'\epsilon_i w_{i,\ell}]\|\leq\|\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)'(\epsilon_i^2+w_{i,\ell}^2)]\|$, this bound on $\|\mathbf{V}_{22}\|$ can be established using Lemma SA-3.9 and the strategy given in Step 3 and Step 4 and that in Lemma SA-3.5.

Step 3: For $\mathbf{V}_3$, in view of Lemma SA-3.1 and SA-3.2, it suffices to show that

$$\sup_{\Delta\in\Pi}\Big\|\mathbb{E}_n[\mathbf{b}_{p,0}(x_i;\Delta)\mathbf{b}_{p,0}(x_i;\Delta)'(\epsilon_i^2-\sigma^2(x_i))]\Big\|\lesssim_{\mathbb{P}}\Big(\frac{J\log J}{n^{\frac{\nu-2}{\nu}}}\Big)^{1/2}.$$

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

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

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

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

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

The VC condition holds by the same argument given in the proof of Lemma SA-3.5. Then, by Proposition 6.1 of Belloni, Chernozhukov, Chetverikov and Kato (2015),

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

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

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

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

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

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-LS, $\sup_{x\in\mathcal{X}}\mathbb{E}[\epsilon_i^2|x_i=x]\lesssim 1$. Then, by the same argument given in the proof of Lemma SA-3.5,

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

Then the proof is complete. □

SA-5.13 Proof of Theorem SA-3.3

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

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

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

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

Now, using Lemmas SA-3.3, SA-3.5 and SA-3.6,

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

since $J/n = o(1)$. By Theorem SA-3.2, the above weak convergence still holds if $\bar{\Omega}(x)$ is replaced by $\widehat{\Omega}(x)$. Now, the desired result follows by Lemmas SA-3.4, SA-3.8 and SA-3.9. □

SA-5.14 Proof of Theorem SA-3.4

Proof. Since $\widehat{\Upsilon}(x,\widehat{\mathbf{w}})$ differs from $\widehat{\mu}(x)$ only when $v = 0$, we will first focus on the IMSE of $\widehat{\mu}^{(v)}(x)$. We rely on the following decomposition:

$$\begin{aligned}
\widehat{\mu}^{(v)}(x)-\mu_0^{(v)}(x) = &\,\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\epsilon_i]+\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{r}_0(x_i)]+\\
&\Big(\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\boldsymbol{\beta}}_0-\mu_0^{(v)}(x)\Big)-\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\mathbf{w}_i'](\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0).
\end{aligned} \qquad \text{(SA-5.5)}$$

The proof is divided into several steps.

Step 1: By Lemma SA-3.9, the variance of the last term is of smaller order, and thus it suffices to characterize the conditional variance of $A(x) := \widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}\epsilon_i]$. By Lemma SA-3.5,

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

In fact, using the argument given in the proof of Lemma SA-3.3, we also have

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

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

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

Step 2: By decomposition (SA-5.5),

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

By Lemma SA-3.8, $\int_{\mathcal{X}}\mathfrak{B}_1(x)^2\omega(x)dx = o_{\mathbb{P}}(J^{-2p-2+2v})$. By Lemma SA-3.9, $\int_{\mathcal{X}}\mathfrak{B}_3(x)^2\omega(x)dx = o_{\mathbb{P}}(J^{-2p-2+2v})$. By Lemma SA-3.4, $\int_{\mathcal{X}}\mathfrak{B}_2(x)^2\omega(x)dx \lesssim_{\mathbb{P}} J^{-2p-2+2v}$. By Cauchy-Schwarz inequality, the integrals of those cross-product terms is of higher-order in the IMSE expansion, and the leading term in the integrated squared bias is

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

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

$$\sup_{x\in\mathcal{X}}\Big|\mu_0^{(v)}(x) - \widehat{\mathbf{b}}^{(v)}_{p,p}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta}) - \frac{\mu^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}_x^{\mathtt{L}}}{\hat{h}_x}\Big)\Big| = o_{\mathbb{P}}(J^{-(p+1-v)}), \quad \text{(SA-5.6)}$$

where for each $m \in \mathbb{Z}_+$, $\mathscr{E}_m(\cdot)$ is the mth Bernoulli polynomial, $\hat{\tau}_x^{\mathtt{L}}$ is the start of the (random) interval in $\widehat{\Delta}$ containing x and $\hat{h}_x$ denotes its length. When $s < p$, $\widehat{\mathbf{b}}_{p,p}(x)'\boldsymbol{\beta}_\infty$ is still an element in the space spanned by $\widehat{\mathbf{b}}_{p,s}(x)$. In other words, it provides a valid approximation of $\mu_0^{(v)}(x)$ in the larger space in terms of sup-norm. Then it follows that

$$\begin{aligned}
&\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\widehat{\boldsymbol{\beta}}_0-\mu_0^{(v)}(x)\\
&=\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\Big(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)']\Big)^{-1}\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,s}(x_i)\mu_0(x_i)]-\mu_0^{(v)}(x)\\
&=\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\Big(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,s}(x_i)\widehat{\mathbf{b}}_{p,s}(x_i)']\Big)^{-1}\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_{p,s}(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!}\hat{h}_{x_i}^{p+1}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}^{\mathtt{L}}_{x_i}}{\hat{h}_{x_i}}\Big)\Big]\\
&\qquad-\frac{\mu_0^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}^{\mathtt{L}}_x}{\hat{h}_x}\Big)+o_{\mathbb{P}}(J^{-p-1+v})\\
&=J^{-p-1}\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_{p,0}(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}^{\mathtt{L}}_{x_i}}{\hat{h}_{x_i}}\Big)\Big]\\
&\qquad-\frac{J^{-p-1+v}\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}^{\mathtt{L}}_x}{\hat{h}_x}\Big)+o_{\mathbb{P}}(J^{-p-1+v}),
\end{aligned}\tag{SA-5.7}$$

where the last step uses Lemmas SA-3.1-SA-3.3 and SA-3.5, and $o_{\mathbb{P}}(\cdot)$ holds uniformly over $x\in\mathcal{X}$. Taking integral of the squared bias and using Assumption SA-DGP and Lemmas SA-3.1–SA-3.3 and SA-3.5 again, we have three leading terms:

$$\begin{aligned}
M_1(x):=&\int_{\mathcal{X}}\bigg(\frac{J^{-p-1+v}\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}^{\mathtt{L}}_x}{\hat{h}_x}\Big)\bigg)^2\omega(x)dx\\
=&\frac{J^{-2p-2+2v}|\mathscr{E}_{2p+2-2v}|}{(2p+2-2v)!}\int_{\mathcal{X}}\Big[\frac{\mu_0^{(p+1)}(x)}{f_X(x)^{p+1-v}}\Big]^2\omega(x)dx+o_{\mathbb{P}}(J^{-2p-2+2v}),\\
M_2(x):=&J^{-2p-2}\int_{\mathcal{X}}\bigg(\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_{p,0}(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}^{\mathtt{L}}_{x_i}}{\hat{h}_{x_i}}\Big)\Big]\bigg)^2\omega(x)dx\\
=&J^{-2p-2}\boldsymbol{\xi}_{0,f}'\mathbf{T}_s'\mathbf{Q}_0^{-1}\Big(\int_{\mathcal{X}}\mathbf{b}_s^{(v)}(x)\mathbf{b}_s^{(v)}(x)'\omega(x)dx\Big)\mathbf{Q}_0^{-1}\mathbf{T}_s\boldsymbol{\xi}_{0,f}+o_{\mathbb{P}}(J^{-2p-2+2v}),\\
M_3(x):=&J^{-2p-2+v}\int_{\mathcal{X}}\bigg\{\bigg(\widehat{\mathbf{b}}^{(v)}_{p,s}(x)'\mathbf{Q}_0^{-1}\mathbf{T}_s\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_{p,0}(x_i)\frac{\mu_0^{(p+1)}(x_i)}{(p+1)!f_X(x_i)^{p+1}}\mathscr{E}_{p+1}\Big(\frac{x_i-\hat{\tau}^{\mathtt{L}}_{x_i}}{\hat{h}_{x_i}}\Big)\Big]\bigg)\\
&\qquad\times\frac{\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\hat{\tau}^{\mathtt{L}}_x}{\hat{h}_x}\Big)\bigg\}\omega(x)dx\\
=&J^{-2p-2+v}\boldsymbol{\xi}_{0,f}'\mathbf{T}_s'\mathbf{Q}_0^{-1}\mathbf{T}_s\boldsymbol{\xi}_{v,\omega}+o_{\mathbb{P}}(J^{-2p-2+2v}),
\end{aligned}$$

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

$$\boldsymbol{\xi}_{v,\lambda}=\int_{\mathcal{X}}\mathbf{b}^{(v)}_{p,0}(x)\frac{\mu_0^{(p+1)}(x)}{(p+1-v)!f_X(x)^{p+1-v}}\mathscr{E}_{p+1-v}\Big(\frac{x-\tau^{\mathtt{L}}_x}{h_x}\Big)\lambda(x)dx.$$

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

Step 3: For $v = 0$, we will have two additional terms $\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ and $(\widehat{\mathbf{w}} - \mathbf{w})'\boldsymbol{\gamma}_0$ in the decomposition of $\widehat{\Upsilon}(x, \widehat{\mathbf{w}}) - \Upsilon_0(x, \mathbf{w})$. By Assumption, $\widehat{\mathbf{w}} - \mathbf{w} = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$, and thus $(\widehat{\mathbf{w}} - \mathbf{w})'\boldsymbol{\gamma}_0$ as a (conditional) bias term is of higher order. The term $\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ can be treated the same way as we analyze $\widehat{\mathbf{b}}_{p,s}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_{p,s}(x_i)\mathbf{w}_i'](\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$. By Lemma SA-3.9, it is also of higher order. Then, the proof is complete. □

SA-5.15 Proof of Corollary SA-3.2

Proof. The proof is divided into two steps.

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

$$\mathbf{Q}_0 = \check{\mathbf{Q}}\mathfrak{D}_f + o_{\mathbb{P}}(1), \; \boldsymbol{\Sigma}_0 = \check{\mathbf{Q}}\mathfrak{D}_{\sigma^2 f} + o_{\mathbb{P}}(1), \; \text{and} \; \int_{\mathcal{X}} \mathbf{b}_{p,0}^{(v)}(x)\mathbf{b}_{p,0}^{(v)}(x)'\omega(x)dx = \check{\mathbf{Q}}_v\mathfrak{D}_\omega + o_{\mathbb{P}}(J^{2v}),$$

where

$$\check{\mathbf{Q}} = \int_{\mathcal{X}} \mathbf{b}_{p,0}(x)\mathbf{b}_{p,0}(x)'dx, \; \check{\mathbf{Q}}_v = \int_{\mathcal{X}} \mathbf{b}_{p,0}^{(v)}(x)\mathbf{b}_{p,0}^{(v)}(x)'dx, \; \mathfrak{D}_f = \operatorname{diag}\{f_X(\check{x}_1), \cdots, f_X(\check{x}_{J(p+1)})\},$$

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

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

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

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

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

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

for $\boldsymbol{\varphi}(z) = (1, z, \dots, z^p)$. Employing Lemma SA-3.1 and letting the trace converge to the Riemann integral, we conclude that

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

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

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

$$r_\infty^{(v)}(x;\widehat{\Delta}) := \mu_0^{(v)}(x) - \widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta}) = \frac{\mu_0^{(p+1)}(x)}{(p+1-v)!}\hat{h}_x^{p+1-v}\mathscr{B}_{p+1-v}\Big(\frac{x-\hat{\tau}_x^{\mathtt{L}}}{\hat{h}_x}\Big) + o_{\mathbb{P}}(J^{-(p+1-v)}),$$

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

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

where the last line follows by change of variables and the orthogonality of Legendre polynomials.

Thus, $r_\infty(x;\widehat{\Delta})$ is approximately orthogonal to the space spanned by $\widehat{\mathbf{b}}_{p,0}(x)$. Immediately, we have

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

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

$$\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x)(r_0(x;\widehat{\Delta}) - r_\infty(x;\widehat{\Delta}))]\| = \|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x)\widehat{\mathbf{b}}_{p,0}(x)'(\boldsymbol{\beta}_\infty(\widehat{\Delta}) - \boldsymbol{\beta}_0(\widehat{\Delta}))]\| = o_{\mathbb{P}}(J^{-p-1}).$$

By Lemma SA-3.5, $\lambda_{\min}(\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_{p,0}(x_i)\widehat{\mathbf{b}}_{p,0}(x_i)]') \gtrsim_{\mathbb{P}} 1$, and thus $\|\boldsymbol{\beta}_\infty(\widehat{\Delta}) - \boldsymbol{\beta}_0(\widehat{\Delta})\| = o_{\mathbb{P}}(J^{-p-1})$. Then,

$$\begin{aligned}&\int_{\mathcal{X}}\Big(\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'(\boldsymbol{\beta}_0(\widehat{\Delta}) - \boldsymbol{\beta}_\infty(\widehat{\Delta}))\Big)^2\omega(x)dx\\ &\leq \lambda_{\max}\Big(\int_{\mathcal{X}}\widehat{\mathbf{b}}_{p,0}^{(v)}(x)\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\omega(x)dx\Big)\|\boldsymbol{\beta}_0(\widehat{\Delta}) - \boldsymbol{\beta}_\infty(\widehat{\Delta})\|^2 = o_{\mathbb{P}}(J^{-2p-2+2v}).\end{aligned}$$

Therefore, we can represent the leading term in the integrated squared bias by L_∞ approximation error: $\int_{\mathcal{X}}\mathfrak{B}_2(x)^2\omega(x)dx = \int_{\mathcal{X}}(\mu_0^{(v)}(x) - \widehat{\mathbf{b}}_{p,0}^{(v)}(x)'\boldsymbol{\beta}_\infty(\widehat{\Delta}))^2\omega(x)dx + o_{\mathbb{P}}(J^{-2p-2+2v})$. Finally, using the results given in Lemma SA-3.1, change of variables and the definition of Riemann integral, we conclude that

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

where

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

Then the proof is complete. □

SA-5.16 Proof of Theorem SA-3.5

Proof. The proof is divided into several steps.

Step 1: Note that

$$\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}} - \frac{\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x)}{\sqrt{\Omega(x)/n}}\right|$$

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

where the last step uses Lemma SA-3.6, Corollary SA-3.1 and Theorem SA-3.2. Then, in view of Lemmas SA-3.4, SA-3.8, SA-3.9 and Theorem SA-3.2 and the rate restriction given in the lemma, we have

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

Step 2: Let us write $\mathscr{K}(x,x_i)=\Omega(x)^{-1/2}\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbf{b}_{p,s}(x_i)$. Now we rearrange $\{x_i\}_{i=1}^n$ as a sequence of order statistics $\{x_{(i)}\}_{i=1}^n$, i.e., $x_{(1)}\leq\cdots\leq x_{(n)}$. Accordingly, $\{\epsilon_i\}_{i=1}^n$ and $\{\sigma^2(x_i)\}_{i=1}^n$ are ordered as concomitants $\{\epsilon_{[i]}\}_{i=1}^n$ and $\{\sigma_{[i]}^2\}_{i=1}^n$ where $\sigma_{[i]}^2=\sigma^2(x_{(i)})$. Clearly, conditional on $\mathbf{X}$, $\{\epsilon_{[i]}\}_{i=1}^n$ is still an independent mean zero sequence. Then by Assumptions SA-DGP, SA-LS and the result of Sakhanenko (1991), there exists a sequence of i.i.d. standard normal random variables $\{\zeta_{[i]}\}_{i=1}^n$ such that

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

Then, using summation by parts,

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

By Lemmas SA-3.3, SA-3.5 and SA-3.6, $\sup_{x\in\mathcal{X}}\sup_{x_i\in\mathcal{X}}|\mathscr{K}(x,x_i)|\lesssim_{\mathbb{P}}\sqrt{J}$, and

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

Then, notice that

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

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

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

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

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

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

Step 3: Note that

$$\begin{aligned}\sup_{x\in\mathcal{X}}|\bar{Z}_p(x)-Z_p(x)|\leq&\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'(\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1})}{\sqrt{\Omega(x)}}\bar{\mathbf{\Sigma}}^{1/2}\mathbf{N}_{K_{p,s}}\Big|\\&+\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}_{p,s}^{(v)}(x)'\mathbf{Q}_0^{-1}}{\sqrt{\Omega(x)}}\Big(\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\Big)\mathbf{N}_{K_{p,s}}\Big|\\&+\sup_{x\in\mathcal{X}}\Big|\frac{\widehat{\mathbf{b}}_{p,0}^{(v)}(x)'(\widehat{\mathbf{T}}_s-\mathbf{T}_s)\mathbf{Q}_0^{-1}}{\sqrt{\Omega(x)}}\mathbf{\Sigma}_0^{1/2}\mathbf{N}_{K_{p,s}}\Big|,\end{aligned}$$

where each term on the right-hand side is a mean-zero Gaussian process conditional on $\mathbf{X}$. By Lemmas SA-3.2 and SA-3.5, $\|\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}_0^{-1}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$ and $\|\widehat{\mathbf{T}}_s-\mathbf{T}_s\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$. Also, using the argument in the proof of Lemma SA-3.5 and Theorem X.3.8 of Bhatia (2013), $\|\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}_0^{1/2}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$. By Gaussian Maximal Inequality (see, e.g., van der Vaart and Wellner, 1996, Corollary 2.2.8),

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

where the last line follows from the imposed rate restriction.

As a reminder, if we drop the third term on the right-hand side, we obtain the same strong approximation result except that the approximating process is

$$\frac{\widehat{\mathbf{b}}^{(v)}_{p,s}(\cdot)'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0^{1/2}}{\sqrt{\Omega(x)}}\mathbf{N}_{K_{p,s}}.$$

Step 4: The above steps have shown the desired result for $v > 0$ already. For $v = 0$,

$$T_p(x) = \frac{\widehat{\Upsilon}(x, \widehat{\mathbf{w}}) - \Upsilon_0(x, \mathbf{w})}{\sqrt{\widehat{\Omega}(x)/n}} = \frac{\widehat{\mu}(x) - \mu_0(x)}{\sqrt{\widehat{\Omega}(x)/n}} + \frac{\widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}} - \mathbf{w}'\boldsymbol{\gamma}_0}{\sqrt{\widehat{\Omega}(x)/n}},$$

where

$$\frac{\widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}} - \mathbf{w}'\boldsymbol{\gamma}_0}{\sqrt{\widehat{\Omega}(x)/n}} = \frac{(\widehat{\mathbf{w}} - \mathbf{w})'\widehat{\boldsymbol{\gamma}}}{\sqrt{\widehat{\Omega}(x)/n}} + \frac{\mathbf{w}'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)}{\sqrt{\widehat{\Omega}(x)/n}} = o_{\mathbb{P}}(a_n^{-1})$$

by Lemma SA-3.9, Theorem SA-3.2 and the condition $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n})$. Therefore, the desired strong approximation for $\widehat{\Upsilon}(x, \widehat{\mathbf{w}})$ follows from the previous steps. Then, the proof is complete. □

SA-5.17 Proof of Theorem SA-3.6

Proof. This conclusion follows from Lemmas SA-3.3 and SA-3.5, Theorem SA-3.2 and Gaussian Maximal Inequality as applied in Step 3 in the proof of Theorem SA-3.5. □

SA-5.18 Proof of Theorem SA-3.7

Proof. We first show that

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

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

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

Then,

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

Now, apply the Anti-Concentration Inequality conditional on $\mathbf{X}$ (see Chernozhukov, Chetverikov and Kato, 2014b) to the second term:

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

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

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

The other side of the inequality follows similarly.

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

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

Then it remains to show that

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)|\leq u\Big)-\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)|\leq u|\mathbf{X}\Big)\Big|=o_{\mathbb{P}}(1). \qquad \text{(SA-5.8)}$$

Now, we can write

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

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 normal 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.

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

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

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,

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

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

For $v > 0$, simply notice that $\widehat{\mathbf{b}}_{p,0}^{(v)}(x) = \widehat{\mathfrak{T}}_v \widehat{\mathbf{b}}_{p,0}(x)$ 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-3.1, 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, e.g., the proof of Theorems SA-3.5 and SA-3.6, we can replace $\widehat{\mathfrak{T}}_v$ in $Z_p(x)$ by $\mathfrak{T}_v$ without affecting the approximation rate. Then the desired result follows by repeating the argument given for $v = 0$ above. □

SA-5.19 Proof of Theorem SA-3.8

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

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

where $c^0(1-\alpha+\xi_{3,n})$ denotes the $(1-\alpha+\xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}} |Z_p(x)|$ (given the partition), the first inequality holds by Theorem SA-3.5, the second by Lemma A.1 of Belloni, Chernozhukov, Chetverikov and Kato (2015), and the third by Anti-Concentration Inequality in Chernozhukov, Chetverikov and Kato (2014b). The other side of the bound follows similarly. □

SA-5.20 Proof of Theorem SA-3.9

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

$$\sup_{x\in\mathcal{X}} |\dot{T}_p(x)| \leq \sup_{x\in\mathcal{X}} \left| \frac{\widehat{\Upsilon}^{(v)}(x,\widehat{\mathbf{w}}) - \Upsilon_0^{(v)}(x,\mathbf{w})}{\sqrt{\widehat{\Omega}(x)/n}} \right| + \sup_{x\in\mathcal{X}} \left| \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} \right|.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\Big[\sup_{x\in\mathcal{X}} |\dot{T}_p(x)| > \mathfrak{c}\Big] &\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |T_p(x)| > \mathfrak{c} - \sup_{x\in\mathcal{X}} \left| \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} \right|\Big] \\
&\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |Z_p(x)| > \mathfrak{c} - \xi_{1,n}/a_n - \sup_{x\in\mathcal{X}} \left| \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} \right|\Big] + o(1) \\
&\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |Z_p(x)| > c^0(1-\alpha-\xi_{3,n}) - (\xi_{1,n}+\xi_{2,n})/a_n - \\
&\qquad \sup_{x\in\mathcal{X}} \left| \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} \right|\Big] + o(1) \\
&\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |Z_p(x)| > c^0(1-\alpha-\xi_{3,n})\Big] + o(1) \\
&= \alpha + o(1)
\end{aligned}$$

where $c^0(1-\alpha-\xi_{3,n})$ denotes the $(1-\alpha-\xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}} |Z_p(x)|$ (given the partition), the second inequality holds by Theorem SA-3.5, the third by Lemma A.1 of Belloni, Chernozhukov, Chetverikov and Kato (2015), the fourth by the fact that $\sup_{x\in\mathcal{X}} \left| \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} \right| = o_{\mathbb{P}}(\frac{1}{\sqrt{\log J}})$ and Anti-Concentration Inequality in Chernozhukov, Chetverikov and Kato (2014b). The other side of the bound follows similarly.

On the other hand, under $\dot{\mathsf{H}}_{\mathrm{A}}$,

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

$$
\begin{aligned}
&= \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \Big| T_p(x) + \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} + \frac{M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big| > \mathfrak{c}\Big]\\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |T_p(x)| < \sup_{x\in\mathcal{X}} \Big| \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} + \frac{M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big| - \mathfrak{c}\Big]\\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |Z_p(x)| \leq \sqrt{\log J} A_n - \xi_{1,n}/a_n\Big] - o(1)\\
&\geq 1 - o(1).
\end{aligned}
$$

where the fourth line holds by Lemma SA-3.6, 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). □

SA-5.21 Proof of Theorem SA-3.10

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. Note that under $\ddot{\mathsf{H}}_0$,

$$
\sup_{x\in\mathcal{X}} \ddot{T}_p(x) \leq \sup_{x\in\mathcal{X}} T_p(x) + \sup_{x\in\mathcal{X}} \frac{|M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})|}{\sqrt{\widehat{\Omega}(x)/n}}.
$$

Then,

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

where $c^0(1-\alpha-\xi_{3,n})$ denotes the $(1-\alpha-\xi_{3,n})$-quantile of $\sup_{x\in\mathcal{X}} Z_p(x)$ (given the partition), the second line holds by Theorem SA-3.5, the third by Lemma A.1 of Belloni, Chernozhukov, Chetverikov and Kato (2015), the fourth by Anti-Concentration Inequality in Chernozhukov, Chetverikov and

Kato (2014b).

On the other hand, under $\ddot{\mathsf{H}}_{\mathrm{A}}$,

$$\begin{aligned}
\mathbb{P}\Big[\sup_{x\in\mathcal{X}} \ddot{T}_p(x) > \mathfrak{c}\Big] &= \mathbb{P}\Big[\sup_{x\in\mathcal{X}}\Big(T_p(x) + \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\gamma})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}\Big) > 0\Big] \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |T_p(x)| < \sup_{x\in\mathcal{X}} \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\gamma})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}, \\
&\qquad \sup_{x\in\mathcal{X}} \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\gamma})}{\sqrt{\widehat{\Omega}(x)/n}} > \mathfrak{c}\Big] \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |T_p(x)| < \sup_{x\in\mathcal{X}} \frac{\Upsilon_0^{(v)}(x,\mathbf{w}) - M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\gamma})}{\sqrt{\widehat{\Omega}(x)/n}} - \mathfrak{c}\Big] - o(1) \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |T_p(x)| < \sqrt{\log J} A_n\Big] - o(1) \\
&\geq \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |Z_p(x)| < \sqrt{\log J} A_n - \xi_{1,n}/a_n\Big] - o(1) \\
&\geq 1 - o(1)
\end{aligned}$$

where the fourth line holds by Lemma SA-3.6, Theorem SA-3.2, Lemma A.1 of Belloni, Chernozhukov, Chetverikov and Kato (2015), the assumptions that $J^v\sqrt{J\log J/n} = o(1)$ and $\sup_{x\in\mathcal{X}}$ $|M^{(v)}(x,\widehat{\mathbf{w}};\widetilde{\boldsymbol{\theta}},\widetilde{\gamma}) - M^{(v)}(x,\mathbf{w};\bar{\boldsymbol{\theta}},\bar{\gamma})| = o_{\mathbb{P}}(1)$, the fifth by definition of A_n, and the sixth by Theorem SA-3.5, and the last by Proposition A.2.7 in van der Vaart and Wellner (1996).

□

References

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato, "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 2015, *186* (2), 345–366.

Bhatia, Rajendra, *Matrix Analysis*, Springer, 2013.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell, "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 2018, *113* (522), 767–779.

_ , _ , and _ , "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," *Bernoulli*, 2022, *28* (4), 2998–3022.

_ , _ , and Rocio Titiunik, "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 2015, *110* (512), 1753–1769.

Cattaneo, Matias D., Max H. Farrell, and Yingjie Feng, "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, 2020, *48* (3), 1718–1741.

_ , Michael Jansson, and Whitney K. Newey, "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 2018, *34* (2), 277–301.

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

_ , Richard K. Crump, Max H. Farrell, and Yingjie Feng, "Nonlinear Binscatter Methods," working paper, 2023.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato, "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, 2014, *42* (4), 1564–1597.

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

de Boor, Carl, *A Practical Guide to Splines*, Springer-Verlag New York, 1978.

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

Giné, Evarist and Richard Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Vol. 40, Cambridge University Press, 2016.

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

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

Schumaker, Larry, *Spline Functions: Basic Theory*, Cambridge University Press, 2007.

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

van der Vaart, Add W. and Jon Wellner, *Weak Convergence and Empirical Processes: With Application to Statistics*, Springer, 1996.