Submitted to Bernoulli

Boundary adaptive local polynomial conditional density estimators

MATIAS D. CATTANEO[1,a], RAJITA CHANDAK[1,b], MICHAEL JANSSON[2,c] and XINWEI MA[3,d]

[1]*Department of Operations Research and Financial Engineering, Princeton University, Princeton NJ, United States,* [a]*cattaneo@princeton.edu,* [b]*rchandak@princeton.edu*
[2]*Department of Economics, UC Berkeley, Berkeley CA, United States,* [c]*mjansson@berkeley.edu*
[3]*Department of Economics, UC San Diego, La Jolla CA, United States,* [d]*x1ma@ucsd.edu*

We begin by introducing a class of conditional density estimators based on local polynomial techniques. The estimators are boundary adaptive and easy to implement. We then study the (pointwise and) uniform statistical properties of the estimators, offering characterizations of both probability concentration and distributional approximation. In particular, we establish uniform convergence rates in probability and valid Gaussian distributional approximations for the Studentized t-statistic process. We also discuss implementation issues such as consistent estimation of the covariance function for the Gaussian approximation, optimal integrated mean squared error bandwidth selection, and valid robust bias-corrected inference. We illustrate the applicability of our results by constructing valid confidence bands and hypothesis tests for both parametric specification and shape constraints, explicitly characterizing their approximation errors. A companion `R` software package implementing our main results is provided.

Keywords: Conditional density estimation; confidence bands; local polynomial methods; specification testing; strong approximation; uniform inference

1. Introduction

Suppose that $(y_1, \mathbf{x}_1^\intercal), (y_2, \mathbf{x}_2^\intercal), \dots, (y_n, \mathbf{x}_n^\intercal)$ is a random sample from a distribution supported on $\mathcal{Y} \times \mathcal{X}$, where $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^d$ are compact. Letting $F(y|\mathbf{x})$ be the conditional cumulative distribution function (CDF) of y_i given $\mathbf{x}_i$, important parameters of interest in statistics, econometrics, and many other data science disciplines, are the conditional probability density function (PDF) and derivatives thereof:

$$f^{(\vartheta)}(y|\mathbf{x}) = \frac{\partial^{1+\vartheta}}{\partial y^{1+\vartheta}} F(y|\mathbf{x}), \qquad \vartheta \in \{0, 1, 2, \dots\},$$

where, in particular, $f(y|\mathbf{x}) = f^{(0)}(y|\mathbf{x})$ is the conditional density function of y_i given $\mathbf{x}_i$.

Estimation and inference methodology for (conditional) PDFs has a long tradition in statistics [e.g., 26,27,29,30, and references therein]. Unfortunately, without specific modifications, smoothing methods employing kernel, series, or other local approximation techniques are invalid at or near boundary points of $\mathcal{Y} \times \mathcal{X}$. To address this challenge, we introduce a boundary adaptive nonparametric estimator of $f^{(\vartheta)}(y|\mathbf{x})$ based on local polynomial techniques [14] and provide an array of distributional approximation results that are valid (pointwise and) uniformly over $\mathcal{Y} \times \mathcal{X}$. In particular, we obtain a uniformly valid stochastic linear representation for the estimator and develop uniform inference methods based on strong approximation techniques leading to, for example, asymptotically valid confidence bands with careful characterization of their associated approximation errors.

To motivate our proposed estimation approach, suppose we start from an estimator of the conditional CDF, $\widehat{F}(\cdot|\mathbf{x})$. Then, for $y \in \mathbb{R}$, a natural estimator of $f^{(\vartheta)}(y|\mathbf{x})$ is obtained via local polynomial regression:

$$\widehat{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\intercal}\widehat{\boldsymbol{\beta}}(y|\mathbf{x}), \qquad \widehat{\boldsymbol{\beta}}(y|\mathbf{x}) = \underset{\mathbf{u}\in\mathbb{R}^{\mathfrak{p}+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(\widehat{F}(y_i|\mathbf{x}) - \mathbf{p}(y_i - y)^{\intercal}\mathbf{u}\Big)^2 K_h(y_i; y), \tag{1}$$

where $\mathfrak{p} \geq 1 + \vartheta$ is the order of the polynomial basis $\mathbf{p}(y) = (1, y/1!, y^2/2!, \ldots, y^{\mathfrak{p}}/\mathfrak{p}!)^{\intercal}$, $\mathbf{e}_l$ is the conformable $(1+l)$-th unit vector, and $K_h(y_i; y) = K((y_i - y)/h)/h$ for some kernel function K and some positive bandwidth h. Since $F(y|\mathbf{x}_i) = \mathbb{E}[\mathbb{1}(y_i \leq y)|\mathbf{x}_i]$, we employ a $\mathfrak{q}$-th order local polynomial regression of the indicator function, $\mathbb{1}(y_i \leq y)$, to form the conditional CDF estimator that will be plugged into (1):

$$\widehat{F}(y|\mathbf{x}) = \mathbf{e}_{\mathbf{0}}^{\intercal}\widehat{\boldsymbol{\gamma}}(y|\mathbf{x}), \qquad \widehat{\boldsymbol{\gamma}}(y|\mathbf{x}) = \underset{\mathbf{v}\in\mathbb{R}^{\mathfrak{q}_d+1}}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbb{1}(y_i \leq y) - \mathbf{q}(\mathbf{x}_i - \mathbf{x})^{\intercal}\mathbf{v})^2 L_b(\mathbf{x}_i; \mathbf{x}).$$

Here, using standard multi-index notation, $\mathbf{q}(\mathbf{x})$ denotes the $(\mathfrak{q}_d + 1)$-dimensional vector collecting the polynomial expansions $\mathbf{x}^{\mathbf{m}}/\mathbf{m}!$ for $0 \leq |\mathbf{m}| \leq \mathfrak{q}$, where $\mathbf{x}^{\mathbf{m}} = x_1^{m_1} x_2^{m_2} \cdots x_d^{m_d}$, $|\mathbf{m}| = m_1 + m_2 + \cdots + m_d$, and $\mathfrak{q}_d = (d + \mathfrak{q})!/(\mathfrak{q}!d!) - 1$. We also let $L_b(\mathbf{x}_i; \mathbf{x}) = L((\mathbf{x}_i - \mathbf{x})/b)/b^d$ be some (multivariate) kernel function L and positive bandwidth b. Our proposed estimator can also be written in closed-form as

$$\widehat{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\intercal}\widehat{\mathbf{S}}_y^{-1}\widehat{\mathbf{R}}_{y,\mathbf{x}}\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}, \tag{2}$$

where the matrices are

$$\widehat{\mathbf{S}}_y = \frac{1}{n}\sum_{i=1}^{n} \mathbf{p}\Big(\frac{y_i - y}{h}\Big)\frac{1}{h}\mathbf{P}\Big(\frac{y_i - y}{h}\Big)^{\intercal}, \qquad \widehat{\mathbf{S}}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{b}\Big)\frac{1}{b^d}\mathbf{Q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{b}\Big)^{\intercal},$$

$$\widehat{\mathbf{R}}_{y,\mathbf{x}} = \frac{1}{n^2 h^{1+\vartheta}}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{1}{h}\mathbf{P}\Big(\frac{y_j - y}{h}\Big)\frac{1}{b^d}\mathbf{Q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{b}\Big)^{\intercal}\mathbb{1}(y_i \leq y_j),$$

with the definitions $\mathbf{P}(y) = \mathbf{p}(y)K(y)$ and $\mathbf{Q}(\mathbf{x}) = \mathbf{q}(\mathbf{x})L(\mathbf{x})$, which absorb the kernel function into the basis. See Appendix A.1 for derivation.

By virtue of being based on a local polynomial smoothing approach, the estimator $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ is not only intuitive, but also boundary adaptive. Furthermore, $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ admits a simple closed-form representation as we have shown in (2), making it easy to implement. These features follow directly from its construction: unlike classical kernel-based conditional density (derivative) estimators, which seek to approximate the conditional PDF indirectly (e.g., by constructing a ratio of two unconditional kernel-based density estimators), our proposed estimator applies local polynomial techniques directly to the conditional CDF estimator $\widehat{F}(y|\mathbf{x})$. In addition, our approach offers an easy way to construct higher-order kernels to reduce misspecification (or smoothing) bias via the choice of polynomial orders $\mathfrak{p}$ and $\mathfrak{q}$.

We present two main uniform results for our proposed estimator. First, we provide precise uniform probability concentration bounds associated with a stochastic linear representation of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ (Lemma 1 and Theorem 1). In addition to being useful for the purposes of characterizing the distributional properties of the conditional density estimator itself, the first main result can be used to analyze multi-step estimation and inference procedures whenever $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ enters as a preliminary step. As a by-product of the development of the first main result, we obtain a related class of conditional density

estimators based on local smoothing. This new approach will require the knowledge of the support $\mathcal{Y}$. On the other hand, it is immune to "low" density regions of y_i. For details, see Appendix A.2.

Our second main result employs the stochastic linear representation of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ to establish a valid strong approximation for the standardized t-statistic stochastic process based on $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ and indexed over $\mathcal{Y} \times \mathcal{X}$ (Theorem 2). This result is established using a powerful result due to Rio [25], which in turn builds on the celebrated Hungarian construction [24]. The t-statistic stochastic processes based on kernel-based nonparametric estimators are not asymptotically tight and, as a consequence, do not converge weakly as a process indexed over $\mathcal{Y} \times \mathcal{X}$ [18,28]. Nevertheless, using strong approximations to such processes, it is possible to deduce distributional approximations for functionals thereof by employing anti-concentration [7]. Combining these ideas, we obtain valid distributional approximations for the suprema of the t-statistic stochastic process (Theorem 3) based on $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ with approximation rates that are faster than those currently available in the literature for the case of $d = 1$ (e.g., Remark 3.1(ii) in [8]).

In addition to our two main uniform estimation and distributional results, we discuss several implementation results that are useful for practice. First, we present a covariance function estimator for the Gaussian approximation and prove its uniform consistency (Lemma 2). This result enables us to estimate the statistical uncertainty underlying the Gaussian approximation for a feasible version of the t-statistic process. Second, in Section 3 we discuss optimal bandwidth selection based on an asymptotic approximation to the integrated mean squared error (IMSE) of the estimator $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. This result allows us to implement our proposed estimator using point estimation optimal data-driven bandwidth selection rules. Finally, we employ robust bias correction [1,2] to develop valid inference methods based on the Gaussian approximation when using the estimated covariance function and IMSE-optimal bandwidth rule.

We illustrate our theoretical and methodological results with three substantive applications in Section 3. To be specific, we construct valid confidence bands for the unknown conditional density function (and derivatives thereof) and we develop valid hypothesis testing procedures for parametric specification and shape constraints of $f^{(\vartheta)}(y|\mathbf{x})$, respectively. All these methods are data-driven and, in some cases, optimal in terms of probability and/or distributional concentration, possibly up to $\log(n)$ factors. Furthermore, thanks to the precise probability approximation errors we obtain via strong approximation and other exponential concentration methods, we are able to characterize precise coverage error and rejection probability error rates for all the feasible inference procedures considered.

Another advantage of our proposed estimation procedure (1) is that it allows for incorporating additional constraints easily. For example, setting $\vartheta = 0$ (PDF), it may be desirable to require that the estimator is non-negative and integrates to 1. In Section 4, we proposed a modified conditional PDF estimator which satisfies these two properties. To be precise, non-negativity can be imposed by solving a constrained version of (1), as the feature is local to the evaluation point. On the other hand, ensuring the estimator integrates to 1 requires imposing a global constraint, which we implement by minimizing the Kullback-Leibler divergence to ensure that the final estimator is a valid conditional density in finite samples. Interestingly, this modified conditional PDF estimator requires introducing a normalization factor that affects the strong approximation in nontrivial ways, leading to a different distributional Gaussian process approximation (Theorem 8).

Proofs of the main results are given in the Appendix. In the supplementary material [5], we consider a more general setup and offer additional technical and methodological results of potential independent interest, including: (i) boundary adaptive estimators for the CDF and its derivatives with respect to the conditioning variable $\mathbf{x}$; (ii) theoretical properties of the local smoothing based conditional PDF and derivatives estimators; (iii) additional details on bandwidth selection; (iv) alternative covariance function estimators. Last but not least, we provide a general purpose `R` software package (`lpcde`) implementing the main results in this paper.

1.1. Related literature

Our paper contributes to the literature on kernel-based conditional density estimation and inference. See Hall, Wolff and Yao [22], De Gooijer and Zerom [10] and Hall, Racine and Li [21] for earlier reviews, and Wand and Jones [29], Wasserman [30], Simonoff [27] and Scott [26] for textbook introductions. Traditional methods for conditional density estimation typically employ ratios of unconditional kernel density estimators, nonlinear kernel-based derivative of distribution function estimators, or local polynomial estimators based on some preliminary density-like approximation. In the leading special case of $\vartheta = 0$, the closest antecedent to our proposed conditional density estimator is the local polynomial conditional density estimator introduced by Fan, Yao and Tong [15], which is formed by a local polynomial regression of $K_h(y_i; y)$ on $\mathbf{x}_i$. Their estimator is valid at the boundary of $\mathcal{X}$, but is generally inconsistent at the boundary of $\mathcal{Y}$. See Appendix A.1 for more discussion.

More generally, classical methods for conditional density estimation are not boundary adaptive without specific modifications, and in some cases do not have a closed-form representation. Boundary adaptivity could be achieved by employing boundary-corrected kernels in some cases, but such conditional density estimation methods do not appear to have been considered in the literature before. Our first contribution is to introduce a novel boundary adaptive, closed-form conditional density (derivative) estimator. Our proposed construction does not rely on boundary-corrected kernels explicitly, but it rather builds on the idea that automatic boundary-adaptive density estimators can be constructed using local polynomial methods to smooth out the (discontinuous) distribution function [3].

We also consider estimation of conditional CDF, as the intercept in Equation (1) is an estimator of $F(y|\mathbf{x})$, that is, $\mathbf{e}_0'\widehat{\boldsymbol{\beta}}(y|\mathbf{x})$. In addition to being boundary adaptive, this CDF estimator is also continuous in y and $\mathbf{x}$. We discuss properties of this estimator (probability concentration, strong approximation, etc.) in the supplementary material. To compare, the conditional CDF estimator $\widehat{F}(y|\mathbf{x})$, which is constructed via a local polynomial regression of the indicators $\mathbb{1}(y_i \le y)$ on $\mathbf{x}_i$, is generally discontinuous in y. Properties of $\widehat{F}(y|\mathbf{x})$, such as the uniform convergence rate, have been studied in the literature [12,16].

1.2. Notation and assumptions

To simplify the presentation, in the remainder of this paper we set L to be the product kernel based on K: $L(\mathbf{x}) = K(x_1)K(x_2)\cdots K(x_d)$ for a vector $\mathbf{x} = (x_1, \ldots, x_d)'$. We also employ the same bandwidth, $b = h$, in the construction of our proposed estimator, and assume $\mathfrak{q} = \mathfrak{p} - \vartheta - 1 \ge 0$ throughout.

For two numbers a and b, let $a \vee b = \max\{a, b\}$. Limits are taken with respect to the sample size tending to infinity (i.e., $n \to \infty$). For two positive sequences a_n and b_n, $a_n \precsim b_n$ means that a_n/b_n is bounded and $a_n \precsim_{\mathbb{P}} b_n$ means that a_n/b_n is bounded in probability. Constants that do not depend on the sample size or the bandwidth will be denoted by $\mathfrak{c}$, $\mathfrak{c}_1$, $\mathfrak{c}_2$, etc.

We introduce the notation $\precsim_{\mathtt{TC}}$, which not only provides an asymptotic order in probability, but also controls the tail probability (TC): $a_n \precsim_{\mathtt{TC}} b_n$ implies that for any $\mathfrak{c}_1 > 0$, there exists some $\mathfrak{c}_2$ such that

$$\limsup_{n\to\infty} n^{\mathfrak{c}_1}\, \mathbb{P}\big[a_n \ge \mathfrak{c}_2 b_n\big] < \infty.$$

Finally, let $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ and $\mathbf{Y} = (y_1, \ldots, y_n)'$ be the data matrices. We make the following assumptions on the joint distribution and the kernel function.

Assumption 1 (DGP). (i) $(y_1, \mathbf{x}_1'), \ldots, (y_n, \mathbf{x}_n')$ is a random sample from an absolutely continuous distribution supported on $\mathcal{Y} \times \mathcal{X} = [0,1]^{1+d}$, and the joint Lebesgue density, $f(y, \mathbf{x})$, is continuous and

bounded away from zero on $\mathcal{Y}\times\mathcal{X}$. (ii) $f^{(\mathfrak{p})}(y|\mathbf{x})$ exists and is continuous. (iii) $\partial^{\boldsymbol{\nu}} f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^{\boldsymbol{\nu}}$ exists and is continuous for all $|\boldsymbol{\nu}| = \mathfrak{p}-\vartheta$.

Assumption 2 (Kernel). K is a symmetric, Lipschitz continuous PDF supported on $[-1,1]$.

Setting $\mathcal{Y}\times\mathcal{X} = [0,1]^{1+d}$ is a normalization without loss of generality: all our results generalize to the case that $\mathcal{Y}\times\mathcal{X}$ is a Cartesian product of closed intervals. Since our method is local in nature, all the pointwise properties (discussed in the supplementary material) continue to hold if the support $\mathcal{Y}\times\mathcal{X}$ is unbounded. Statements of uniform properties will also remain valid for compact subsets.

We also follow the literature to classify evaluation points as interior or (near) boundary (for example, Section 2.1.2 of [6]). To be precise, let $\text{Cube}_h(y,\mathbf{x}) = [y-h, y+h]\times[x_1-h, x_1+h]\times\cdots\times[x_d-h, x_d+h]$ be the cube of length $2h$ centered at $(y,\mathbf{x})$. Then $(y,\mathbf{x})$ is interior if $\text{Cube}_h(y,\mathbf{x})\subseteq\mathcal{Y}\times\mathcal{X}$. Otherwise it is called (near) boundary. This classification stems from properties of our estimator: as discussed in Appendix A.4, the equivalent kernel is compactly supported, meaning that the estimator only employs observations in an h-neighborhood of the evaluation point.

2. Main results

This section presents four main theoretical results. First, we provide a stochastic linearization of our estimator (Lemma 1). Based on this representation, we obtain a uniform probability concentration result for $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ (Theorem 1). Next, we obtain valid strong approximation results for the standardized t-process based on $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ (Theorem 2). Finally, we develop a feasible distributional approximation for the suprema of the Studentized t-process (Theorem 3). We obtain a uniform consistency result for an estimator of the covariance function (Lemma 2) to establish Theorem 3.

2.1. Stochastic linearization and uniform probability concentration

We first define the large-sample limits of the matrices $\widehat{\mathbf{S}}_y$ and $\widehat{\mathbf{S}}_{\mathbf{x}}$:

$$\mathbf{S}_y = \int_{\mathcal{Y}} \mathbf{p}\Big(\frac{u-y}{h}\Big)\frac{1}{h}\mathbf{P}\Big(\frac{u-y}{h}\Big)^{\intercal}\mathrm{d}F_y(u) \qquad \text{and} \qquad \mathbf{S}_{\mathbf{x}} = \int_{\mathcal{X}} \mathbf{q}\Big(\frac{\mathbf{v}-\mathbf{x}}{h}\Big)\frac{1}{h^d}\mathbf{Q}\Big(\frac{\mathbf{v}-\mathbf{x}}{h}\Big)^{\intercal}\mathrm{d}F_{\mathbf{x}}(\mathbf{v}),$$

with F_y and $F_{\mathbf{x}}$ denoting the CDFs of y_i and $\mathbf{x}_i$, respectively. The following uniform stochastic linear representation holds for $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$.

Lemma 1 (Stochastic linearization). *Suppose Assumptions 1 and 2 hold. If* $nh^{1+d}/\log(n)\to\infty$ *and* $h\to 0$*, then*

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\Big|\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x}) - \bar{f}^{(\vartheta)}(y|\mathbf{x})\Big| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{SL}}, \quad \mathtt{r}_{\mathtt{SL}} = h^{\mathfrak{p}-\vartheta} + \frac{\log(n)}{\sqrt{n^2h^{1+2\vartheta+d+(2\vee d)}}},$$

where $\bar{f}^{(\vartheta)}(y|\mathbf{x}) = n^{-1}\sum_{i=1}^n \mathscr{K}^{\circ}_{\vartheta,h}(y_i,\mathbf{x}_i;y,\mathbf{x})$*, and*

$$\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x}) = \frac{1}{h^{1+\vartheta}}\mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\int_{\mathcal{Y}}\Big(\mathbb{1}(a\le u) - F(u|\mathbf{b})\Big)\frac{1}{h}\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}F_y(u)\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{b}-\mathbf{x}}{h}\right)^{\intercal}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}.$$

The proof, given in Appendix A.3, involves showing that the matrices $\widehat{\mathbf{S}}_y$, $\widehat{\mathbf{S}}_{\mathbf{x}}$ and $\widehat{\mathbf{R}}_{y,\mathbf{x}}$ concentrate. $\widehat{\mathbf{S}}_y$ and $\widehat{\mathbf{S}}_{\mathbf{x}}$ concentrate in probability (and $\mathtt{TC}$ sense), uniformly in y and $\mathbf{x}$ respectively, around $\mathbf{S}_y$ and $\mathbf{S}_{\mathbf{x}}$. Characterizing the large-sample behavior of the matrix $\widehat{\mathbf{R}}_{y,\mathbf{x}}$ in (2) requires a little more care, but the end result can be combined with the results for $\widehat{\mathbf{S}}_y$ and $\widehat{\mathbf{S}}_{\mathbf{x}}$ to obtain the uniform stochastic linear representation for $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$.

Lemma 1 implies that the properties of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ are thus governed by the properties of the stochastic linear representation. In Appendix A.4, we first characterize the leading variance of $\bar{f}^{(\vartheta)}(y|\mathbf{x})$ (Lemma 4). Define $\mathsf{V}_\vartheta(y,\mathbf{x}) := \mathbb{V}[\bar{f}^{(\vartheta)}(y|\mathbf{x})]$, then

$$\mathsf{V}_\vartheta(y,\mathbf{x}) = \frac{1}{nh^{1+d+2\vartheta}} f(y|\mathbf{x}) \left(\mathbf{e}_{1+\vartheta}^\intercal \mathbf{S}_y^{-1}\mathbf{T}_y\mathbf{S}_y^{-1}\mathbf{e}_{1+\vartheta}\right)\left(\mathbf{e}_{\mathbf{0}}^\intercal \mathbf{S}_{\mathbf{x}}^{-1}\mathbf{T}_{\mathbf{x}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}\right) + O\left(\frac{1}{nh^{d+2\vartheta}}\right),$$

$$\text{where} \quad \mathbf{T}_y = \iint_{\mathcal{Y}\times\mathcal{Y}} \frac{\min(u_1,u_2)-y}{h}\frac{1}{h^2}\mathbf{P}\Big(\frac{u_1-y}{h}\Big)\mathbf{P}\Big(\frac{u_2-y}{h}\Big)^\intercal \mathrm{d}F_y(u_1)\mathrm{d}F_y(u_2),$$

$$\mathbf{T}_{\mathbf{x}} = \int_{\mathcal{X}} \frac{1}{h^d}\mathbf{Q}\Big(\frac{\mathbf{v}-\mathbf{x}}{h}\Big)\mathbf{Q}\Big(\frac{\mathbf{v}-\mathbf{x}}{h}\Big)^\intercal \mathrm{d}F_{\mathbf{x}}(\mathbf{v}). \tag{3}$$

Based on the stochastic linearization result in Lemma 1 and the above leading variance characterization, we can obtain a pointwise (in y and $\mathbf{x}$) convergence rate of our estimator: $h^{\mathfrak{p}-\vartheta} + 1/\sqrt{nh^{1+d+2\vartheta}}$. In Theorem 1 below we will establish a uniform convergence rate and a probability concentration result.

Appendix A.4 establishes additional important features of $\mathscr{K}^\circ_{\vartheta,h}$, such as boundedness and Lipschitz continuity which will play a crucial role in our strong approximation results. We also bound the uniform covering number for the class of functions formed by varying the evaluation point. This uniform covering number result takes into account the fact that the shape of $\mathscr{K}^\circ_{\vartheta,h}$ changes across different evaluation points. To this end, we provide in Appendix A.10 a generic result on covering number calculation for function classes formed by kernels, which may be of independent interest. This result allows the kernel functions to take different shapes as well as to depend on a range of bandwidths — the latter feature can be useful for establishing consistency and distributional approximation that are uniform in bandwidth (for example, [13]). However, we do not further pursue along this uniform-in-bandwidth direction to avoid obscuring the main message of the paper.

The following theorem gives a uniform probability concentration result for our conditional density and derivative estimator. The proof is in Appendix A.5.

Theorem 1 (Probability concentration). *Suppose Assumptions 1 and 2 hold. If $h \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then*

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x})\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{PC}}, \quad \mathtt{r}_{\mathtt{PC}} = h^{\mathfrak{p}-\vartheta} + \sqrt{\frac{\log(n)}{nh^{1+d+2\vartheta}}}.$$

The $h^{\mathfrak{p}-\vartheta}$ in Theorem 1 stems from a bias term whose magnitude coincides with that of the pointwise bias at interior evaluation points. As a consequence, the theorem implies that the estimator is boundary adaptive. The other term represents "noise," whose magnitude is larger than its counterpart in Lemma 1, reflecting the fact that the estimation error $\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x})$ can be characterized by the bias and the randomness in $\bar{f}^{(\vartheta)}(y|\mathbf{x})$. By setting $h = (\log(n)/n)^{\frac{1}{1+d+2\mathfrak{p}}}$, it follows from the theorem that the estimator achieves the minimax optimal uniform convergence rate [23], namely $(\log(n)/n)^{\frac{\mathfrak{p}-\vartheta}{1+d+2\mathfrak{p}}}$.

2.2. Strong approximation

We study the distributional properties of the standardized process $\widehat{\mathbb{S}}_\vartheta(y,\mathbf{x})$:

$$\widehat{\mathbb{S}}_\vartheta(y,\mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x})}{\sqrt{\mathsf{V}_\vartheta(y,\mathbf{x})}}. \tag{4}$$

Using elementary tools, Theorem 2.1 in the supplementary material obtains a pointwise Gaussian approximation to $\widehat{\mathbb{S}}_\vartheta(y,\mathbf{x})$. However, the process $\widehat{\mathbb{S}}_\vartheta$ is not asymptotically tight and hence it does not converge weakly to a Gaussian process in $\ell^\infty(\mathcal{Y}\times\mathcal{X})$, the set of uniformly bounded real-valued functions on $\mathcal{Y}\times\mathcal{X}$ equipped with the uniform norm [18,28]. To obtain a uniform distributional approximation, we use the result of Rio [25] and establish a strong approximation result for $(\widehat{\mathbb{S}}_\vartheta(y,\mathbf{x}) : y\in\mathcal{Y}, \mathbf{x}\in\mathcal{X})$. To state the result, define the correlation function

$$\rho_\vartheta(y,\mathbf{x},y',\mathbf{x}') = \mathsf{C}_\vartheta(y,\mathbf{x},y',\mathbf{x}') \Big/ \sqrt{\mathsf{V}_\vartheta(y,\mathbf{x})\mathsf{V}_\vartheta(y',\mathbf{x}')},$$

where $\mathsf{C}_\vartheta(y,\mathbf{x},y',\mathbf{x}') = n^{-1}\mathbb{E}[\mathscr{K}^\circ_{\vartheta,h}(y_i,\mathbf{x}_i;y,\mathbf{x})\mathscr{K}^\circ_{\vartheta,h}(y_i,\mathbf{x}_i;y',\mathbf{x}')]$.

Theorem 2 (Strong approximation). *Suppose Assumptions 1 and 2 hold. If $nh^{1+d+2\mathfrak{p}}\to 0$ and if $nh^{1+d}/\log(n)\to\infty$, then there exist two stochastic processes, $\widehat{\mathbb{S}}'_\vartheta$ and $\mathbb{G}_\vartheta$, in a possibly enlarged probability space, such that:*

(i) *$\widehat{\mathbb{S}}_\vartheta$ and $\widehat{\mathbb{S}}'_\vartheta$ have the same distribution,*
(ii) *$\mathbb{G}_\vartheta$ is a centered Gaussian process with unit variance and correlation ρ_ϑ;*
(iii) *the following holds:*

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\widehat{\mathbb{S}}'_\vartheta(y,\mathbf{x}) - \mathbb{G}_\vartheta(y,\mathbf{x})\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{SA}}, \quad \mathtt{r}_{\mathtt{SA}} = \sqrt{nh^{1+d+2\mathfrak{p}}} + \left(\frac{\log^{1+d}(n)}{nh^{1+d}}\right)^{\frac{1}{2+2d}}.$$

The theorem provides a Gaussian approximation for the entire stochastic process $\widehat{\mathbb{S}}_\vartheta$ rather than for a particular functional thereof. Later we will employ this result to approximate the distribution of the suprema of the process, based on which uniform confidence bands can be constructed.

2.3. Variance-covariance estimation and suprema approximation

Because both the process $\widehat{\mathbb{S}}_\vartheta$ and the correlation function ρ_ϑ depend on unknown features of the underlying data generating process (namely, the covariance function C_ϑ), Theorem 2 in isolation cannot be used for inference. In this subsection we first propose an estimator of the covariance function, and then demonstrate how to obtain a feasible distributional approximation for the suprema of the Studentized t-process.

The covariance function C_ϑ can be expressed as a functional of two unknowns: the conditional CDF of y_i given $\mathbf{x}_i$ and the marginal CDF of y_i. Replacing $F(y|\mathbf{x})$ and $F_y(y)$ with $\widehat{F}(y|\mathbf{x})$ and $\widehat{F}_y(y) = n^{-1}\sum_{i=1}^n \mathbb{1}(y_i\le y)$, respectively, we obtain the following plug-in covariance function estimator:

$$\widehat{\mathsf{C}}_\vartheta(y,\mathbf{x},y',\mathbf{x}') = \frac{1}{n^2}\sum_{i=1}^n \widehat{\mathscr{K}^\circ_{\vartheta,h}}\Big(y_i,\mathbf{x}_i;y,\mathbf{x}\Big)\widehat{\mathscr{K}^\circ_{\vartheta,h}}\Big(y_i,\mathbf{x}_i;y',\mathbf{x}'\Big),$$

where

$$\widehat{\mathscr{K}}^{\circ}_{\vartheta,h}\Big(a,\mathbf{b};y,\mathbf{x}\Big) = \frac{1}{h^{1+\vartheta}}\mathbf{e}_{1+\vartheta}^{\mathsf{T}}\widehat{\mathbf{S}}_y^{-1}\left[\frac{1}{n}\sum_{j=1}^{n}\Big(\mathbb{1}(a\le y_j)-\widehat{F}(y_j|\mathbf{b})\Big)\frac{1}{h}\mathbf{P}\Big(\frac{y_j-y}{h}\Big)\right]\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{b}-\mathbf{x}}{h}\right)^{\mathsf{T}}\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}.$$

The corresponding estimators of V_ϑ and ρ_ϑ are given by $\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x}) = \widehat{\mathsf{C}}_\vartheta(y,\mathbf{x},y,\mathbf{x})$ and

$$\widehat{\rho}_\vartheta(y,\mathbf{x},y',\mathbf{x}') = \widehat{\mathsf{C}}_\vartheta(y,\mathbf{x},y',\mathbf{x}')\Big/\sqrt{\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})\widehat{\mathsf{V}}_\vartheta(y',\mathbf{x}')}.$$

Lemma 2 establishes a uniform probability concentration result for $\widehat{\mathsf{V}}_\vartheta$ and $\widehat{\rho}_\vartheta$. We relegate the proof to the supplementary material as it is quite involved.

Lemma 2 (Covariance estimation). *Suppose Assumptions 1 and 2 hold. If $h\to 0$ and if $nh^{1+d}/\log(n)\to\infty$, then*

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\frac{\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})-\mathsf{V}_\vartheta(y,\mathbf{x})}{\mathsf{V}_\vartheta(y,\mathbf{x})}\right| \lesssim_{\mathtt{TC}} \mathtt{r}_{\mathtt{VE}},\quad \sup_{y,y'\in\mathcal{Y},\mathbf{x},\mathbf{x}'\in\mathcal{X}}\Big|\widehat{\rho}_\vartheta(y,\mathbf{x},y',\mathbf{x}')-\rho_\vartheta(y,\mathbf{x},y',\mathbf{x}')\Big| \lesssim_{\mathtt{TC}} \mathtt{r}_{\mathtt{VE}},$$

$$\textit{where } \mathtt{r}_{\mathtt{VE}} = h^{\mathfrak{p}-\vartheta-\frac{1}{2}}+\sqrt{\frac{\log(n)}{nh^{1+d}}}.$$

With a valid covariance (and variance) estimator, we replacing $\mathsf{V}_\vartheta(y,\mathbf{x})$ with $\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})$ in (4) to obtain the Studentized t-process,

$$\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}(y|\mathbf{x})-f^{(\vartheta)}(y|\mathbf{x})}{\sqrt{\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})}}.$$

By Theorem 2 and Lemma 2, the law of $(\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x}) : y\in\mathcal{Y},\mathbf{x}\in\mathcal{X})$ can be approximated by that of a centered Gaussian process with unit variance and correlation function ρ_ϑ, where the latter is estimated by $\widehat{\rho}_\vartheta$. As a consequence, functionals of $\widehat{\mathbb{T}}_\vartheta$ admit feasible distributional approximations. To illustrate this general phenomenon, the following theorem gives a result for the supremum of $\big|\widehat{\mathbb{T}}_\vartheta\big|$. We define $\widehat{\mathbb{G}}_\vartheta$ as a process whose law, conditional on the data, is a centered Gaussian with unit variance and correlation function $\widehat{\rho}_\vartheta$.

Theorem 3 (Kolmogorov-Smirnov distance: suprema). *Suppose Assumptions 1 and 2 hold. If $n\log(n)h^{1+d+2\mathfrak{p}}\to 0$ and if $nh^{1+d}/\log(n)\to\infty$, then*

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\big|\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x})\big|\le u\Big]-\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\big|\widehat{\mathbb{G}}_\vartheta(y,\mathbf{x})\big|\le u\Big|\mathbf{X},\mathbf{Y}\Big]\right| \lesssim_{\mathbb{P}} \mathtt{r}_{\mathtt{KS}}$$

$$\textit{where } \mathtt{r}_{\mathtt{KS}} = \sqrt{n\log(n)h^{1+d+2\mathfrak{p}}}+\left(\frac{\log^{2+2d}(n)}{nh^{1+d}}\right)^{\frac{1}{2+2d}}+\left(\frac{\log^5(n)}{nh^{1+d}}\right)^{\frac{1}{4}}.$$

To compare the rate of distributional approximation with existing results, we follow the literature and ignore the first (smoothing bias) term. Then, the rate matches what Chernozhukov, Chetverikov and Kato [8] obtained when $d=2$ (see their Remark 3.1(ii)), but it is strictly faster when $d=1$.

3. Applications

This section illustrates our theoretical and methodological results by means of three applications. Before turning to these applications, we discuss bandwidth selection, a necessary step for implementation. It is customary to select the bandwidth by minimizing an approximation to the IMSE of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. Employing Lemma 1 and assuming that $\mathfrak{p}-\vartheta$ is even (as outlined in the local polynomial regression literature [14]), we propose to select the bandwidth by minimizing a feasible analogue of the integrated mean squared error (IMSE)

$$h_{\mathfrak{p}}^{\star} = \operatorname*{argmin}_{h>0} \iint_{\mathcal{Y}\times\mathcal{X}} \left(h^{2\mathfrak{p}-2\vartheta} B_{\vartheta}(y,\mathbf{x})^2 + \frac{1}{nh^{1+2\vartheta+d}} V_{\vartheta}(y,\mathbf{x}) \right) \mathrm{d}y\mathrm{d}\mathbf{x},$$

where $B_{\vartheta}(y,\mathbf{x})$ and $V_{\vartheta}(y,\mathbf{x})$ are the constants in the leading bias and variance, respectively, defined as

$$B_{\vartheta}(y,\mathbf{x}) = f^{(\mathfrak{p})}(y|\mathbf{x})\mathbf{e}_{1+\vartheta}^{\mathsf{T}}\mathbf{S}_y^{-1}\mathbf{c}_{y,\mathfrak{p}+1} + \sum_{|\boldsymbol{\nu}|=\mathfrak{p}-\vartheta} \frac{\partial^{\boldsymbol{\nu}}}{\partial \mathbf{x}^{\boldsymbol{\nu}}} f^{(\vartheta)}(y|\mathbf{x})\mathbf{e}_{\mathbf{0}}^{\mathsf{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{c}_{\mathbf{x},\boldsymbol{\nu}},$$

$$V_{\vartheta}(y,\mathbf{x}) = f(y|\mathbf{x})\Big(\mathbf{e}_{1+\vartheta}^{\mathsf{T}}\mathbf{S}_y^{-1}\mathbf{T}_y\mathbf{S}_y^{-1}\mathbf{e}_{1+\vartheta}\Big)\Big(\mathbf{e}_{\mathbf{0}}^{\mathsf{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{T}_{\mathbf{x}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}\Big),$$

with

$$\mathbf{c}_{y,\mathfrak{p}+1} = \int_{\mathcal{Y}} \frac{1}{(\mathfrak{p}+1)!}\Big(\frac{u-y}{h}\Big)^{\mathfrak{p}+1}\frac{1}{h}\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}F_y(u), \quad \mathbf{c}_{\mathbf{x},\boldsymbol{\nu}} = \int_{\mathcal{X}} \frac{1}{\boldsymbol{\nu}!}\Big(\frac{\mathbf{v}-\mathbf{x}}{h}\Big)^{\boldsymbol{\nu}}\frac{1}{h^d}\mathbf{Q}\Big(\frac{\mathbf{v}-\mathbf{x}}{h}\Big)\mathrm{d}F_{\mathbf{x}}(\mathbf{v}).$$

Both $B_{\vartheta}(y|\mathbf{x})$ and $V_{\vartheta}(y|\mathbf{x})$ involve the conditional PDF and its derivatives, which can be estimated with our proposed method. Other unknown quantities in the IMSE expression have the sample analogues:

$$\widehat{\mathbf{c}}_{y,\mathfrak{p}+1} = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{(\mathfrak{p}+1)!}\left(\frac{y_i-y}{h}\right)^{\mathfrak{p}+1}\mathbf{P}\left(\frac{y_i-y}{h}\right)^{\mathsf{T}}, \quad \hat{\mathbf{c}}_{\mathbf{x},\boldsymbol{\nu}} = \frac{1}{nh^d}\sum_{i=1}^{n}\frac{1}{\boldsymbol{\nu}!}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right)^{\boldsymbol{\nu}}\mathbf{Q}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right),$$

$$\widehat{\mathbf{T}}_y = \frac{1}{n^2h^3}\sum_{i,j=1}^{n}\big(\min(y_i,y_j)-y\big)\mathbf{P}\Big(\frac{y_i-y}{h}\Big)\mathbf{P}\Big(\frac{y_j-y}{h}\Big)^{\mathsf{T}}, \quad \widehat{\mathbf{T}}_{\mathbf{x}} = \frac{1}{nh^d}\sum_{i=1}^{n}\mathbf{Q}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right)\mathbf{Q}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right)^{\mathsf{T}}.$$

The bandwidth that minimizes the approximate IMSE, $h_{\mathfrak{p}}^{\star}$, is proportional to $n^{-\frac{1}{1+d+2\mathfrak{p}}}$. Although this bandwidth delivers estimates that are approximately IMSE-optimal, a non-vanishing bias will be present in their asymptotic distribution, complicating statistical inference. To address this well-known problem, our construction of confidence bands and test statistics for parametric or shape restrictions employs robust bias correction [1,2]: one first constructs an IMSE-optimal point estimator, and then bias corrects the estimator and adjust the covariance function estimator accordingly to obtain a valid distributional approximation. More precisely, given an IMSE-optimal point estimator $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$, robust bias correction relies on a test statistic of the form

$$\frac{\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - \widehat{\mathsf{Bias}}\big[\widehat{f}^{(\vartheta)}(y|\mathbf{x})\big]}{\sqrt{\widehat{\mathsf{Var}}\Big[\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - \widehat{\mathsf{Bias}}\big[\widehat{f}^{(\vartheta)}(y|\mathbf{x})\big]\Big]}},$$

where $\widehat{\mathsf{Bias}}[\widehat{f}^{(\vartheta)}(y|\mathbf{x})]$ denotes a bias correction estimate of the IMSE-optimal point estimator $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$, and $\widehat{\mathsf{Var}}\big[\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - \widehat{\mathsf{Bias}}[\widehat{f}^{(\vartheta)}(y|\mathbf{x})]\big]$ denotes an estimator of the variance of the bias-corrected estimate. The key idea underlying robust bias correction is to Studentize by the variance of the

bias corrected estimate as opposed to by the variance of the original point estimator, an approach that leads to better distributional approximations [1,2]. Similarly, uniform robust bias correction constructs an estimator of the correlation function $\rho_\vartheta(y,\mathbf{x},y',\mathbf{x}')$ taking into account the additional variability introduced by the bias correction.

A simple and intuitive way of operationalizing robust bias correction in local polynomial settings is by increasing the polynomial order $\mathfrak{p}$ (recall that we set $\mathfrak{q}=\mathfrak{p}-\vartheta-1$). That is, we first compute the bandwidth $h^\star_\mathfrak{p}$, and then form the final estimator with a local polynomial order of $\mathfrak{p}+1$. To make the procedure precise, we augment the notation so that it reflects the local polynomial order and the bandwidth used as needed. For example, the conditional density estimator using polynomial order $\mathfrak{p}$ and employing the bandwidth h is written as $\widehat{f}^{(\vartheta)}_\mathfrak{p}(y|\mathbf{x};h)$.

3.1. Confidence bands

Confidence bands can be constructed using the process $(\widehat{\mathbb{T}}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}) : y\in\mathcal{Y},\mathbf{x}\in\mathcal{X})$, where

$$\widehat{\mathbb{T}}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})=\frac{\widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x};h^\star_\mathfrak{p})-f^{(\vartheta)}(y|\mathbf{x})}{\sqrt{\widehat{\mathrm{V}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x};h^\star_\mathfrak{p})}},$$

By Theorem 3, the distribution of $\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{T}}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})|$ is approximated by the conditional (on the data) distribution of $\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{G}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})|$, with $\widehat{\mathbb{G}}_{\vartheta,\mathfrak{p}+1}$ being a centered Gaussian process whose law, conditionally on the data, is Gaussian with unit variance and correlation $\widehat{\rho}_{\vartheta,\mathfrak{p}+1}(\cdot;h^\star_\mathfrak{p})$. Accordingly, let

$$\mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha)=\left[\ \widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x};h^\star_\mathfrak{p})\ \pm\ \mathtt{cv}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha)\sqrt{\widehat{\mathrm{V}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x};h^\star_\mathfrak{p})}\ :\ y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}\right],$$

where

$$\mathtt{cv}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha)=\inf\left\{u\in\mathbb{R}_+:\mathbb{P}\left[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\widehat{\mathbb{G}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})\right|\le u\,\middle|\,\mathbf{X},\mathbf{Y}\right]\ge 1-\alpha\right\}.$$

As the notation suggests, $\mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha)$ is a $100(1-\alpha)\%$ confidence band. To be specific, we have the following theorem.

Theorem 4 (Confidence bands). *Suppose Assumptions 1 and 2 hold, $f^{(\mathfrak{p}+1)}(y|\mathbf{x})$ exists and is continuous, and $\partial^\nu f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^\nu$ exists and is continuous for all $|\nu|=\mathfrak{p}+1-\vartheta$. Then*

$$\left|\mathbb{P}\left[f^{(\vartheta)}\in\mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha)\right]-(1-\alpha)\right|\precsim\log^{\frac{5}{4}}(n)\mathtt{r}_{\mathtt{CB}},$$

where $\mathtt{r}_{\mathtt{CB}}=n^{-\frac{1}{1+d+2\mathfrak{p}}}+n^{-\frac{2\mathfrak{p}-2\vartheta+1}{4(1+d+2\mathfrak{p})}}+n^{-\frac{\mathfrak{p}}{(1+d+2\mathfrak{p})(1+d)}}$.

The confidence band $\mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha)$ is easy to construct because, by discretizing the index set of the Gaussian process, the critical value $\mathtt{cv}_{\vartheta,\mathfrak{p}+1}(1-\alpha)$ can be computed by simulation from a conditionally (on the data) multivariate Gaussian distribution. We illustrate the performance of our proposed confidence bands using simulated and real data in Section 5.

Theorem 4 provides a formal, theoretical justification for employing strong approximation methods to construct confidence bands instead of relying on extreme value theory for approximating the

distribution of the suprema of the process $\widehat{\mathbb{T}}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}$. More specifically, the coverage error rate $\mathtt{r}_{\mathtt{CB}}$ is polynomial in n for the former inference approach, while the latter inference approach would have a logarithmic in n convergence rate [see, e.g., 19,20, and references therein]. The same remark applies to the upcoming Theorems 5 and 6, which characterize the error in rejection probability of two different classes of hypothesis testing procedures.

3.2. Parametric specification testing

Suppose the researcher postulates that the conditional density (derivative) belongs to the parametric class $\{f^{(\vartheta)}(y|\mathbf{x};\boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \Gamma_\vartheta\}$, where Γ_ϑ is some parameter space. Abstracting away from the specifics of the estimation technique, we assume that the researcher also picks some estimator $\widehat{\boldsymbol{\gamma}}$ (e.g., maximum likelihood or minimum distance), which is assumed to converge in probability to some $\bar{\boldsymbol{\gamma}} \in \Gamma_\vartheta$. A natural statistic for the problem of testing

$$\mathsf{H}_0^{\mathrm{PS}} : \ f^{(\vartheta)}(y|\mathbf{x};\bar{\boldsymbol{\gamma}}) = f^{(\vartheta)}(y|\mathbf{x}) \qquad \text{for all } (y,\mathbf{x}) \in \mathcal{Y}\times\mathcal{X}$$

is

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{\mathbb{T}}^{\mathrm{PS}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})\right|, \qquad \widehat{\mathbb{T}}^{\mathrm{PS}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x};h^\star_{\mathfrak{p}}) - f^{(\vartheta)}(y|\mathbf{x};\widehat{\boldsymbol{\gamma}})}{\sqrt{\widehat{\mathrm{V}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x};h^\star_{\mathfrak{p}})}}.$$

Assuming the estimation error of $\widehat{\boldsymbol{\gamma}}$ is asymptotically negligible, a valid $100\alpha\%$ critical value is given by $\mathtt{cv}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha)$. To be specific, we have:

Theorem 5 (Parametric specification testing). *Suppose Assumptions 1 and 2 hold, $f^{(\mathfrak{p}+1)}(y|\mathbf{x})$ exists and is continuous, and $\partial^{\boldsymbol{\nu}} f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^{\boldsymbol{\nu}}$ exists and is continuous for all $|\boldsymbol{\nu}| = \mathfrak{p}+1-\vartheta$. If*

$$n^{\frac{\mathfrak{p}-\vartheta}{1+d+2\mathfrak{p}}} \sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|f^{(\vartheta)}(y|\mathbf{x};\widehat{\boldsymbol{\gamma}}) - f^{(\vartheta)}(y|\mathbf{x};\bar{\boldsymbol{\gamma}})\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{CB}},$$

then, under $\mathsf{H}_0^{\mathrm{PS}}$,

$$\left|\mathbb{P}\left[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} |\widehat{\mathbb{T}}^{\mathrm{PS}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})| > \mathtt{cv}^{\mathrm{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha)\right] - \alpha\right| \precsim \log^{\frac{5}{4}}(n)\,\mathtt{r}_{\mathtt{CB}},$$

where $\mathtt{r}_{\mathtt{CB}}$ is defined in Theorem 4.

3.3. Testing shape restrictions

As a third application, suppose the researcher wants to test shape restrictions on $f^{(\vartheta)}$. Letting c_ϑ be a pre-specified function, consider the problem of testing

$$\mathsf{H}_0^{\mathrm{SR}} : f^{(\vartheta)}(y|\mathbf{x}) \le c_\vartheta(y|\mathbf{x}) \qquad \text{for all } (y,\mathbf{x}) \in \mathcal{Y}\times\mathcal{X}.$$

For example, if $\vartheta = 0$ and if $c_\vartheta(y|\mathbf{x})$ is some (positive) constant value c, the testing problem refers to whether the conditional density exceeds c somewhere on its support. As another example, if $\vartheta = 1$ and if $c_\vartheta(y|\mathbf{x}) = 0$, then the testing problem refers to whether the conditional density is non-increasing in

y for all values of $\mathbf{x}$. More generally, the testing problem above can be used to test for monotonicity, convexity, and other shape features of the conditional density, possibly relative to the function $c_\vartheta(y|\mathbf{x})$.

A natural testing procedure rejects $\mathsf{H}_0^{\mathrm{SR}}$ whenever the test statistic

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \mathbb{T}^{\mathrm{SR}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}), \qquad \mathbb{T}^{\mathrm{SR}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x};h^\star_{\mathfrak{p}}) - c_\vartheta(y|\mathbf{x})}{\sqrt{\widehat{\mathrm{V}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x};h^\star_{\mathfrak{p}})}}$$

exceeds a critical value of the form

$$\mathtt{cv}^{\mathrm{SR}}_{\vartheta,\mathfrak{p}+1}(\alpha) = \inf\Big\{u\in\mathbb{R}_+ : \mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \widehat{\mathbb{G}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}) \le u \,\Big|\, \mathbf{X},\mathbf{Y}\Big] \ge 1-\alpha\Big\}.$$

Theorem 6 (Testing shape restriction). *Suppose Assumptions 1 and 2 hold, $f^{(\mathfrak{p}+1)}(y|\mathbf{x})$ exists and is continuous, and $\partial^{\boldsymbol{\nu}} f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^{\boldsymbol{\nu}}$ exists and is continuous for all $|\boldsymbol{\nu}| = \mathfrak{p}+1-\vartheta$. Then, under $\mathsf{H}_0^{\mathrm{SR}}$,*

$$\Big|\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \widehat{\mathbb{T}}^{\mathrm{SR}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}) > \mathtt{cv}^{\mathrm{SR}}_{\vartheta,\mathfrak{p}+1}(\alpha)\Big] - \alpha\Big| \precsim \log^{\frac{5}{4}}(n)\,\mathtt{r}_{\mathtt{CB}},$$

where $\mathtt{r}_{\mathtt{CB}}$ is defined in Theorem 4.

4. Imposing additional constraints for density estimation

Specific applications may require additional constraints on the estimates. For example, setting $\vartheta = 0$ (PDF), it may be desirable to require that the estimator is non-negative and integrates to one. The nonnegativity constraint can be directly incorporated into the local polynomial regression (1):

$$\widehat{f}_{\mathtt{N}}(y|\mathbf{x}) = \mathbf{e}_1^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\mathtt{N}}(y|\mathbf{x}), \qquad \widehat{\boldsymbol{\beta}}_{\mathtt{N}}(y|\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^{\mathfrak{p}+1}:\ \mathbf{e}_1^{\mathsf{T}}\mathbf{u}\ge 0} \sum_{i=1}^n \Big(\widehat{F}(y_i|\mathbf{x}) - \mathbf{p}(y_i-y)^{\mathsf{T}}\mathbf{u}\Big)^2 K_h(y_i;y),$$

where the subscript "$\mathtt{N}$" stands for "non-negative." While $\widehat{f}_{\mathtt{N}}(y|\mathbf{x})$ is non-negative by construction, it does not necessarily integrate to one. This follows from the fact that the estimator only exploits local features of the data and not global constraints. To address the second constraint, we propose and study the following enhanced estimator based on minimizing Kullback-Leibler divergence (the subscript "$\mathtt{I}$" stands for "integrating to one"):

$$\widehat{f}_{\mathtt{I}}(y|\mathbf{x}) = \operatorname*{argmin}_{g\in\mathcal{G}} \mathrm{KL}\big(g \,\big\|\, \widehat{f}_{\mathtt{N}}(\cdot|\mathbf{x})\big), \qquad \text{where } \mathrm{KL}(g\,\|\,f) = \int_{\mathcal{Y}} g(y)\log\Big(\frac{g(y)}{f(y)}\Big)\mathrm{d}y,$$

and $\mathcal{G} = \{g\ge 0 : \int_{\mathcal{Y}} g(y)\mathrm{d}y = 1,\ g(y) = 0 \text{ for } y\notin\mathcal{Y}\}$. It follows that our proposed conditional PDF estimator, $\widehat{f}_{\mathtt{I}}(y|\mathbf{x})$, is non-negative and integrates to one. Furthermore, both $\widehat{f}_{\mathtt{N}}(y|\mathbf{x})$ and $\widehat{f}_{\mathtt{I}}(y|\mathbf{x})$ can be written in closed form (see Appendix A.8):

$$\widehat{f}_{\mathtt{I}}(y|\mathbf{x}) = \frac{\widehat{f}_{\mathtt{N}}(y|\mathbf{x})}{\int_{\mathcal{Y}} \widehat{f}_{\mathtt{N}}(u|\mathbf{x})\mathrm{d}u} \qquad \text{and} \qquad \widehat{f}_{\mathtt{N}}(y|\mathbf{x}) = \max\big\{\widehat{f}(y|\mathbf{x})\ ,\ 0\big\}. \tag{5}$$

In practice, the support $\mathcal{Y}$ might be unknown, and in this case one can naturally replace it by the empirical support: $\widehat{\mathcal{Y}} = [y_{(1)}, y_{(n)}]$, defined by the smallest $(y_{(1)})$ and largest $(y_{(n)})$ order statistics of

the observed $y_1, y_2, \ldots, y_n$. Since $\widehat{\mathcal{Y}} \subseteq \mathcal{Y}$, all the theoretical results discussed below remain valid on the empirical support $\widehat{\mathcal{Y}}$.

We first establish stochastic linearization for both, $\widehat{f}_{\mathrm{N}}(y|\mathbf{x})$ and $\widehat{f}_{\mathrm{I}}(y|\mathbf{x})$.

Lemma 3 (Stochastic linearization). *Suppose Assumptions 1 and 2 hold. If $nh^{1+d}/\log(n) \to \infty$ and $h \to 0$, then*

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{f}_{\mathrm{N}}(y|\mathbf{x}) - f(y|\mathbf{x}) - \bar{f}^{(0)}(y|\mathbf{x})\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{SL}},$$

and

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{f}_{\mathrm{I}}(y|\mathbf{x}) - f(y|\mathbf{x}) - \left(\bar{f}^{(0)}(y|\mathbf{x}) - f(y|\mathbf{x})\int_{\mathcal{Y}} \bar{f}^{(0)}(u|\mathbf{x})\mathrm{d}u\right)\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{SL}},$$

where $\bar{f}^{(0)}(y|\mathbf{x})$ and $\mathtt{r}_{\mathtt{SL}}$ are defined in Lemma 1 by setting $\vartheta = 0$.

The lemma provides a more refined stochastic linearization for $\widehat{f}_{\mathrm{I}}(y|\mathbf{x})$. We will show that the normalization in $\widehat{f}_{\mathrm{I}}(y|\mathbf{x})$ does not affect the uniform rate of convergence of the estimator. For distributional approximation, however, it is crucial to employ different Gaussian processes for the two estimators. In particular, we show that failing to capture the asymptotic contribution of the normalization in $\widehat{f}_{\mathrm{I}}(y|\mathbf{x})$ may lead to a slower rate for strong approximation.

Theorem 7 (Probability concentration). *Suppose Assumptions 1 and 2 hold. If $h \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then*

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{f}_{\mathrm{N}}(y|\mathbf{x}) - f(y|\mathbf{x})\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{PC}}, \quad \sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{f}_{\mathrm{I}}(y|\mathbf{x}) - f(y|\mathbf{x})\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{PC}},$$

where $\mathtt{r}_{\mathtt{PC}}$ is defined in Theorem 1 (with $\vartheta = 0$).

Finally, to state a strong approximation result, we define the following standardized processes

$$\widehat{\mathbb{S}}_{\mathrm{N}}(y,\mathbf{x}) = \frac{\widehat{f}_{\mathrm{N}}(y|\mathbf{x}) - f(y|\mathbf{x})}{\sqrt{\mathrm{V}_0(y,\mathbf{x})}}, \quad \widehat{\mathbb{S}}_{\mathrm{I}}(y,\mathbf{x}) = \frac{\widehat{f}_{\mathrm{I}}(y|\mathbf{x}) - f(y|\mathbf{x})}{\sqrt{\mathrm{V}_0(y,\mathbf{x})}}.$$

Theorem 8 (Strong approximation). *Suppose Assumptions 1 and 2 hold. If $nh^{1+d+2\mathfrak{p}} \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then there exist three stochastic processes, $\widehat{\mathbb{S}}'_{\mathrm{N}}$, $\widehat{\mathbb{S}}'_{\mathrm{I}}$, and $\mathbb{G}$, in a possibly enlarged probability space, such that:*

(i) *$\widehat{\mathbb{S}}_{\mathrm{N}}$ and $\widehat{\mathbb{S}}'_{\mathrm{N}}$ have the same distribution; $\widehat{\mathbb{S}}_{\mathrm{I}}$ and $\widehat{\mathbb{S}}'_{\mathrm{I}}$ have the same distribution*

(ii) *$\mathbb{G}$ is a centered Gaussian process with unit variance and correlation ρ_0;*

(iii) *the following holds:*

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{\mathbb{S}}'_{\mathrm{N}}(y,\mathbf{x}) - \mathbb{G}(y,\mathbf{x})\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{SA}},$$

and

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{\mathbb{S}}'_{\mathrm{I}}(y,\mathbf{x}) - \left(\mathbb{G}(y,\mathbf{x}) - f(y|\mathbf{x})\int_{\mathcal{Y}} \sqrt{\frac{\mathrm{V}_0(u,\mathbf{x})}{\mathrm{V}_0(y,\mathbf{x})}}\mathbb{G}(u,\mathbf{x})\mathrm{d}u\right)\right| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{SA}},$$

where $\mathtt{r}_{\mathtt{SA}}$ is defined in Theorem 2.

The different Gaussian processes needed for distributional approximation to $\widehat{\mathbb{S}}_{\mathtt{N}}$ and $\widehat{\mathbb{S}}_{\mathtt{I}}$ in Theorem 8 is due to the normalization in $\widehat{\mathbb{S}}_{\mathtt{I}}$. Of course, it is possible to couple $\widehat{\mathbb{S}}'_{\mathtt{I}}$ with $\mathbb{G}$ directly, but a slower rate may arise, particularly $\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} |\widehat{\mathbb{S}}'_{\mathtt{I}} - \mathbb{G}(y,\mathbf{x})| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{SA}} + \sqrt{\log(n)h}$.

Constructing analogues of Lemma 2 and Theorem 3 from Section 2.3 for the constrained estimators $\widehat{f}_{\mathtt{N}}(y|\mathbf{x})$ and $\widehat{f}_{\mathtt{I}}(y|\mathbf{x})$ now follows directly. Additionally, confidence bands and hypothesis testing procedures as in Section 3 can also be easily developed when employing the constrained density estimators. We omit details to avoid repetition.

5. Numerical Evidence

We illustrate the effectiveness of our proposed methods with two Monte Carlo experiments, where we set $d = 1$ and simulate $\mathbf{x}$ and y from a joint normal distribution with variance 2 and covariance -0.1, truncated on $[-1, 1]^2$. We employ 1000 Monte Carlo repetitions, each with the sample size $n = 5000$. Replication files, additional simulation results, and details of the companion `R` package, `lpcde`, can be found at https://nppackages.github.io/lpcde/ and in our companion software article [4].

In the first simulation experiment, we estimate the conditional PDF for 20 equally spaced points on $[-1, 1]$ for y. Table 1 presents the simulation results at three different conditioning values: (a) interior ($\mathbf{x} = 0$), (b) near-boundary ($\mathbf{x} = 0.8$), and (c) at-boundary ($\mathbf{x} = 1$). See the discussion at the end of Section 1 for a classification of interior and (near) boundary evaluation points.

Table 1 reports average estimated bandwidth in column "$\widehat{h}$", and average bias and standard error in the "bias" and "se" columns, respectively. We consider bands formed by pointwise confidence intervals (columns "pointwise"), which are not uniformly valid and hence should exhibit considerable under coverage, as well as the uniform confidence bands discussed in Section 3 (columns "uniform"). We report their empirical uniform coverage probabilities (column "Coverage") and the average width (column "Width"). For the non-bias corrected rows ("NBC"), the polynomial orders for bandwidth selection, point estimation and statistical inference are $\mathfrak{p} = 2$ and $\mathfrak{q} = 1$, while those for robust bias-corrected statistical inference rows ("RBC") are $\mathfrak{p} = 3$ and $\mathfrak{q} = 2$.

Table 1. Empirical uniform coverage probabilities.

					Coverage		Width	
		$\widehat{h}$	bias	se	pointwise	uniform	pointwise	uniform
$\mathbf{x} = 0$	NBC	0.32	0.09	0.03	62.6	74.8	0.01	0.02
	RBC	0.32	0.09	0.09	83.4	93.9	0.05	0.05
$\mathbf{x} = 0.8$	NBC	0.30	0.10	0.04	72.8	89.4	0.02	0.03
	RBC	0.30	0.10	0.18	86.9	94.3	0.13	0.19
$\mathbf{x} = 1.0$	NBC	0.32	0.10	0.06	74.9	91.3	0.02	0.05
	RBC	0.32	0.10	0.20	88.1	93.2	0.11	0.23

The simulation results in Table 1 support our main theoretical findings. First, robust bias correction leads to uniformly better performance of the inference procedures, both pointwise and uniformly over $\mathcal{Y}$. Second, our uniform distributional approximation leads to feasible confidence bands with good finite sample performance, when coupled with robust bias correction methods.

Table 2. Comparison between FYT and our method for conditional PDF estimation.

		FYT				This paper: $\widehat{f}(y\|\mathbf{x})$					
	$\times h_{\text{MSE}}$	h	bias	se	rmse	h	bias	se	rmse	NBC CI	RBC CI
$(y=0 \quad, \mathbf{x}=0)$	0.8	0.32	0.08	0.13	0.15	0.23	0.07	0.07	0.11	0.87	0.99
	0.9	0.36	0.09	0.12	0.15	0.26	0.07	0.05	0.09	0.74	0.98
	1	0.39	0.09	0.12	0.15	0.29	0.07	0.04	0.08	0.56	0.96
	1.1	0.43	0.09	0.12	0.15	0.32	0.06	0.03	0.07	0.39	0.89
	1.2	0.46	0.10	0.11	0.15	0.35	0.07	0.02	0.07	0.25	0.81
$(y=0.8, \mathbf{x}=0)$	0.8	0.29	0.14	0.10	0.18	0.26	0.03	0.02	0.04	0.90	0.99
	0.9	0.33	0.14	0.10	0.17	0.30	0.03	0.01	0.03	0.83	0.98
	1	0.36	0.14	0.09	0.17	0.33	0.03	0.01	0.03	0.75	0.93
	1.1	0.39	0.14	0.09	0.17	0.36	0.03	0.01	0.04	0.70	0.87
	1.2	0.43	0.14	0.08	0.16	0.40	0.03	0.01	0.04	0.64	0.80
$(y=1 \quad, \mathbf{x}=0)$	0.8	0.27	0.18	0.07	0.20	0.40	0.04	0.04	0.06	0.93	1.00
	0.9	0.30	0.18	0.07	0.19	0.45	0.04	0.03	0.05	0.73	0.99
	1	0.33	0.20	0.06	0.20	0.50	0.04	0.02	0.04	0.51	0.96
	1.1	0.36	0.21	0.06	0.21	0.55	0.04	0.01	0.04	0.36	0.89
	1.2	0.39	0.23	0.05	0.24	0.60	0.04	0.01	0.04	0.22	0.80

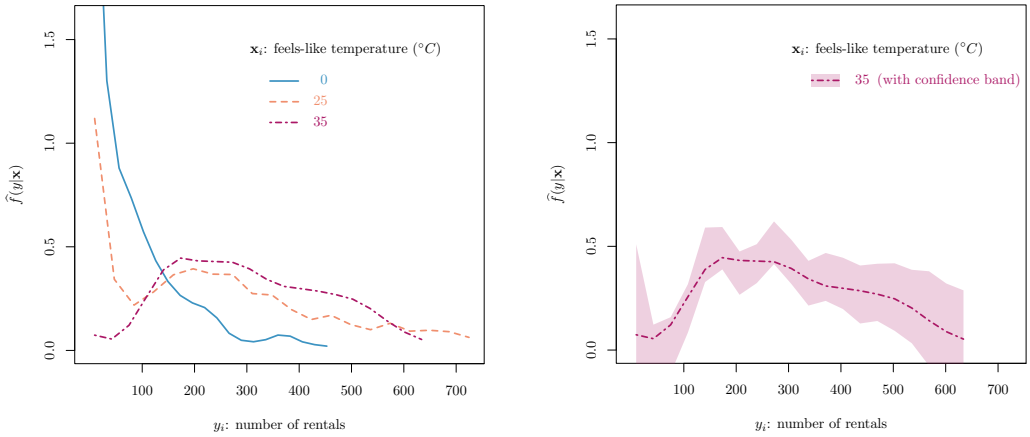
For example, for $\mathbf{x}=0$, the averaged (across simulations) estimated approximate IMSE-optimal bandwidth choice is $\widehat{h}=0.32$, with $\mathfrak{p}=2$ and $\mathfrak{q}=\mathfrak{p}-1$. Bands constructed with pointwise confidence intervals have empirical uniform coverage of 62.6% without bias correction, and 83.4% with robust bias correction, both are substantially below the 95% nominal level because they are not uniformly valid over the range of y. The feasible confidence bands are designed to address that issue: our proposed confidence bands have empirical coverage of 93.9% when robust bias correction is employed. It also highlights the importance of addressing the misspecification (smoothing) bias for statistical inference. Without bias correction, the uniform confidence bands only cover the true conditional PDF with probability 74.8%.

The second simulation study compares our estimator (lpcde) to the estimator proposed by Fan, Yao and Tong [15] (FYT, see Appendix A.1 for details). Table 2 presents the simulation results for the conditional PDF at three distinct evaluation points. For a fair comparison, we first compute the MSE optimal bandwidth (h_{MSE}) for the two estimators at each evaluation point. We then investigate the performance of the two estimators over a grid of bandwidths, ranging from $0.8\times h_{\text{MSE}}$ (under smoothing) to $1.2\times h_{\text{MSE}}$ (over-smoothing).

For each of the two estimators we report the average bandwidth, bias, standard error, and root mean squared error. Additionally, for our estimator we report the pointwise empirical coverage probabilities, both with and without bias correction. Since FYT do not provide theory for statistical inference, we do not report confidence interval information for the estimator. Results in Table 2 suggest that our local polynomial conditional PDF estimator perform well across all three evaluation points, and the confidence intervals constructed thereof exhibits satisfactory empirical coverage property. In particular, at the boundary evaluation point $(y=1,\mathbf{x}=0)$, our estimator has accurate coverage while FYT suffers from boundary bias.

Finally, we illustrate the performance of our estimator in Figure 1 with real data. The data we employ is from Capital Bikeshare (available at https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset). The outcome variable y_i is the total number of bike rentals, and the covariate $\mathbf{x}_i$ is the "feels-like"

temperature in Celsius. Panel (a) shows the estimated conditional PDFs for three temperature levels, $\mathbf{x}_i = 0$, 25, and 35 $^\circ C$. From the conditional density plots, more bike rental activities happen in warmer days (i.e., the conditional distribution moves toward right). It is worth mentioning that the outcome variable has a lower boundary at 0, and using a standard kernel density estimator for conditional PDF estimation will lead to a severe under-estimation bias for $f(y|\mathbf{x})$ whenever the evaluation point y is close to zero. To avoid overcrowding the figure, we illustrate the confidence band with robust bias correction in panel (b).



(a) Estimated conditional PDFs.

(b) Illustration of confidence band.

Figure 1: Estimated relationship (conditional PDF) between bike rental counts and temperature.

6. Conclusion

We introduced a new boundary adaptive estimator of the conditional density and derivatives thereof. This estimator is conceptually distinct from prior proposals in the literature, as it relies on two (nested) local polynomial estimators. Our proposed estimation approach has several appealing features, most notably automatic boundary adaptivity. We provided an array of uniform estimation and distributional results, including a valid uniform equivalent kernel representation and uniform distributional approximations. Our methods are applicable in data science settings either where the conditional density or its derivatives are the main object of interest, or where they are preliminary estimands entering a multi-step statistical procedure.

Appendix

A.1. Derivation of (2) and an alternative expression

To start, the conditional CDF estimation step is a weighted least squares problem, and has the solution

$$\widehat{F}(y_j|\mathbf{x}) = \mathbf{e}_0^\intercal \Big(\sum_{i=1}^n \mathbf{q}(\mathbf{x}_i - \mathbf{x}) \mathbf{q}(\mathbf{x}_i - \mathbf{x})^\intercal L_b(\mathbf{x}_i; \mathbf{x}) \Big)^{-1} \Big(\sum_{i=1}^n \mathbf{q}(\mathbf{x}_i - \mathbf{x}) L_b(\mathbf{x}_i; \mathbf{x}) \mathbb{1}(y_i \leq y_j) \Big).$$

The second local polynomial regression takes $\widehat{F}(y_j|\mathbf{x})$ as the "dependent variable," and therefore the final estimator takes the form

$$\widehat{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\intercal}\Big(\sum_{j=1}^{n}\mathbf{p}(y_j-y)\mathbf{p}(y_j-y)^{\intercal}K_h(y_j;y)\Big)^{-1}\Big(\sum_{j=1}^{n}\mathbf{p}(y_j-y)K_h(y_j;y)\widehat{F}(y_j|\mathbf{x})\Big).$$

The final expression in (2) then follows from re-normalizing $\mathbf{x}_i-\mathbf{x}$ to $(\mathbf{x}_i-\mathbf{x})/b$ and y_j-y to $(y_j-y)/h$, leading to the multiplicative factor $h^{-1-\vartheta}$. By changing the order of summation in $\widehat{\mathbf{R}}_{y,\mathbf{x}}$, we can also write $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ as

$$\widehat{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{\mathbf{0}}^{\intercal}\Big(\sum_{i=1}^{n}\mathbf{q}(\mathbf{x}_i-\mathbf{x})\mathbf{q}(\mathbf{x}_i-\mathbf{x})^{\intercal}L_b(\mathbf{x}_i;\mathbf{x})\Big)^{-1}\Big(\sum_{i=1}^{n}\mathbf{q}(\mathbf{x}_i-\mathbf{x})L_b(\mathbf{x}_i;\mathbf{x})\widehat{K}_h(y_i,y)\Big),$$

where

$$\widehat{K}_h(y_i,y) = \mathbf{e}_{1+\vartheta}^{\intercal}\Big(\sum_{j=1}^{n}\mathbf{p}(y_j-y)\mathbf{p}(y_j-y)^{\intercal}K_h(y_j;y)\Big)^{-1}\Big(\sum_{j=1}^{n}\mathbf{p}(y_j-y)K_h(y_j;y)\mathbb{1}(y_i\le y_j)\Big).$$

The above alternative expression shows that our proposed estimator can be understood as first forming $\widehat{K}_h(y_i,y)$, which is a data-driven kernel re-weighting of y_i and then conducting local polynomial regression on $\mathbf{x}_i$. To compare, the density estimator ($\vartheta=0$) introduced by Fan, Yao and Tong [15] takes the form

$$\widehat{f}_{\mathrm{FYT}}(y|\mathbf{x}) = \mathbf{e}_{\mathbf{0}}^{\intercal}\Big(\sum_{i=1}^{n}\mathbf{q}(\mathbf{x}_i-\mathbf{x})\mathbf{q}(\mathbf{x}_i-\mathbf{x})^{\intercal}L_b(\mathbf{x}_i;\mathbf{x})\Big)^{-1}\Big(\sum_{i=1}^{n}\mathbf{q}(\mathbf{x}_i-\mathbf{x})L_b(\mathbf{x}_i;\mathbf{x})K_h(y_i,y)\Big),$$

where $K_h(y_i,y)=K((y_i-y)/h)/h$ for some (second-order) kernel function K. The estimator, $\widehat{f}_{\mathrm{FYT}}(y|\mathbf{x})$, is consistent at the boundary of $\mathcal{X}$ (due to the local polynomial regression step on $\mathbf{x}_i$), but is generally inconsistent at the boundary of $\mathcal{Y}$. Unlike their proposal, our estimator remains consistent at the boundaries of both $\mathcal{X}$ and $\mathcal{Y}$.

A.2. A local smoothing based estimator

In this appendix we introduce a local smoothing based estimator for the conditional PDF and its derivatives. Recall from Section 1 that $\widehat{F}(y|\mathbf{x})$ is the estimated conditional CDF formed by a $\mathfrak{q}$-th order local polynomial regression. Now let G be some nonnegative measure such that the Radon-Nikodym derivative with respect to the Lebesgue measure is continuous. Then instead of employing a local polynomial regression as in (1), we form a conditional PDF (and derivatives) estimator by local smoothing:

$$\check{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\intercal}\check{\boldsymbol{\beta}}(y|\mathbf{x}), \qquad \check{\boldsymbol{\beta}}(y|\mathbf{x}) = \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^{\mathfrak{p}+1}}\int_{\mathcal{Y}}\Big(\widehat{F}(u|\mathbf{x})-\mathbf{p}(u-y)^{\intercal}\mathbf{v}\Big)^2K_h(u;y)\mathrm{d}G(u),$$

which has the closed-form expression: $\check{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\bar{\mathbf{R}}_{y,\mathbf{x}}\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}$. Here we define

$$\mathbf{S}_y = \int_{\mathcal{Y}}\mathbf{p}\Big(\frac{u-y}{h}\Big)\frac{1}{h}\mathbf{P}\Big(\frac{u-y}{h}\Big)^{\intercal}\mathrm{d}G(u),$$

$$\bar{\mathbf{R}}_{y,\mathbf{x}} = \frac{1}{nh^{1+\vartheta}}\sum_{i=1}^{n}\Big(\int_{\mathcal{Y}}\mathbb{1}(y_i\le u)\frac{1}{h}\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}G(u)\Big)\frac{1}{b^d}\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{b}\Big)^{\intercal}.$$

Compared to $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$, the above local smoothing based estimator requires knowledge of the support $\mathcal{Y}$. On the other hand, $\check{f}^{(\vartheta)}(y|\mathbf{x})$ has the advantage that it is immune to low density regions of y_i; that is, the new estimator remains valid even when the density of y_i is close to zero. Intuitively, this is because $\check{f}^{(\vartheta)}(y|\mathbf{x})$ employs a nonrandom local smoothing in the second step, while $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ is based on two local polynomial regressions.

Due to space limitations, we investigate the theoretical properties of this estimator in the supplementary material [5].

A.3. Proof of Lemma 1

Define

$$u_{i,j} = \left(\Big(\mathbb{1}(y_i \le y_j) - F(y_j|\mathbf{x}_i) \Big) \mathbf{P}\Big(\frac{y_j - y}{h}\Big) - \int_{\mathcal{Y}} \Big[\mathbb{1}(y_i \le u) - F(u|\mathbf{x}_i) \Big] \mathbf{P}\Big(\frac{u-y}{h}\Big) \mathrm{d}F_y(u) \right) \mathbf{Q}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)^{\intercal}.$$

We write

$$\widehat{f}^{(\vartheta)}(y|\mathbf{x}) = \frac{1}{nh^{2+d+\vartheta}} \mathbf{e}_{1+\vartheta}^{\intercal} \widehat{\mathbf{S}}_y^{-1} \left(\sum_{i=1}^{n} \int_{\mathcal{Y}} \Big(\mathbb{1}(y_i \le u) - F(u|\mathbf{x}_i) \Big) \mathbf{P}\Big(\frac{u-y}{h}\Big) \mathrm{d}F_y(u) \mathbf{Q}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)^{\intercal} \right) \widehat{\mathbf{S}}_{\mathbf{x}}^{-1} \mathbf{e_0} \qquad \text{(I)}$$

$$+ \frac{1}{n^2 h^{2+d+\vartheta}} \mathbf{e}_{1+\vartheta}^{\intercal} \widehat{\mathbf{S}}_y^{-1} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} F(y_j|\mathbf{x}_i) \mathbf{P}\Big(\frac{y_j - y}{h}\Big) \mathbf{Q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)^{\intercal} \right) \widehat{\mathbf{S}}_{\mathbf{x}}^{-1} \mathbf{e_0} \qquad \text{(II)}$$

$$+ \frac{1}{n^2 h^{2+d+\vartheta}} \mathbf{e}_{1+\vartheta}^{\intercal} \widehat{\mathbf{S}}_y^{-1} \left(\sum_{i=1}^{n} u_{i,i} \right) \widehat{\mathbf{S}}_{\mathbf{x}}^{-1} \mathbf{e_0} \quad + \quad \frac{1}{n^2 h^{2+d+\vartheta}} \mathbf{e}_{1+\vartheta}^{\intercal} \widehat{\mathbf{S}}_y^{-1} \left(\sum_{i,j=1, i \ne j}^{n} u_{i,j} \right) \widehat{\mathbf{S}}_{\mathbf{x}}^{-1} \mathbf{e_0}. \qquad \text{(III + IV)}$$

We first provide probability concentration results for the matrices $\widehat{\mathbf{S}}_{\mathbf{x}}$ and $\widehat{\mathbf{S}}_y$. We will then show that term (II) encompasses the target parameter $f^{(\vartheta)}(y|\mathbf{x})$ and the smoothing bias. Next, we establish probabilistic orders for (III) and (IV). We analyze term (I) as the last step, which will close the proof.

Convergence of $\widehat{\mathbf{S}}_{\mathbf{x}}$ ***and*** $\widehat{\mathbf{S}}_y$. To start, note that $\mathcal{X}$ is compact, then for any $\eta_n > 0$, one can find $\{\mathbf{x}_\ell : 1 \le \ell \le M_n\}$, such that $\mathcal{X} \subseteq \cup_{1 \le \ell \le M_n} B_\ell$, where $B_\ell := B(\mathbf{x}_\ell, \eta_n)$ is the Euclidean ball centered at $\mathbf{x}_\ell$ with radius η_n. Define $\mathtt{r} = \sqrt{\log(n)/(nh^d)}$. Then,

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| \widehat{\mathbf{S}}_{\mathbf{x}} - \mathbf{S}_{\mathbf{x}} \right| \le \max_{1 \le \ell \le M_n} \left| \widehat{\mathbf{S}}_{\mathbf{x}_\ell} - \mathbf{S}_{\mathbf{x}_\ell} \right| + \sup_{1 \le \ell \le M_n} \sup_{\mathbf{x} \in B_\ell} \left| \widehat{\mathbf{S}}_{\mathbf{x}} - \widehat{\mathbf{S}}_{\mathbf{x}_\ell} \right| + \sup_{1 \le \ell \le M_n} \sup_{\mathbf{x} \in B_\ell} \left| \mathbf{S}_{\mathbf{x}} - \mathbf{S}_{\mathbf{x}_\ell} \right|.$$

Consider the last term on the RHS. It is straightforward to show that $\mathbf{S}_{\mathbf{x}}$ is continuous with Lipschitz constant of order h^{-1}, which implies that $\sup_{1 \le \ell \le M_n} \sup_{\mathbf{x} \in B_\ell} |\mathbf{S}_{\mathbf{x}} - \mathbf{S}_{\mathbf{x}_\ell}| \precsim \eta_n / h$. Similar technique applies to the second term on the RHS: the matrix $\widehat{\mathbf{S}}_{\mathbf{x}}$ is the average of continuous functions with Lipschitz constant of order h^{-1-d}, which means $\sup_{1 \le \ell \le M_n} \sup_{\mathbf{x} \in B_\ell} |\widehat{\mathbf{S}}_{\mathbf{x}} - \widehat{\mathbf{S}}_{\mathbf{x}_\ell}| \precsim \eta_n / h^{1+d}$.

Now consider the first term. By employing the union bound, we have that, for any constant $\mathfrak{c}_1 > 0$,

$$\mathbb{P}\Big[\max_{1 \le \ell \le M_n} \left| \widehat{\mathbf{S}}_{\mathbf{x}_\ell} - \mathbf{S}_{\mathbf{x}_\ell} \right| > \mathfrak{c}_1 \mathtt{r} \Big] \le M_n \max_{1 \le \ell \le M_n} \mathbb{P}\Big[\left| \widehat{\mathbf{S}}_{\mathbf{x}_\ell} - \mathbf{S}_{\mathbf{x}_\ell} \right| > \mathfrak{c}_1 \mathtt{r} \Big].$$

To proceed, we recall the formula of $\widehat{\mathbf{S}}_{\mathbf{x}}$, and it follows that the summands satisfy

$$\mathbb{V}\Big[\frac{1}{h^d} \mathbf{q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big) \mathbf{Q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)^{\intercal} \Big] \le C' h^{-d}, \qquad \Big| \frac{1}{h^d} \mathbf{q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big) \mathbf{Q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)^{\intercal} \Big| \le C' h^{-d},$$

where C' is a constant that does not depend on n, h or the evaluation point $\mathbf{x}$. Applying Bernstein's inequality,

$$M_n \max_{1\le \ell\le M_n} \mathbb{P}\Big[\big|\widehat{\mathbf{S}}_{\mathbf{x}_\ell} - \mathbf{S}_{\mathbf{x}_\ell}\big| > \mathfrak{c}_1 \mathtt{r}\Big] \le 2\exp\Big\{-\frac{1}{2}\frac{\mathfrak{c}_1^2\log(n)}{C' + \frac{1}{3}\mathfrak{c}_1 C'\mathtt{r}} + \log(M_n)\Big\}.$$

To complete the proof, we note that M_n is at most polynomial in n as long as η_n is also polynomial in n. Therefore, one can choose η_n sufficiently small so that η_n/h^{1+d} become negligible, and hence for some constants $\mathfrak{c}_1$, $\mathfrak{c}_2$, and $\mathfrak{c}_3$,

$$\mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}}\big|\widehat{\mathbf{S}}_{\mathbf{x}} - \mathbf{S}_{\mathbf{x}}\big| > \mathfrak{c}_1\mathtt{r}\Big] \le \mathfrak{c}_2 n^{-\mathfrak{c}_3},$$

and $\mathfrak{c}_3$ can be made arbitrarily large with appropriate choices of $\mathfrak{c}_1$. In other words, we have shown that $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\mathbf{S}}_{\mathbf{x}} - \mathbf{S}_{\mathbf{x}}| \precsim_{\mathtt{TC}} \sqrt{\log(n)/(nh^d)}$. Analogously, we can show the probability concentration result $\sup_{y\in\mathcal{Y}}|\widehat{\mathbf{S}}_y - \mathbf{S}_y| \precsim_{\mathtt{TC}} \sqrt{\log(n)/(nh)}$.

Term (II), and the smoothing bias calculation. We start with a Taylor expansion of the conditional CDF up to some order s:

$$F(y_j|\mathbf{x}_i) = \sum_{\ell+|\mathbf{m}|\le s} \frac{\partial^\ell}{\partial y^\ell}\frac{\partial^{\mathbf{m}}}{\partial \mathbf{x}^{\mathbf{m}}}F(y|\mathbf{x})\frac{1}{\ell!\mathbf{m}!}(y_j-y)^\ell(\mathbf{x}_i-\mathbf{x})^{\mathbf{m}} + o\bigg(\sum_{\ell+|\mathbf{m}|=s}|y_j-y|^\ell|\mathbf{x}_i-\mathbf{x}|^{\mathbf{m}}\bigg).$$

Then,

$$\begin{aligned}
&\frac{1}{n^2h^{2+d+\vartheta}}\sum_{i,j=1}^n \mathbf{e}_{1+\vartheta}^\intercal \widehat{\mathbf{S}}_y^{-1} F(y_j|\mathbf{x}_i)\mathbf{P}\Big(\frac{y_j-y}{h}\Big)\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)^\intercal \widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e_0}\\
=&\frac{1}{n^2h^{2+d+\vartheta}}\sum_{i,j=1}^n \mathbf{e}_{1+\vartheta}^\intercal \widehat{\mathbf{S}}_y^{-1}\sum_{\ell+|\mathbf{m}|\le s}\frac{\partial^\ell}{\partial y^\ell}\frac{\partial^{\mathbf{m}}}{\partial\mathbf{x}^{\mathbf{m}}}F(y|\mathbf{x})\frac{1}{\ell!\mathbf{m}!}(y_j-y)^\ell(\mathbf{x}_i-\mathbf{x})^{\mathbf{m}}\mathbf{P}\Big(\frac{y_j-y}{h}\Big)\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)^\intercal\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e_0}\\
&+o\bigg(\frac{1}{n^2h^{2+d+\vartheta}}\mathbf{e}_{1+\vartheta}^\intercal\widehat{\mathbf{S}}_y^{-1}\sum_{i,j=1}^n\sum_{\ell+|\mathbf{m}|=s}|y_j-y|^\ell|\mathbf{x}_i-\mathbf{x}|^{\mathbf{m}}\Big|\mathbf{P}\Big(\frac{y_j-y}{h}\Big)\Big|\Big|\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)\Big|\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e_0}\bigg)\\
=&f^{(\vartheta)}(y|\mathbf{x}) + O_{\mathbb{P}}(h^{\mathfrak{q}+1}+h^{\mathfrak{p}-\vartheta}).
\end{aligned}$$

To understand the stochastic order, we notice that the first nonzero term in the summation corresponds to $\ell = 1+\vartheta$ and $\mathbf{m} = \mathbf{0}$, which gives rise to the target parameter $f^{(\vartheta)}(y|\mathbf{x})$. The next nonzero terms in the summation will be the leading smoothing bias, and correspond to $\ell = 1+\vartheta$ and $|\mathbf{m}| = \mathfrak{q}+1$, or $\ell = \mathfrak{p}+1$ and $\mathbf{m} = \mathbf{0}$. The leading bias terms will involve random vectors and matrices that are sample averages, whose probabilistic orders can be established using the earlier method of combining discretization, union bound, and Bernstein's inequality.

Term (III), the leave-in bias. This term arises because the same observation is used twice: y_i is used to construct the conditional CDF estimator $\widehat{F}(y|\mathbf{x})$, and later as an evaluation point in the second step local polynomial regression. Term (III) takes the form of a sample average, and using the earlier method of combining discretization, union bound, and Bernstein's inequality, it is straightforward to show that

it has the order

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} |(\text{III})| \precsim_{\text{TC}} \frac{1}{nh^{1+\vartheta}}\left(1+\sqrt{\frac{\log(n)}{nh^{1+d}}}\right).$$

Term (IV). Term (IV) is a degenerate U-statistic. Take C and C' to be some large constant, and we set

$$A = C', \qquad B^2 = C'nh, \qquad D^2 = C'n^2h^{d+1}, \qquad t = C(\log(n))\sqrt{n^2h^{d+1}}.$$

We apply Lemmas 8 and 9, which give (the value of C' may change for each line)

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\Big|\sum_{i,j=1,i\neq j}^{n} u_{i,j}\Big| > t\Big] \leq C'\exp\left\{-\frac{1}{C'}\min\left[\frac{t}{\sqrt{n^2h^{d+1}}},\ \frac{t^{2/3}}{(nh)^{1/3}},\ t^{1/2}\right] + \log(n)\right\}\\
&= C'\exp\Big\{-\frac{\sqrt{C}}{C'}\min\Big[\log(n),\ \left(\log^2(n)nh^d\right)^{\frac{1}{3}},\ \left(\log^2(n)n^2h^{1+d}\right)^{\frac{1}{4}}\Big] + \log(n)\Big\}.
\end{aligned}$$

As a result,

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \big|(\text{IV})\big| \precsim_{\text{TC}} \frac{\log(n)}{\sqrt{n^2h^{3+d+2\vartheta}}}.$$

Term (I). To close the proof, we write $(\text{I}) - \bar{f}^{(\vartheta)}(y|\mathbf{x}) = (\text{I.1}) + (\text{I.2})$, where

$$\begin{aligned}
(\text{I.1}) &= \frac{1}{nh^{2+d+\vartheta}}\mathbf{e}_{1+\vartheta}^{\intercal}(\widehat{\mathbf{S}}_y^{-1} - \mathbf{S}_y^{-1})\Big(\sum_{i=1}^{n}\int_{\mathcal{Y}}\Big(\mathbb{1}(y_i\leq u) - F(u|\mathbf{x}_i)\Big)\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}F_y(u)\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)^{\intercal}\Big)\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}},\\
(\text{I.2}) &= \frac{1}{nh^{2+d+\vartheta}}\mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\Big(\sum_{i=1}^{n}\int_{\mathcal{Y}}\Big(\mathbb{1}(y_i\leq u) - F(u|\mathbf{x}_i)\Big)\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}F_y(u)\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)^{\intercal}\Big)(\widehat{\mathbf{S}}_{\mathbf{x}}^{-1} - \mathbf{S}_{\mathbf{x}}^{-1})\mathbf{e}_{\mathbf{0}}.
\end{aligned}$$

To analyze term (I.1), we have shown that $\sup_{y\in\mathcal{Y}}|\widehat{\mathbf{S}}_y - \mathbf{S}_y| \precsim_{\text{TC}} \sqrt{\log(n)/(nh)}$ and $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\mathbf{S}}_{\mathbf{x}}| \precsim_{\text{TC}} 1+\sqrt{\log(n)/(nh^d)}$. Notice that both $\mathbf{S}_y$ and $\mathbf{S}_{\mathbf{x}}$ are invertible, which means the same rates apply after inverting the matrices. The middle matrix in (I.1) is a sample average that is mean zero and has variance of order nh^{2+d}. We can therefore apply the earlier technique of discretization, union bound, and Bernstein's inequality to show that the middle matrix has the order $\sqrt{\log(n)nh^{2+d}}$. Therefore,

$$(\text{I.1}) \precsim_{\text{TC}} \frac{1}{nh^{2+d+\vartheta}}\sqrt{\frac{\log(n)}{nh}}\sqrt{\log(n)nh^{2+d}}\Big(1+\sqrt{\frac{\log(n)}{nh^d}}\Big) \precsim \frac{\log(n)}{\sqrt{n^2h^{3+d+2\vartheta}}}.$$

To analyze term (I.2), we us the fact that $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\mathbf{S}}_{\mathbf{x}} - \mathbf{S}_{\mathbf{x}}| \precsim_{\text{TC}} \sqrt{\log(n)/(nh^d)}$ and the rest of the term is mean zero conditional on $\mathbf{x}_i$. It remains to compute the variance.

$$\begin{aligned}
&\mathbb{V}\Big[\mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\int_{\mathcal{Y}}\Big(\mathbb{1}(y_i\leq u) - F(u|\mathbf{x}_i)\Big)\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}F_y(u)\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)^{\intercal}\Big]\\
&= h^2\mathbb{E}\Big[\iint_{(\mathcal{Y}-y)/h}\Big(F(y+h(u_1\wedge u_2)|\mathbf{x}_i) - F(y+hu_1|\mathbf{x}_i)F(y+hu_2|\mathbf{x}_i)\Big)f_y(y+hu_1)f_y(y+hu_2)\\
&\qquad\qquad \mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\mathbf{P}(u_1)\mathbf{P}(u_2)^{\intercal}\mathbf{S}_y^{-1}\mathbf{e}_{1+\vartheta}\mathrm{d}u_1\mathrm{d}u_2\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)\mathbf{Q}\Big(\frac{\mathbf{x}_i-\mathbf{x}}{h}\Big)^{\intercal}\Big],
\end{aligned}$$

where f_y represents the marginal PDF of y_i. By a standard Taylor expansion (in h) exercise, one can show that the leading term is zero, which means the variance has the order h^{3+d}. We can therefore apply the earlier technique (discretization, union bound, and Bernstein's inequality) to show that

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\Big(\sum_{i=1}^{n}\int_{\mathcal{Y}}\Big(\mathbb{1}(y_i\le u)-F(u|\mathbf{x}_i)\Big)\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}F_y(u)\mathbf{Q}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right)^{\intercal}\Big)\right|\lesssim_{\mathtt{TC}}\sqrt{\log(n)nh^{3+d}}.$$

As a result,

$$\text{(I.2)}\lesssim_{\mathtt{TC}}\frac{1}{nh^{3+d+\vartheta}}\sqrt{\log(n)nh^{3+d}}\sqrt{\frac{\log(n)}{(nh^d)}}=\frac{\log(n)}{\sqrt{n^2h^{1+2d+2\vartheta}}}.$$

A.4. Properties of the equivalent kernel

In this appendix we prove some useful properties of the equivalent kernel function $\mathscr{K}^{\circ}_{\vartheta,h}$, which will be employed to establish the strong approximation result in Theorem 2.

Lemma 4 (Leading variance). *Suppose Assumptions 1 and 2 hold. If $h\to 0$ and if $nh^{1+d}/\log(n)\to\infty$, then (3) holds.*

Proof of Lemma 4. To save notation, let $\mathbf{c}_1=\mathbf{S}_y^{-1}\mathbf{e}_{1+\vartheta}$ and $\mathbf{c}_2=\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}$. Then

$$\begin{aligned}&\mathbb{V}\left[\int_{\mathcal{Y}}\big(\mathbb{1}(y_i\le u)-F(u|\mathbf{x}_i)\big)\mathbf{c}_1^{\intercal}\frac{1}{h}\mathbf{P}\left(\frac{u-y}{h}\right)f_y(u)\mathrm{d}u\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right)^{\intercal}\mathbf{c}_2\right]\\&=\mathbb{E}\Big[\iint_{\frac{\mathcal{Y}-y}{h}}\Big(F(y+h(u_1\wedge u_2)|\mathbf{x}_i)-F(y+hu_1|\mathbf{x}_i)F(y+hu_2|\mathbf{x}_i)\Big)f_y(y+hu_1)f_y(y+hu_2)\\&\qquad\mathbf{c}_1^{\intercal}\mathbf{P}(u_1)\,\mathbf{c}_1^{\intercal}\mathbf{P}(u_2)\,\mathrm{d}u_1\mathrm{d}u_2\Big(\mathbf{c}_2^{\intercal}\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right)\Big)^2\Big].\end{aligned}\tag{I}$$

We make a further expansion:

$$\begin{aligned}&F(y+h(u_1\wedge u_2)|\mathbf{x}_i)-F(y+hu_1|\mathbf{x}_i)F(y+hu_2|\mathbf{x}_i)\\&=F(y|\mathbf{x}_i)(1-F(y|\mathbf{x}_i))+h(u_1\wedge u_2)f(y|\mathbf{x}_i)-h(u_1+u_2)f(y|\mathbf{x}_i)F(y|\mathbf{x}_i)+O(h^2).\end{aligned}$$

Note that the remainder term, $O(h^2)$, holds uniformly for $y\in\mathcal{Y}$ and $\mathbf{x}_i\in\mathcal{X}$ since the conditional distribution function is assumed to have bounded second derivatives. In addition, it is straightforward to verify that with the above Taylor expansion, the first term in (I) is zero, meaning that the leading variance term is

$$\text{(I)}=h\left(\mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\mathbf{T}_y\mathbf{S}_y^{-1}\mathbf{e}_{1+\vartheta}\right)\mathbb{E}\Big[f(y|\mathbf{x}_i)\Big(\mathbf{c}_2^{\intercal}\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right)\Big)^2\Big]+O\Big(\frac{1}{h^{d-2}}\Big).$$

To conclude the proof, we compute the expectation,

$$\text{(I)}=\frac{1}{h^{d-1}}f(y|\mathbf{x})\left(\mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\mathbf{T}_y\mathbf{S}_y^{-1}\mathbf{e}_{1+\vartheta}\right)\left(\mathbf{e}_{\mathbf{0}}^{\intercal}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{T}_{\mathbf{x}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}\right)+O\left(\frac{1}{h^{d-2}}\right).$$

Therefore, (3) holds. ■

Lemma 5 (Properties of $\mathscr{K}^{\circ}_{\vartheta,h}$). *Let Assumptions 1 and 2 hold. Then*
(i) $\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x})$ is bounded: $\sup_{a,\mathbf{b},y,\mathbf{x}}|\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x})| \precsim h^{-1-d-\vartheta}$.
(ii) $\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x})$ is Lipschitz continuous:

$$\sup_{|a-a'|+|\mathbf{b}-\mathbf{b}'|>0,y,\mathbf{x}} \frac{\left|\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x}) - \mathscr{K}^{\circ}_{\vartheta,h}(a',\mathbf{b}';y,\mathbf{x})\right|}{|a-a'|+|\mathbf{b}-\mathbf{b}'|} = O\left(h^{-2-d-\vartheta}\right),$$

$$\sup_{a,\mathbf{b},|y-y'|+|\mathbf{x}-\mathbf{x}'|>0} \frac{\left|\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x}) - \mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y',\mathbf{x}')\right|}{|y-y'|+|\mathbf{x}-\mathbf{x}'|} = O\left(h^{-2-d-\vartheta}\right).$$

Proof of Lemma 5. *Part (i).* We first rewrite the kernel using change-of-variable. Then, $h^{1+d+\vartheta}\mathscr{K}^{\circ}_{\vartheta,h}$ takes the form

$$\mathbf{e}_{1+\vartheta}^{\mathsf{T}}\mathbf{S}_y^{-1}\Big[\int_{\frac{\mathcal{Y}-y}{h}}\Big(\mathbb{1}(a\le y+hv)-F(y+hv|\mathbf{b})\Big)\mathbf{P}(v)\,f_y(y+hv)\mathrm{d}v\Big]\mathbf{Q}\left(\frac{\mathbf{b}-\mathbf{x}}{h}\right)^{\mathsf{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}.$$

It should be clear that the above is bounded.

Part (ii). From the expression in part (i), it is clear that $h^{1+d+\vartheta}\mathscr{K}^{\circ}_{\vartheta,h}$ is Lipschitz continuous in $\mathbf{b}$ with a Lipschitz constant of order h^{-1}. Next consider the directions a. We have

$$\begin{aligned}
&\sup_{\mathbf{b},y,\mathbf{x}} h^{1+d+\vartheta}|\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x}) - \mathscr{K}^{\circ}_{\vartheta,h}(a',\mathbf{b};y,\mathbf{x})| \\
&\precsim \sup_y\left|\int_{\frac{\mathcal{Y}-y}{h}}\Big(\mathbb{1}(a\le y+hv)-\mathbb{1}(a'\le y+hv)\Big)\mathbf{P}(v)\,f_y(y+hv)\mathrm{d}v\right| \\
&\precsim \sup_y\left|\int_{\frac{\mathcal{Y}-y}{h}\cap[-1,1]\cap\left[\frac{a-y}{h},\frac{a'-y}{h}\right]}\mathbf{P}(v)\,f_y(y+hv)\mathrm{d}v\right|.
\end{aligned}$$

Therefore, the kernel is also Lipschitz-h^{-1} continuous with respect to a.

To conclude the proof, it is straightforward to show that $\mathbf{S}_{\mathbf{x}}$ and $\mathbf{S}_y$ are Lipschitz continuous with respect to $\mathbf{x}$ and y, with the Lipschitz constant of order $1/h$. The same holds for their inverses. ■

Lemma 6 (Covering number). *Define $\mathcal{K}=\{h^{1+d+\vartheta}\mathscr{K}^{\circ}_{\vartheta,h}(\cdot,\cdot;y,\mathbf{x}):\ y\in\mathcal{Y},\ \mathbf{x}\in\mathcal{X}\}$. Let Assumptions 1 and 2 hold. Then*

$$\sup_P N\Big(\varepsilon,\ \mathcal{K},\ L^1(P)\Big)\le \mathfrak{c}\frac{1}{\varepsilon^{d+2}}+1,$$

where the supremum is taken over all probability measures on $[0,1]^{d+1}$, and the constant $\mathfrak{c}$ does not depend on the bandwidth h.

Proof of Lemma 6. To show this result, it suffices to consider the uncentered kernel function,

$$\begin{aligned}
h^{1+d+\vartheta}\mathscr{K}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x}) &= \mathbf{e}_{1+\vartheta}^{\mathsf{T}}\mathbf{S}_y^{-1}\int_{\mathcal{Y}}\mathbb{1}(a\le u)\frac{1}{h}\mathbf{P}\Big(\frac{u-y}{h}\Big)\mathrm{d}F_y(u)\mathbf{Q}\left(\frac{\mathbf{b}-\mathbf{x}}{h}\right)^{\mathsf{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}} \\
&= \mathbf{e}_{1+\vartheta}^{\mathsf{T}}\mathbf{S}_y^{-1}\Big[\int_{\frac{\mathcal{Y}-y}{h}}\mathbb{1}(a\le y+hv)\mathbf{P}(v)\,f_y(y+hv)\mathrm{d}v\Big]\mathbf{Q}\left(\frac{\mathbf{b}-\mathbf{x}}{h}\right)^{\mathsf{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}.
\end{aligned}$$

We will first show that it has compact support. Consider two cases. If $(a-y)/h > 1$, then the integrand $\mathbb{1}(a \le y+hv)\mathbf{P}(v)$ will be zero because $\mathbf{P}(v)$ is zero for $v \ge 1$. Therefore, the kernel defined above will be zero as well. For the case that $(a-y)/h \le -1$, we can simply drop the indicator, as again $\mathbf{P}(v)$ will be zero for $v \le -1$. Then the kernel becomes

$$h^{1+d+\vartheta}\mathscr{K}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\intercal}\mathbf{S}_y^{-1}\Big[\int_{\frac{y-y}{h}} \mathbf{P}(v)\, f_y(y+hv)\mathrm{d}v\Big]\mathbf{Q}\left(\frac{\mathbf{b}-\mathbf{x}}{h}\right)^{\intercal}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}, \qquad a \le -1.$$

Note that the matrix, $\mathbf{S}_y$, can be written as $\mathbf{S}_y = \int_{\frac{y-y}{h}} \mathbf{P}(v)\mathbf{p}(v)^{\intercal} f_y(y+hv)\mathrm{d}v$, which means its first column is $\int_{\frac{y-y}{h}} \mathbf{P}(v)\, f_y(y+hv)\mathrm{d}v$, showing that the expression above is zero. As for the second argument, $\mathbf{b}$, we note that $\mathbf{Q}((\mathbf{b}-\mathbf{x})/h)$ is zero if $\mathbf{b}$ lies outside of an h-cube around $\mathbf{x}$.

With the above result, we can simply apply Lemmas 5 and 7 to conclude the covering number result for the class $\{h^{1+d+\vartheta}\mathscr{K}_{\vartheta,h}(\cdot,\cdot;y,\mathbf{x}) : \ y \in \mathcal{Y},\ \mathbf{x} \in \mathcal{X}\}$ (note that the boundedness and Lipschitz continuity results in Lemma 5 also apply to $\mathscr{K}_{\vartheta,h}$). The same covering number then holds for $\mathcal{K}$, as the two classes differ only by a centering. ■

A.5. Proof of Theorem 1

Given Lemma 1, we will only need to provide a probability concentration for $\bar{f}^{(\vartheta)}(y|\mathbf{x})$. We have established in Lemma 4 that

$$\mathbb{V}[\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x})] \le C'\frac{1}{h^{1+d+2\vartheta}}, \quad |\mathscr{K}^{\circ}_{\vartheta,h}(a,\mathbf{b};y,\mathbf{x})| \le C'\frac{1}{h^{1+d+\vartheta}}.$$

Then we apply the technique used in the proof of Lemma 1 (discretization, union bound, and Bernstein's inequality), which leads to $\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\bar{f}^{(\vartheta)}(y|\mathbf{x})| \precsim_{\mathtt{TC}} \sqrt{\log(n)/(nh^{1+d+2\vartheta})}$. To conclude the proof, we notice that the second component in $\mathtt{r}_{\mathtt{SL}}$ satisfies

$$\frac{\log(n)}{\sqrt{n^2h^{1+2\vartheta+d+(2\vee d)}}} = \sqrt{\frac{\log(n)}{nh^{1+d+2\vartheta}}}\sqrt{\frac{\log(n)}{nh^{2\vee d}}} = o\left(\sqrt{\frac{\log(n)}{nh^{1+d+2\vartheta}}}\right).$$

A.6. Proof of Theorem 2

It suffices to consider the process $\tilde{\mathbb{S}}_\vartheta(y,\mathbf{x}) = \sum_{i=1}^n h^{1+d+\vartheta}\mathscr{K}^{\circ}_{\vartheta,h}(y_i,\mathbf{x}_i;y,\mathbf{x})/\sqrt{n}$, which is the empirical process indexed by the function class $\mathcal{K}$ (defined in Lemma 6 above). From Lemma 5, the functions in the above class are uniformly bounded. Lemma 6 shows that the function class above is of VC type, and the covering number does not depend on the bandwidth. The measurability condition required in Lemma 10 also holds, as our function class is indexed by $(y,\mathbf{x}) \in [0,1]^{d+1}$, and the functions in $\mathcal{K}$ are continuous in y and $\mathbf{x}$.

Now the only missing ingredient is the total variation of the functions in $\mathcal{K}$. First, note that the function $h^{1+d+\vartheta}\mathscr{K}^{\circ}_{\vartheta,h}(\cdot,\cdot;y,\mathbf{x})$ is Lipschitz continuous with respect to the arguments, and the Lipschitz constant is of order h^{-1}. Therefore, its total variation is bounded by

$$\mathrm{TV}_{(y,\mathbf{x})} = \mathrm{TV}\Big(h^{1+d+\vartheta}\mathscr{K}^{\circ}_{\vartheta,h}(\cdot,\cdot;y,\mathbf{x})\Big) \precsim \frac{1}{h}\mathrm{vol}\Big(\mathrm{supp}\Big(\mathscr{K}_{\vartheta,h}(\cdot,\cdot;y,\mathbf{x})\Big)\Big),$$

where $\mathrm{vol}(\mathrm{supp}(\cdot))$ denotes the Euclidean volume of the support, and $\mathscr{K}_{\vartheta,h}$ is defined in the proof of Lemma 6. We also showed in the proof of Lemma 6 that $\mathscr{K}_{\vartheta,h}$ has compact support, leading to $\mathrm{TV}_{\mathcal{K}} = \sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \mathrm{TV}_{(y,\mathbf{x})} \precsim h^d$.

Putting all pieces together, we conclude that there exists a centered Gaussian process, $\tilde{\mathbb{G}}_\vartheta$ which has the same covariance kernel as $\tilde{\mathbb{S}}_\vartheta$, such that

$$\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \big|\tilde{\mathbb{S}}'_\vartheta(y,\mathbf{x}) - \tilde{\mathbb{G}}_\vartheta(y,\mathbf{x})\big| \geq \mathfrak{c}_1\Big(\sqrt{\frac{h^d \log n}{n^{\frac{1}{d+1}}}} + \sqrt{\frac{\log^3 n}{n}}\Big)\Big] \leq \mathfrak{c}_2 n^{-\mathfrak{c}_3},$$

where $\tilde{\mathbb{S}}'_\vartheta(y,\mathbf{x})$ is a copy of $\tilde{\mathbb{S}}_\vartheta(y,\mathbf{x})$.

A.7. Proof of Theorem 3

First consider $\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x})$. The difference between $\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x})$ and $\widehat{\mathbb{S}}_\vartheta(y,\mathbf{x})$ is

$$\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x}) - \widehat{\mathbb{S}}_\vartheta(y,\mathbf{x}) = \Big(\sqrt{\frac{\mathsf{V}_\vartheta(y,\mathbf{x})}{\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})}} - 1\Big)\widehat{\mathbb{S}}_\vartheta(y,\mathbf{x}).$$

With Theorem 1, Lemma 2 and the variance bound in (3) (also see Lemma 4 in Appendix A.4), we have

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \Big|\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x}) - \widehat{\mathbb{S}}_\vartheta(y,\mathbf{x})\Big| \precsim_{\mathtt{TC}} \mathtt{r}_{\mathtt{VE}}\Big(h^{\mathfrak{p}-\vartheta} + \sqrt{\frac{\log(n)}{nh^{1+d+2\vartheta}}}\Big)\sqrt{nh^{1+d+2\vartheta}} \precsim \sqrt{\log(n)}\,\mathtt{r}_{\mathtt{VE}}.$$

Next, we establish a Gaussian comparison result. Consider an ε discretization of $\mathcal{Y}\times\mathcal{X}$, which is denoted by $\mathcal{A}_\varepsilon = \{(y_\ell, \mathbf{x}_\ell^\intercal) : \ 1 \leq \ell \leq L\}$. Then one can define two Gaussian vectors, $\mathbf{z}, \widehat{\mathbf{z}} \in \mathbb{R}^L$, such that

$$\mathbb{C}\mathrm{ov}[z_\ell, z_{\ell'}] = \rho_\vartheta(y_\ell,\mathbf{x}_\ell,y_{\ell'},\mathbf{x}_{\ell'}), \quad \mathbb{C}\mathrm{ov}[\widehat{z}_\ell, \widehat{z}_{\ell'}|\mathbf{X},\mathbf{Y}] = \widehat{\rho}_\vartheta(y_\ell,\mathbf{x}_\ell,y_{\ell'},\mathbf{x}_{\ell'}).$$

Then we apply the Gaussian comparison result in Lemma 11 and the correlation estimation error rate in Lemma 2, which lead to

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big[\sup_{1\leq\ell\leq L}|\widehat{\mathbb{G}}_\vartheta(y_\ell,\mathbf{x}_\ell)| \leq u\Big|\mathbf{Y},\mathbf{X}\Big] - \mathbb{P}\Big[\sup_{1\leq\ell\leq L}|\mathbb{G}_\vartheta(y_\ell,\mathbf{x}_\ell)| \leq u\Big]\Big| \precsim_{\mathbb{P}} \sqrt{\mathtt{r}_{\mathtt{VE}}}\log\Big(\frac{1}{\varepsilon}\Big).$$

Since ε only enters the above error bound logarithmically, one can choose $\varepsilon = n^{-\mathfrak{c}}$ for some $\mathfrak{c}$ large enough, so that the error that arises from discretization becomes negligible. In other words, we have

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{G}}_\vartheta(y,\mathbf{x})| \leq u\Big|\mathbf{Y},\mathbf{X}\Big] - \mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\mathbb{G}_\vartheta(y,\mathbf{x})| \leq u\Big]\Big| \precsim_{\mathbb{P}} \log(n)\sqrt{\mathtt{r}_{\mathtt{VE}}}.$$

Now consider $\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x})$ again. Given the bound on the difference, $\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x}) - \widehat{\mathbb{S}}_\vartheta(y,\mathbf{x})|$, and the strong approximation in Theorem 2, we clearly have

$$\begin{aligned}\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\mathbb{G}_\vartheta(y,\mathbf{x})| \leq u - \mathfrak{c}_1(\sqrt{\log(n)}\mathtt{r}_{\mathtt{VE}} + \mathtt{r}_{\mathtt{SA}})\Big] - \mathfrak{c}_2 n^{-\mathfrak{c}_3} &\leq \mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x})| \leq u\Big] \\ &\leq \mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\mathbb{G}_\vartheta(y,\mathbf{x})| \leq u + \mathfrak{c}_1(\sqrt{\log(n)}\mathtt{r}_{\mathtt{VE}} + \mathtt{r}_{\mathtt{SA}})\Big] + \mathfrak{c}_2 n^{-\mathfrak{c}_3}.\end{aligned}$$

Finally, we apply the Gaussian comparison result, which implies that

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{T}}_{\vartheta}(y,\mathbf{x})|\le u\Big]-\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{G}}_{\vartheta}(y,\mathbf{x})|\le u\Big|\mathbf{Y},\mathbf{X}\Big]\right|$$
$$\lesssim_{\mathbb{P}} \mathfrak{c}_2 n^{-\mathfrak{c}_3}+\log(n)\sqrt{\mathtt{r}_{\mathrm{VE}}}+\sup_{u\in\mathbb{R}}\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\mathbb{G}_{\vartheta}(y,\mathbf{x})|\in[u,u+\mathfrak{c}_1(\sqrt{\log(n)}\mathtt{r}_{\mathrm{VE}}+\mathtt{r}_{\mathrm{SA}})]\Big].$$

Finally, due to Lemma 12, we have

$$\sup_{u\in\mathbb{R}}\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\mathbb{G}_{\vartheta}(y,\mathbf{x})|\in[u,u+\mathfrak{c}_1(\sqrt{\log(n)}\mathtt{r}_{\mathrm{VE}}+\mathtt{r}_{\mathrm{SA}})]\Big]\lesssim\sqrt{\log(n)}(\sqrt{\log(n)}\mathtt{r}_{\mathrm{VE}}+\mathtt{r}_{\mathrm{SA}}).$$

A.8. Derivation of (5)

First consider $\widehat{f}_{\mathrm{N}}(y|\mathbf{x})$. If the unconstrained estimator, $\widehat{f}(y|\mathbf{x})$, is already nonnegative, then the constraint in the least squares problem is not binding, which means in this case $\widehat{f}_{\mathrm{N}}(y|\mathbf{x})=\widehat{f}(y|\mathbf{x})$. Now assume $\widehat{f}(y|\mathbf{x})<0$. Since the least squares objective function is strictly convex, the solution will be on the boundary of the set $\{\mathbf{u}\in\mathbb{R}^{\mathrm{p}+1}:\mathbf{e}_1^{\mathsf{T}}\mathbf{u}\ge 0\}$, leading to $\widehat{f}_{\mathrm{N}}(y|\mathbf{x})=0$. Therefore, we have the expression $\widehat{f}_{\mathrm{N}}(y|\mathbf{x})=\max\{0,\widehat{f}(y|\mathbf{x})\}$ in (5).

The expression of $\widehat{f}_{\mathrm{I}}(y|\mathbf{x})$ in (5) follows from Jensen's inequality, which is binding if and only if $g(y)/\widehat{f}_{\mathrm{N}}(y|\mathbf{x})$ is constant (in y).

A.9. Proof of Lemma 3, Theorems 7 and 8

We write $\widehat{f}_{\mathrm{N}}(y|\mathbf{x})=\widehat{f}(y|\mathbf{x})-\mathbb{1}(\widehat{f}(y|\mathbf{x})<0)\cdot\widehat{f}(y|\mathbf{x})$. We first study the indicator function. Take $\mathtt{r}$ to be any sequence shrinking to 0, and $\mathfrak{c}_1$ some positive constant. Then

$$\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\mathbb{1}\left(\widehat{f}(y|\mathbf{x})<0\right)>\mathtt{r}\mathfrak{c}_1\Big]\le\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\widehat{f}(y|\mathbf{x})-f(y|\mathbf{x})\right|>\inf_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}f(y|\mathbf{x})\Big].$$

Then by the probability concentration in Theorem 1, it should be obvious that the the above probability vanishes faster than any polynomials of n (recall that we assume the conditional density is uniformly bounded away from zero); that is, $\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\mathbb{1}(\widehat{f}(y|\mathbf{x})<0)\lesssim_{\mathrm{TC}}\mathtt{r}$ for any vanishing sequence $\mathtt{r}$. This shows that $\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{f}_{\mathrm{N}}(y|\mathbf{x})-\widehat{f}(y|\mathbf{x})|\lesssim_{\mathrm{TC}}\mathtt{r}$. By letting $\mathtt{r}$ shrinking to 0 fast enough, we have the stochastic linearization for $\widehat{f}_{\mathrm{N}}(y|\mathbf{x})$.

Next, For $\widehat{f}_{\mathrm{I}}(y|\mathbf{x})$, we employ the following decomposition:

$$\widehat{f}_{\mathrm{I}}(y|\mathbf{x})=\widehat{f}_{\mathrm{N}}(y|\mathbf{x})-\frac{\widehat{f}_{\mathrm{N}}(y|\mathbf{x})}{\int_{\mathcal{Y}}\widehat{f}_{\mathrm{N}}(u|\mathbf{x})\mathrm{d}u}\int_{\mathcal{Y}}\big(\widehat{f}_{\mathrm{N}}(u|\mathbf{x})-f(u|\mathbf{x})\big)\mathrm{d}u.$$

Then we can write

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\widehat{f}_{\mathrm{I}}(y|\mathbf{x})-f(y|\mathbf{x})-\left(\bar{f}^{(0)}(y|\mathbf{x})-f(y|\mathbf{x})\int_{\mathcal{Y}}\bar{f}^{(0)}(u|\mathbf{x})\mathrm{d}u\right)\right|$$
$$\le\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\widehat{f}_{\mathrm{N}}(y|\mathbf{x})-f(y|\mathbf{x})-\bar{f}^{(0)}(y|\mathbf{x})\right|+\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\frac{\widehat{f}_{\mathrm{N}}(y|\mathbf{x})}{\int_{\mathcal{Y}}\widehat{f}_{\mathrm{N}}(u|\mathbf{x})\mathrm{d}u}-f(y|\mathbf{x})\right|\cdot\left|\int_{\mathcal{Y}}\bar{f}^{(0)}(u|\mathbf{x})\mathrm{d}u\right|$$

$$+ \sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\frac{\widehat{f}_{\mathrm{N}}(y|\mathbf{x})}{\int_{\mathcal{Y}}\widehat{f}_{\mathrm{N}}(u|\mathbf{x})\mathrm{d}u}\right| \cdot \left|\int_{\mathcal{Y}} \big(\widehat{f}_{\mathrm{N}}(u|\mathbf{x}) - f(u|\mathbf{x}) - \bar{f}^{(0)}(u|\mathbf{x})\big)\mathrm{d}u\right|$$

$$\lesssim_{\mathrm{TC}} \mathtt{r}_{\mathrm{SL}} + \left(h^{\mathfrak{p}} + \sqrt{\frac{\log(n)}{nh^{1+d}}}\right)\left(\sqrt{\frac{\log(n)}{nh^{d}}}\right) \precsim \mathtt{r}_{\mathrm{SL}}.$$

In the above, we have used the result that $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{V}[\int_{\mathcal{Y}} \bar{f}^{(0)}(u|\mathbf{x})\mathrm{d}u] \precsim (nh^d)^{-1}$, which shows that $\int_{\mathcal{Y}} \bar{f}^{(0)}(u|\mathbf{x})\mathrm{d}u$ has a smaller asymptotic order compared to $\bar{f}^{(0)}(y|\mathbf{x})$.

To prove Theorem 7, we combine the results in Lemma 3 and Theorem 1. The strong approximation for $\widehat{\mathbb{S}}_{\mathrm{N}}$ in Theorem 8 follows from Lemma 3 and Theorem 2. The strong approximation for $\widehat{\mathbb{S}}_{\mathrm{I}}$ also follows from Lemma 3, as the stochastic linearization of $\widehat{f}_{\mathrm{I}}$ is a linear functional of $\bar{f}^{(0)}$.

A.10. A result on covering number

In this appendix, we prove a general result on the uniform covering number for function classes consisting of kernels. Importantly, we allow the kernels in the function class to take different shapes and to depend on a range of bandwidths.

Lemma 7 (Covering number). *Let $h > 0$, and $\mathfrak{c} > 0$ be a (large) generic constant which does not depend on h. Define the class of functions*

$$\mathcal{G} = \left\{ g_{\mathbf{z}}\left(\frac{\cdot - \mathbf{z}}{ah}\right) : \ \mathbf{z} \in [0,1]^d, \ 1 \le a \le \mathfrak{c} \right\}.$$

Assume (i) boundedness: $\sup_{\mathbf{z},\mathbf{z}'} |g_{\mathbf{z}}(\mathbf{z}')| \le \mathfrak{c}$. *(ii) $g_{\mathbf{z}}(\cdot)$ is supported in $[-1,1]^d$ for all $\mathbf{z}$. (iii) Lipschitz continuity:* $\sup_{\mathbf{z}} |g_{\mathbf{z}}(\mathbf{z}') - g_{\mathbf{z}}(\mathbf{z}'')| \le \mathfrak{c}|\mathbf{z}' - \mathbf{z}''|$ *and* $\sup_{\mathbf{z}} |g_{\mathbf{z}'}(\mathbf{z}) - g_{\mathbf{z}''}(\mathbf{z})| \le \mathfrak{c}h^{-1}|\mathbf{z}' - \mathbf{z}''|$. *Then, for any probability measure P, the $L^1(P)$-covering number of the class $\mathcal{G}$ satisfies*

$$N\big((2\mathfrak{c}+1)^{d+1}\varepsilon, \ \mathcal{G}, \ L^1(P)\big) \le \mathfrak{c}' \frac{1}{\varepsilon^{d+2}} + 1,$$

where $\mathfrak{c}'$ is some constant that depends only on $\mathfrak{c}$ and d.

This rate, ε^{-d-2}, is clearly suboptimal for very small ε. The reason is that when we fix h and consider how the covering number changes as $\varepsilon \downarrow 0$, the optimal rate is ε^{-d-1}, as in this case the class of functions is fixed (c.f. Theorem 2.7.11 in [28]). Such suboptimality is introduced because we prefer a covering number that depends only on ε (but not h). The result we derived performs better for moderate and large ε (relative to h).

Now consider how the above (a sharper result for moderate and large ε) manifests itself in our proof below. Take a fixed ε. As the bandwidth shrinks to 0, we will be employing finer partitions of $[0,1]^d$. However, not all of the sets in the partition matter for bounding the covering number, because there are at most ε^{-1} sets carrying a probability mass larger than ε. Given that the functions we consider have compact support, most of them become irrelevant in our calculation of the covering number. Indeed, a function only makes a nontrivial contribution if its support intersects with some set in the (very fine) partition whose P-measure exceeds ε. Therefore, instead of considering all h^{-d} partitions, we only need to focus on ε^{-1} of them, which is why an extra ε^{-1} term is introduced.

Finally, from the definition of $\mathcal{G}$, it is clear that the covering number obtained above allows for a range of bandwidths (captured by ah with $1 \le a \le \mathfrak{c}$). If we instead consider the restricted function class, $\{g_{\mathbf{z}}((\cdot - \mathbf{z})/h) : \ \mathbf{z} \in [0,1]^d\}$, then a sharper bound will apply: $\mathfrak{c}'\varepsilon^{-d-1} + 1$.

Proof of Lemma 7. This proof strategy is motivated by Lemma 4.1 in [25]. Take $\ell = \lfloor 1/h \rfloor$, and partition each coordinate $[0,1]$ into ℓ intervals of equal length. This will lead to a partition $\mathcal{A} = \{A_j : 1 \le j \le \ell^d\}$ of $[0,1]^d$. Next, consider sets whose P-measure exceeds ε,

$$\mathcal{A}_{P,\varepsilon} = \{A \in \mathcal{A} : P[A] > \varepsilon\},$$

and their $\mathfrak{c}h$-enlargements

$$\mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon} = \{A + [-\mathfrak{c}h, \mathfrak{c}h]^d : A \in \mathcal{A}_{P,\varepsilon}\}.$$

Case 1: z does not belong to any set in $\mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon}$. This implies that the support of the function $g_{\mathbf{z}}\left(\frac{\cdot-\mathbf{z}}{ah}\right)$ will not intersect with any set in $\mathcal{A}_{P,\varepsilon}$. We also notice that

$$\int \left| g_{\mathbf{z}}\Big(\frac{\cdot-\mathbf{z}}{ah}\Big)\right| \mathrm{d}P \le \mathfrak{c}P\big[ah \cdot \mathrm{supp}(g_{\mathbf{z}}(\cdot)) + \mathbf{z}\big] \le \mathfrak{c}P\big[\mathfrak{c}h \cdot \mathrm{supp}(g_{\mathbf{z}}(\cdot)) + \mathbf{z}\big].$$

Define the complement of $\mathcal{A}_{P,\varepsilon}$ as $\mathcal{A}^{\perp}_{P,\varepsilon} = \{A \in \mathcal{A} : P[A] \le \varepsilon\}$, then the set $\mathfrak{c}h \cdot \mathrm{supp}(g_{\mathbf{z}}(\cdot)) + \mathbf{z}$ will be completely covered by sets in $\mathcal{A}^{\perp}_{P,\varepsilon}$. To determine the maximum number of intersections between $\mathfrak{c}h \cdot \mathrm{supp}(g_{\mathbf{z}}(\cdot)) + \mathbf{z}$ and sets in $\mathcal{A}^{\perp}_{P,\varepsilon}$, it suffices to consider the Euclidean volume of the enlarged set $\mathfrak{c}h \cdot \mathrm{supp}(g_{\mathbf{z}}(\cdot)) + \mathbf{z} + [-\ell^{-1}, \ell^{-1}]^d$, which is $(2\mathfrak{c}h + \ell^{-1})^d$. The Euclidean volume of each set in $\mathcal{A}^{\perp}_{P,\varepsilon}$ is ℓ^{-d}. Therefore, the set $\mathfrak{c}h \cdot \mathrm{supp}(g_{\mathbf{z}}(\cdot)) + \mathbf{z}$ can intersect with at most

$$\frac{(2\mathfrak{c}h + \ell^{-1})^d}{\ell^{-d}} = (2\mathfrak{c}h\ell + 1)^d \le (2\mathfrak{c}+1)^d$$

sets in $\mathcal{A}^{\perp}_{P,\varepsilon}$. As a result, we conclude that $\int |g_{\mathbf{z}}\left(\frac{\cdot-\mathbf{z}}{ah}\right)| \mathrm{d}P \le \mathfrak{c}\,(2\mathfrak{c}+1)^d\,\varepsilon$. This leads to our first result. Let $A^{\mathfrak{c}h}_{P,\varepsilon} = \cup\mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon}$ be the union of sets in $\mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon}$, then

$$N\Big((2\mathfrak{c}+1)^{d+1}\varepsilon,\ \mathcal{G}_1,\ L^1(P)\Big) = 1, \quad \text{where } \mathcal{G}_1 = \left\{ g_{\mathbf{z}}\left(\frac{\cdot-\mathbf{z}}{ah}\right) : \mathbf{z} \notin A^{\mathfrak{c}h}_{P,\varepsilon},\ 1 \le a \le \mathfrak{c}\right\}.$$

As remark, we note that the function class $\mathcal{G}_1$ changes with respect to h, ε, as well as the probability measure P.

Case 2: z belongs to some set in $\mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon}$. Each set in $\mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon}$ is a cube with edge length $\ell^{-1} + 2\mathfrak{c}h \le 2(\mathfrak{c}+1)h$, because $h\ell \ge 0.5$. Then the covering number of $A^{\mathfrak{c}h}_{P,\varepsilon}$ (under the Euclidean distance) is

$$N\Big(h\varepsilon,\ A^{\mathfrak{c}h}_{P,\varepsilon},\ |\cdot|\Big) \le \sum_{A \in \mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon}} N(h\varepsilon,\ A,\ |\cdot|) \le \mathrm{card}(\mathcal{A}^{\mathfrak{c}h}_{P,\varepsilon}) \cdot \mathfrak{c}'\frac{1}{\varepsilon^d} \le \mathfrak{c}'\frac{1}{\varepsilon^{d+1}}.$$

Here, $\mathfrak{c}'$ is some fixed number that only depends on $\mathfrak{c}$ and d. Using the Lipschitz property, we have

$$\int \left| g_{\mathbf{z}}\Big(\frac{\cdot-\mathbf{z}}{ah}\Big) - g_{\mathbf{z}'}\left(\frac{\cdot-\mathbf{z}'}{a'h}\right)\right| \mathrm{d}P \le 2\mathfrak{c}h^{-1}|\mathbf{z}-\mathbf{z}'| + \mathfrak{c}^2|a-a'|.$$

Now define $\mathcal{G}_2 = \mathcal{G}\backslash\mathcal{G}_1 = \{g_{\mathbf{z}}((\cdot-\mathbf{z})/(ah)) : \mathbf{z} \in A^{\mathfrak{c}h}_{P,\varepsilon},\ 1 \le a \le \mathfrak{c}\}$, then

$$N\Big((2\mathfrak{c}+1)^{d+1}\varepsilon,\ \mathcal{G}_2,\ L^1(P)\Big) \le N\Big(\frac{(2\mathfrak{c}+1)^{d+1}}{4\mathfrak{c}}h\varepsilon,\ A^{\mathfrak{c}h}_{P,\varepsilon},\ |\cdot|\Big) N\Big(\frac{(2\mathfrak{c}+1)^{d+1}}{2\mathfrak{c}^2}\varepsilon,\ [1,\mathfrak{c}],\ |\cdot|\Big) \le \frac{\mathfrak{c}'}{\varepsilon^{d+2}}.$$

This closes the proof. ■

A.11. Technical lemmas

Lemma 8 (Equation (3.5) in [17]). *Let $\{z_i, 1 \le i \le n\}$ be independent random variables, and $\{\tilde{z}_i, 1 \le i \le n\}$ be an independent copy of $\{z_i, 1 \le i \le n\}$. For a degenerate and decoupled second order U-statistic, $\sum_{i,j=1,i\neq j}^{n} u_{ij}(z_i, \tilde{z}_j)$, the following holds:*

$$\mathbb{P}\Big[\Big|\sum_{i,j,i\neq j}^{n} u_{ij}(z_i,\tilde{z}_j)\Big| > t\Big] \le \mathfrak{c}\exp\Big\{-\frac{1}{\mathfrak{c}}\min\Big[\frac{t}{A}, \Big(\frac{t}{B}\Big)^{\frac{2}{3}}, \Big(\frac{t}{C}\Big)^{\frac{1}{2}}\Big]\Big\},$$

where $\mathfrak{c}$ is some absolute constant, and A, B and C are any constants satisfying

$$A^2 \ge \sum_{i,j=1,i\neq j}^{n} \mathbb{E}[u_{ij}(z_i,\tilde{z}_j)^2], \quad B^2 \ge \max_{1\le i,j\le n}\Big[\sup_w\Big|\sum_{i=1}^{n}\mathbb{E}[u_{ij}(z_i,w)^2]\Big|, \sup_v\Big|\sum_{j=1}^{n}\mathbb{E}[u_{ij}(v,\tilde{z}_j)^2]\Big|\Big],$$
$$C \ge \max_{1\le i,j\le n}\sup_{v,w}|u_{ij}(v,w)|.$$

To apply the above lemma, an additional decoupling step is usually needed. Fortunately, the decoupling step only introduces an extra constant, but will not affect the order of the tail probability bound. Formally,

Lemma 9 ([11]). *Consider the setting of Lemma 8. Then*

$$\mathbb{P}\Big[\Big|\sum_{i,j,i\neq j}^{n} u_{ij}(z_i,z_j)\Big| > t\Big] \le \mathfrak{c}\cdot\mathbb{P}\Big[\mathfrak{c}\Big|\sum_{i,j,i\neq j}^{n} u_{ij}(z_i,\tilde{z}_j)\Big| > t\Big],$$

where $\mathfrak{c}$ is an absolute constant.

As a result, we will apply Lemma 8 without explicitly mentioning the decoupling step or the extra constant it introduces.

Lemma 10 (Theorem 1.1 in [25]). *Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ be iid random vectors with continuous and strictly positive density on $[0,1]^d$, and $d \ge 2$. Let $\mathcal{G}$ be a class of functions from $[0,1]^d$ to $[-1,1]$, satisfying $\sup_P N(\varepsilon, \mathcal{G}, L^1(P)) \le \mathfrak{c}_1\varepsilon^{-\mathfrak{c}_2}$, where the supremum is taken over all probability measures on $[0,1]^d$, and $\mathfrak{c}_1$ and $\mathfrak{c}_2$ are constants that can depend on $\mathcal{G}$. In addition, assume the following measurability condition holds: there exists a Suslin space $\mathcal{S}$ and a mapping $\mathbb{F}: \mathcal{S} \to \mathcal{G}$, such that $(s,\mathbf{z}) \mapsto \mathbb{F}(s,\mathbf{z})$ is measurable. Let*

$$\mathrm{TV}_{\mathcal{G}} = \sup_{g\in\mathcal{G}}\sup_{\phi\in C_1^\infty([0,1]^d)}\int_{[0,1]^d} g(\mathbf{z})\mathrm{div}\phi(\mathbf{z})\mathrm{d}\mathbf{z},$$

where div *is the divergence operator, and $C_1^\infty([0,1]^d)$ is the collection of infinitely differentiable functions with values in $\mathbb{R}^d$, support included in $[0,1]^d$, and supremum norm bounded by 1. Then on a possibly enlarged probability space, there exists a centered Gaussian process, $\mathbb{G}$, indexed by $\mathcal{G}$, such that (i) $\mathbb{C}\mathrm{ov}[\mathbb{G}(g), \mathbb{G}(g')] = \mathbb{C}\mathrm{ov}[g(\mathbf{z}_i), g'(\mathbf{z}_i)]$, and (ii) for any $t \ge \mathfrak{c}_3 \log n$,*

$$\mathbb{P}\Big[\sqrt{n}\sup_{g\in\mathcal{G}}|\mathbb{B}(g) - \mathbb{G}(g)| \ge \mathfrak{c}_3\sqrt{n^{\frac{d-1}{d}}\, t\, \mathrm{TV}_{\mathcal{G}}} + \mathfrak{c}_3 t\sqrt{\log(n)}\Big] \le e^{-t}.$$

In the above, $\mathbb{B} = \sum_{i=1}^{n}(g(\mathbf{z}_i) - \mathbb{E}[g(\mathbf{z}_i)])/\sqrt{n}$ *is the empirical process indexed by* $\mathcal{G}$*, and* $\mathfrak{c}_3$ *is some constant that only depends on* d*,* $\mathfrak{c}_1$*, and* $\mathfrak{c}_2$*.*

Lemma 11 (Corollary 5.1 in Chernozhukov *et al.* [9]). *Let* $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{\ell_n}$ *be two mean-zero Gaussian random vectors with covariance matrices* $\mathbf{\Omega}_1$ *and* $\mathbf{\Omega}_2$*, respectively. Further assume that the diagonal elements in* $\mathbf{\Omega}_1$ *are all one. Then*

$$\sup_{A \text{ rectangular}} |\mathbb{P}[\mathbf{z}_1 \in A] - \mathbb{P}[\mathbf{z}_2 \in A]| \leq \mathfrak{c}\sqrt{\|\mathbf{\Omega}_1 - \mathbf{\Omega}_2\|_\infty \log(\ell_n)},$$

where $\|\cdot\|_\infty$ *denotes the supremum norm, and* $\mathfrak{c}$ *is an absolute constant.*

Lemma 12 (Theorem 2.1 in [8]). *Let* $\mathbb{G}$ *be a centered and separable Gaussian process indexed by* $g \in \mathcal{G}$ *such that* $\mathbb{V}[\mathbb{G}(g)] = 1$ *for all* $g \in \mathcal{G}$*. Assume* $\sup_{g\in\mathcal{G}} \mathbb{G}(g) < \infty$ *almost surely. Define* $C_{\mathcal{G}} = \mathbb{E}[\sup_{g\in\mathcal{G}} \mathbb{G}(g)]$*. Then for all* $\varepsilon > 0$*,*

$$\sup_{u\in\mathbb{R}} \mathbb{P}\left[\left|\sup_{g\in\mathcal{G}} \mathbb{G}(g) - u\right| \leq \varepsilon\right] \leq 4\varepsilon(C_{\mathcal{G}} + 1).$$

Acknowledgments

The authors thank the editor, two anonymous reviewers, Jianqing Fan, Jason Klusowski, Will Underwood, Jingshen Wang, and Rae Yu for their thoughtful discussions and valuable feedback.

Funding

Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805 and DMS-2210561, and from the National Institute of Health (R01 GM072611-16).

Jansson gratefully acknowledges financial support from the National Science Foundation through grant SES-1947662 and the research support of CREATES.

Supplementary Material

Supplementary material to "Boundary adaptive local polynomial conditional density estimators"
The supplementary material [5] contains general theoretical results encompassing those discussed in the main paper, includes proofs of those general results, and discusses additional methodological and technical results.

References

[1] Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *J. Am. Stat. Assoc.* **113** 767-779.

[2] Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2022). Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli* **28** 2998–3022.

[3] Cattaneo, M. D., Jansson, M. and Ma, X. (2020). Simple local polynomial density estimators. *J. Am. Stat. Assoc.* **115** 1449–1455.

[4] CATTANEO, M. D., CHANDAK, R., JANSSON, M. and MA, X. (2022). `lpcde`: Local polynomial conditional density estimation and inference. Working paper.

[5] CATTANEO, M. D., CHANDAK, R., JANSSON, M. and MA, X. (2023). Supplementary material to "Boundary adaptive local polynomial conditional density estimators".

[6] CHENG, M.-Y. (1994). On boundary effects of smooth curve estimators. PhD thesis, Dept. Statistics, Univ. North Carolina, Chapel Hill.

[7] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *Ann. Stat.* **42** 1787–1818.

[8] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014b). Gaussian approximation of suprema of empirical processes. *Ann. Stat.* **42** 1564–1597.

[9] CHERNOZHUKOV, V., CHETVERIKOV, D., KATO, K. and KOIKE, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *Ann. Stat.* **50** 2562–2586.

[10] DE GOOIJER, J. G. and ZEROM, D. (2003). On conditional density estimation. *Stat. Neerl.* **57** 159–176.

[11] DE LA PEÑA, V. H. and MONTGOMERY-SMITH, S. J. (1995). Decoupling inequalities for the tail probabilities of multivariate U-statistics. *Ann. Probab.* **23** 806–816.

[12] EINMAHL, U. and MASON, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theor. Probab.* **13** 1–37.

[13] EINMAHL, U. and MASON, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Stat.* **33** 1380–1403.

[14] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.

[15] FAN, J., YAO, Q. and TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206.

[16] FERRIGNO, S., MAUMY-BERTRAND, M. and MULLER, A. (2010). Uniform law of the logarithm for the local linear estimator of the conditional distribution function. *C. R. Math.* **348** 1015–1019.

[17] GINÉ, E., LATAŁA, R. and ZINN, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II* Springer.

[18] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-dimensional Statistical Models*. Cambridge University Press.

[19] HALL, P. (1979). On the rate of convergence of normal extremes. *J. Appl. Probab.* **16** 433–439.

[20] HALL, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **55** 291–304.

[21] HALL, P., RACINE, J. and LI, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Am. Stat. Assoc.* **99** 1015–1026.

[22] HALL, P., WOLFF, R. C. and YAO, Q. (1999). Methods for estimating a conditional distribution function. *J. Am. Stat. Assoc.* **94** 154–163.

[23] KHAS'MINSKII, R. Z. (1979). A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theor. Probability Ap* **23** 794–798.

[24] KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV'-s, and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131.

[25] RIO, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21–45.

[26] SCOTT, D. W. (2015). *Multivariate density estimation: Theory, practice, and visualization*. John Wiley & Sons.

[27] SIMONOFF, J. S. (2012). *Smoothing Methods in Statistics*. Springer.

[28] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

[29] WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman & Hall/CRC.

[30] WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer.