Mathematics and Statistics Department Faculty Publication Series

Mathematics and Statistics

2020

Correcting an estimator of a multivariate monotone function with isotonic regression

Ted Westling

Mark J. van der Laan

Marco Carone

Electronic Journal of Statistics
Vol. 14 (2020) 3032–3069
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1740

Correcting an estimator of a multivariate monotone function with isotonic regression

Ted Westling[*]

Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, Massachusetts, USA
e-mail: twestling@math.umass.edu

Mark J. van der Laan[†]

Division of Biostatistics
University of California, Berkeley
Berkeley, California, USA
e-mail: laan@stat.berkeley.edu

Marco Carone[‡]

Department of Biostatistics
University of Washington
Seattle, Washington, USA
e-mail: mcarone@uw.edu

Abstract: In many problems, a sensible estimator of a possibly multivariate monotone function may fail to be monotone. We study the correction of such an estimator obtained via projection onto the space of functions monotone over a finite grid in the domain. We demonstrate that this corrected estimator has no worse supremal estimation error than the initial estimator, and that analogously corrected confidence bands contain the true function whenever the initial bands do, at no loss to band width. Additionally, we demonstrate that the corrected estimator is asymptotically equivalent to the initial estimator if the initial estimator satisfies a stochastic equicontinuity condition and the true function is Lipschitz and strictly monotone. We provide simple sufficient conditions in the special case that the initial estimator is asymptotically linear, and illustrate the use of these results for estimation of a G-computed distribution function. Our stochastic equicontinuity condition is weaker than standard uniform stochastic equicontinuity, which has been required for alternative correction procedures. This allows us to apply our results to the bivariate correction of the local linear estimator of a conditional distribution function known to be monotone in its conditioning argument. Our experiments suggest that the projection step can yield significant practical improvements.

[*] Supported by NIAID grant UM1AI068635.
[†] Supported by NIAID grant R01AI074345.
[‡] Supported by NHLBI grant R01HL137808.

MSC2020 subject classifications: Primary 62G20; secondary 60G15.
Keywords and phrases: Asymptotic linearity, confidence band, kernel smoothing, projection, shape constraint, stochastic equicontinuity.

Received December 2019.

Contents

1. Introduction

1.1. Background

In many scientific problems, the parameter of interest is a component-wise monotone function. In practice, an estimator of this function may have several desirable statistical properties, yet fail to be monotone. This often occurs when the estimator is obtained through the pointwise application of a statistical procedure over the domain of the function. For instance, we may be interested in estimating a conditional cumulative distribution function θ_0, defined pointwise as $\theta_0(a, y) = P_0(Y \leq y \mid A = a)$, over its domain $\mathcal{D} \subset \mathbb{R}^2$. Here, Y may represent an outcome and A an exposure. The map $y \mapsto \theta_0(a, y)$ is necessarily monotone for each fixed a. In some scientific contexts, it may be known that $a \mapsto \theta_0(a, y)$

is also monotone for each y, in which case θ_0 is a bivariate component-wise monotone function. An estimator of θ_0 can be constructed by estimating the regression function $(a, y) \mapsto E_{P_0}[I(Y \leq y) \mid A = a]$ for each (a, y) on a finite grid using kernel smoothing, and performing suitable interpolation elsewhere. For some types of kernel smoothing, including the Nadaraya-Watson estimator, the resulting estimator is necessarily monotone as a function of y for each value of a, but not necessarily monotone as a function of a for each value of y. For other types of kernel smoothing, including the local linear estimator, which often has smaller asymptotic bias than the Nadaraya-Watson estimator, the resulting estimator need not be monotone in either component.

Whenever the function of interest is component-wise monotone, failure of an estimator to itself be monotone can be problematic. This is most apparent if the monotonicity constraint is probabilistic in nature – that is, the parameter mapping is monotone under all possible probability distributions. This is the case, for instance, if θ_0 is a distribution function. In such settings, returning a function estimate that fails to be monotone is nonsensical, like reporting a probability estimate outside the interval $[0, 1]$. However, even if the monotonicity constraint is based on scientific knowledge rather than probabilistic constraints, failure of an estimator to be monotone can be an issue. For example, if the parameter of interest represents average height or weight among children as a function of age, scientific collaborators would likely be unsatisfied if presented with an estimated curve that were not monotone. Finally, as we will see, there are often finite-sample performance benefits to ensuring that the monotonicity constraint is respected.

Whenever this phenomenon occurs, it is natural to seek an estimator that respects the monotonicity constraint but nevertheless remains close to the initial estimator, which may otherwise have good statistical properties. A monotone estimator can be naturally constructed by projecting the initial estimator onto the space of monotone functions with respect to some norm. A common choice is the L_2-norm, which amounts to using multivariate isotonic regression to correct the initial estimator.

1.2. Contribution and organization of the article

In this article, we discuss correcting an initial estimator of a multivariate monotone function by computing the isotonic regression of the estimator over a finite grid in the domain, and interpolating between grid points. We also consider correcting an initial confidence band by using the same procedure applied to the upper and lower limits of the band. We provide three general results regarding this simple procedure.

1. Building on the results of Robertson, Wright and Dykstra (1988) and Chernozhukov, Fernández-Val and Galichon (2009), we demonstrate that the corrected estimator is at least as good as the initial estimator, meaning:
   (a) its uniform error over the grid used in defining the projection is less than or equal to that of the initial estimator for every sample;

   (b) its uniform error over the entire domain is less than or equal to that of the initial estimator asymptotically;
   (c) the corrected confidence band contains the true function on the projection grid whenever the initial band does, at no cost in terms of average or uniform band width.

2. We provide high-level sufficient conditions under which the uniform difference between the initial and corrected estimators is $o_P(r_n^{-1})$ for a generic sequence $r_n \to \infty$.
3. We provide simpler lower-level sufficient conditions in two special cases:
   (a) when the initial estimator is uniformly asymptotically linear, in which case the appropriate rate is $r_n = n^{1/2}$;
   (b) when the initial estimator is kernel-smoothed with bandwidth h_n, in which case the appropriate rate is $r_n = (nh_n)^{1/2}$ for univariate kernel smoothing.

We apply our theoretical results to two sets of examples: nonparametric efficient estimation of a G-computed distribution function for a binary exposure, and local linear estimation of a conditional distribution function with a continuous exposure.

Other authors have considered the correction of an initial estimator using isotonic regression. To name a few, Mukarjee and Stern (1994) used a projection-like procedure applied to a kernel smoothing estimator of a regression function, whereas Patra and Sen (2016) used the projection procedure applied to a univariate cumulative distribution function in the context of a mixture model. These articles addressed the properties of the projection procedure in their specific applications. In contrast, we provide general results that are applicable broadly.

1.3. Alternative projection procedures

The projection approach is not the only possible correction procedure. Dette, Neumeyer and Pilz (2006), Chernozhukov, Fernández-Val and Galichon (2009), and Chernozhukov, Fernández-Val and Galichon (2010) studied a correction based on monotone rearrangements. However, monotone rearrangements do not generalize to the multivariate setting as naturally as projections — for example, Chernozhukov, Fernández-Val and Galichon (2009) proposed averaging a variety of possible multivariate monotone rearrangements to obtain a final monotone estimator. In contrast, the L_2 projection of an initial estimator onto the space of component-wise monotone functions is uniquely defined, even in the context of multivariate functions.

Daouia and Park (2013) proposed an alternative correction procedure that consists of taking a convex combination of upper and lower monotone envelope functions, and they demonstrated conditions under which their estimator is asymptotically equivalent in supremum norm to the initial estimator. There are several differences between our contributions and those of Daouia and Park (2013). For instance, Daouia and Park (2013) did not study correction of confidence bands, which we consider in Section 2.3, or the important special case of asymptotically linear estimators, which we consider in Section 3.1. Our results in these two sections apply equally well to our correction procedure and to the correction procedure considered by Daouia and Park (2013).

Perhaps the most important theoretical contribution of our work beyond that of existing research is the weaker form of stochastic equicontinuity that we require for establishing asymptotic equivalence of the initial and projected estimators. In contrast, Daouia and Park (2013) explicitly required the usual uniform asymptotic equicontinuity, while application of the Hadamard differentiability results of Chernozhukov, Fernández-Val and Galichon (2010) requires weak convergence to a tight limit, which is stronger than uniform asymptotic equicontinuity. Our weaker condition allows us to use our general results to tackle a broader range of initial estimators, including kernel smoothed estimators, which are typically not uniformly asymptotically equicontinuous at useful rates, but nevertheless can frequently be shown to satisfy our condition. We discuss this in detail in Section 3.2. We illustrate this general contribution in Section 4.2 by studying the bivariate correction of a conditional distribution function estimated using local linear regression, which would not be possible using the stronger asymptotic equicontinuity condition. In numerical studies, we find that the projected estimator and confidence bands can offer substantial finite-sample improvements over the initial estimator and bands in this example.

2. Main results

2.1. *Definitions and statistical setup*

Let $\mathcal{M}$ be a statistical model of probability measures on a probability space $(\mathcal{X}, \mathcal{B})$. Let $\theta : \mathcal{M} \to \ell^\infty(\mathcal{T})$ be a parameter of interest on $\mathcal{M}$, where $\mathcal{T} := [0, 1]^d$ and $\ell^\infty(\mathcal{T})$ is the Banach space of bounded functions from $\mathcal{T}$ to $\mathbb{R}$ equipped with supremum norm $\| \cdot \|_{\mathcal{T}}$. We have specified this particular $\mathcal{T}$ for simplicity, but the results established here apply to any bounded rectangular domain $\mathcal{T} \subset \mathbb{R}^d$. For each $P \in \mathcal{M}$, denote by θ_P the evaluation of θ at P and note that θ_P is a bounded real-valued function on $\mathcal{T}$. For any $t \in \mathcal{T}$, denote by $\theta_P(t) \in \mathbb{R}$ the evaluation of θ_P at t.

For any vector $t \in \mathbb{R}^d$ and $1 \leq j \leq d$, denote by t_j the j^{th} component of t. Define the partial order $\leq$ on $\mathbb{R}^d$ by setting $t \leq t'$ if and only if $t_j \leq t'_j$ for each $1 \leq j \leq d$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (component-wise) monotone non-decreasing if $t \leq t'$ implies that $f(t) \leq f(t')$. Denote $\|t\| = \max_{1 \leq j \leq d} |t_j|$ for any vector $t \in \mathbb{R}^d$. Additionally, denote by $\boldsymbol{\Theta} \subset \ell^\infty(\mathcal{T})$ the convex set of bounded monotone non-decreasing functions from $\mathcal{T}$ to $\mathbb{R}$. For concreteness, we focus on non-decreasing functions, but all results established here apply equally to non-increasing functions.

Let $\mathcal{M}_0 := \{P \in \mathcal{M} : \theta_P \in \boldsymbol{\Theta}\} \subseteq \mathcal{M}$ and suppose that $\mathcal{M}_0$ is nonempty. Generally, this inclusion is strict only if, rather than being implied by the rules of probability, the monotonicity constraint stems at least in part from prior scientific knowledge. Also, define $\mathbf{\Theta}_0 := \{\theta \in \mathbf{\Theta} : \theta = \theta_P \text{ for some } P \in \mathcal{M}\} \subseteq \mathbf{\Theta}$. We are primarily interested in settings where $\mathbf{\Theta}_0 = \mathbf{\Theta}$, since in this case there is no additional knowledge about θ encoded by $\mathcal{M}$, and in particular there is no danger of yielding a corrected estimator that is compatible with no $P \in \mathcal{M}$.

Suppose that observations $X_1, X_2, \dots, X_n$ are sampled independently from an unknown distribution $P_0 \in \mathcal{M}_0$, and that we wish to estimate $\theta_0 := \theta_{P_0}$ based on these observations. Suppose that, for each $t \in \mathcal{T}$, we have access to an estimator $\theta_n(t)$ of $\theta_0(t)$ based on $X_1, X_2, \dots, X_n$. We note that the assumption that the data are independent and identically distributed is not necessary for Theorems 1 and 2 below. For any suitable $f : \mathcal{X} \to \mathbb{R}$, we define $Pf := \int f(x)\, P(dx)$ and $\mathbb{G}_n f := n^{1/2} \int f(x)(\mathbb{P}_n - P_0)(dx)$, where $\mathbb{P}_n$ is the empirical distribution based on $X_1, X_2, \dots, X_n$.

The central premise of this article is that $\theta_n(t)$ may have desirable statistical properties for each t or even uniformly in t, but that θ_n as an element of $\ell^\infty(\mathcal{T})$ may not fall in $\mathbf{\Theta}$ for any finite n or even with probability tending to one. Our goal is to provide a corrected estimator θ_n^* that necessarily falls in $\mathbf{\Theta}$, and yet retains the statistical properties of θ_n. A natural way to accomplish this is to define θ_n^* as the closest element of $\mathbf{\Theta}$ to θ_n in some norm on $\mathcal{T}$. Ideally, we would prefer to take θ_n^* to minimize $\|\theta - \theta_n\|_{\mathcal{T}}$ over $\theta \in \mathbf{\Theta}$. However, this is not tractable for two reasons. First, optimization over the entirety of $\mathcal{T}$ is an infinite-dimensional optimization problem, and is hence frequently computationally intractable. To resolve this issue, for each n, we let $\mathcal{T}_n = \{t_1, t_2, \dots, t_{m_n}\} \subseteq \mathcal{T}$ be a finite rectangular lattice in $\mathcal{T}$ over which we will perform the optimization, and define and consider $\|\cdot\|_{\mathcal{T}_n}$ as the supremum norm over $\mathcal{T}_n$. While it is now computationally feasible to define $\theta_{n,\infty}^*$ as a minimizer over $\theta \in \mathbf{\Theta}$ of the finite-dimensional objective function $\|\theta - \theta_n\|_{\mathcal{T}_n}$, this objective function is challenging due to its non-differentiability. Instead, we define

$$\theta_n^* \in \operatorname*{argmin}_{\theta \in \mathbf{\Theta}} \sum_{t \in \mathcal{T}_n} [\theta(t) - \theta_n(t)]^2 \ . \tag{2.1}$$

The squared-error objective function is smooth in its arguments. In dimension $d = 1$, θ_n^* thus defined is simply the isotonic regression of θ_n on the grid $\mathcal{T}_n$, which has a closed-form representation as the greatest convex minorant of the so-called cumulative sum diagram. Furthermore, since $\|\theta_n^* - \theta_n\|_{\mathcal{T}_n} \geq \|\theta_{n,\infty}^* - \theta_n\|_{\mathcal{T}_n}$, many of our results also apply to $\theta_{n,\infty}^*$.

We note that θ_n^* is only uniquely defined on $\mathcal{T}_n$. To completely characterize θ_n^*, we must monotonically interpolate function values between elements of $\mathcal{T}_n$. We will permit any monotonic interpolation that satisfies a weak condition. By the definition of a rectangular lattice, every $t \in \mathcal{T}$ can be assigned a hyper-rectangle whose vertices $\{s_1, s_2 \dots, s_{2^d}\}$ are elements of $\mathcal{T}_n$ and whose interior has empty intersection with $\mathcal{T}_n$. If multiple such hyper-rectangles exist for t, such as when t lies on the boundary of two or more such hyper-rectangles, one can be assigned arbitrarily. We will assume that, for $t \notin \mathcal{T}_n$,

$$\theta_n^*(t) = \sum_k \lambda_{k,n}(t)\theta_n^*(s_k)$$

for weights $\lambda_{1,n}(t), \lambda_{2,n}(t), \dots, \lambda_{2^d,n}(t) \in (0,1)$ such that $\sum_k \lambda_{k,n}(t) = 1$. In words, we assume that $\theta_n^*(t)$ is a convex combination of the values of θ_n^* on the vertices of the hyper-rectangle containing t. A simple interpolation approach consists of setting $\theta_n^*(t) = \theta_n^*(t')$ with t' the element of $\mathcal{T}_n$ closest to t, and choosing any such element if there are multiple elements of $\mathcal{T}_n$ equally close to t. This particular scheme satisfies our requirement.

Finally, for each n, we let $\ell_n(t) \leq u_n(t)$ denote lower and upper endpoints of a confidence band for $\theta_0(t)$. We then define ℓ_n^* and u_n^* as the corrected versions of ℓ_n and u_n using the same projection and interpolation procedure defined above for obtaining θ_n^* from θ_n.

In dimension $d = 1$, $\theta_n^*(t)$, $\ell_n^*(t)$, and $u_n^*(t)$ can be obtained for $t \in \mathcal{T}_n$ via the Pool Adjacent Violators Algorithm (Ayer et al., 1955), as implemented in the `R` command `isoreg` (R Core Team, 2018). In dimension $d = 2$, the corrections can be obtained using the algorithm described in Bril et al. (1984), which is implemented in the `R` command `biviso` in the package `Iso` (Turner, 2015). In dimension $d \geq 3$, Kyng, Rao and Sachdeva (2015) provides algorithms for computing the isotonic regression based on embedding the points in a directed acyclic graph. Alternatively, general-purpose algorithms for minimization of quadratic criteria over convex cones have been developed and implemented in the `R` package `coneproj` and may be used in this case (Meyer, 1999; Liao and Meyer, 2014).

2.2. Properties of the projected estimator

The projected estimator θ_n^* is the isotonic regression of θ_n over the grid $\mathcal{T}_n$. Hence, many existing finite-sample results on isotonic regression can be used to deduce properties of θ_n^*. Theorem 1 below collects a few of these properties, building upon the results of Barlow et al. (1972) and Chernozhukov, Fernández-Val and Galichon (2009). We denote $\omega_n := \sup_{t\in\mathcal{T}} \min_{s\in\mathcal{T}_n} \|t-s\|$ as the mesh of $\mathcal{T}_n$ in $\mathcal{T}$.

Theorem 1. *(i) It holds that* $\|\theta_n^* - \theta_0\|_{\mathcal{T}_n} \leq \|\theta_n - \theta_0\|_{\mathcal{T}_n}$.

(ii) If $\omega_n = o_P(1)$ *and* θ_0 *is continuous on* $\mathcal{T}$, *then*

$$\|\theta_n^* - \theta_0\|_{\mathcal{T}} \leq \|\theta_n - \theta_0\|_{\mathcal{T}_n} + o_P(1) \ .$$

(iii) If there exists some $\alpha > 0$ *for which* $\sup_{s,t\in\mathcal{T}:\|t-s\|\leq\delta} |\theta_0(t) - \theta_0(s)| = o(\delta^\alpha)$ *as* $\delta \to 0$, *then*

$$\|\theta_n^* - \theta_0\|_{\mathcal{T}} \leq \|\theta_n - \theta_0\|_{\mathcal{T}} + o_P(\omega_n^\alpha) \ .$$

(iv) If $\theta_0(t) \in [\ell_n(t), u_n(t)]$ *for all* $t \in \mathcal{T}_n$, *then* $\theta_0(t) \in [\ell_n^*(t), u_n^*(t)]$ *for all* $t \in \mathcal{T}_n$.

(v) It holds that $\|u_n^* - \ell_n^*\|_{\mathcal{T}_n} \leq \|u_n - \ell_n\|_{\mathcal{T}_n}$ *and*

$$\sum_{t\in\mathcal{T}_n} [u_n^*(t) - \ell_n^*(t)] = \sum_{t\in\mathcal{T}_n} [u_n(t) - \ell_n(t)] \ .$$

Theorem 1 is proved in Appendix A.1. We remark briefly on the implications of Theorem 1. Part (i) says that the estimation error of θ_n^* over the grid $\mathcal{T}_n$ is never worse than that of θ_n, whereas parts (ii) and (iii) provide bounds on the estimation error of θ_n^* on all of $\mathcal{T}$ in supremum norm. In particular, part (ii) indicates that θ_n^* is uniformly consistent on $\mathcal{T}$ as long as θ_n is uniformly consistent on $\mathcal{T}$, θ_0 is continuous on $\mathcal{T}$, and $\omega_n = o_P(1)$. Part (iii) provides an upper bound on the uniform rate of convergence of $\theta_n^* - \theta_0$, and indicates that if θ_0 is known to lie in a Hölder class, then ω_n can be chosen in such a way as to guarantee that the estimation error of θ_n^* on all of $\mathcal{T}$ is asymptotically no worse than the estimation error of θ_n on $\mathcal{T}_n$ in supremum norm. We note that parts (i)–(iii) also hold for the L_p norm with respect to uniform measure on $\mathcal{T}$ for any $p \in [1, \infty)$. Part (iv) guarantees that the isotonized band $[\ell_n^*, u_n^*]$ never has worse coverage than the original band over $\mathcal{T}_n$. Finally, part (v) states that the potential increase in coverage comes at no cost to the average or supremum width of the bands over $\mathcal{T}_n$. We note that parts (i), (iv) and (v) hold true for each n.

While comprehensive in scope, Theorem 1 does not rule out the possibility that θ_n^* performs strictly better, even asymptotically, than θ_n, or that the band $[\ell_n^*, u_n^*]$ is asymptotically strictly more conservative than $[\ell_n, u_n]$. In order to construct confidence intervals or bands with correct asymptotic coverage, a stronger result is needed: it must be that $\|\theta_n^* - \theta_n\|_{\mathcal{T}} = o_P(r_n^{-1})$, where r_n is a diverging sequence such that $r_n \|\theta_n - \theta_0\|_{\mathcal{T}}$ converges in distribution to a nondegenerate limit distribution. Then, we would have that $r_n \|\theta_n^* - \theta_0\|_{\mathcal{T}}$ converges in distribution to this same limit, and hence confidence bands constructed using approximations of this limit distribution would have correct coverage when centered around θ_n^*, as we discuss more below.

We consider the following conditions on θ_0 and the initial estimator θ_n:

(A) there exists a deterministic sequence r_n tending to infinity such that, for all $\delta > 0$,

$$\sup_{\|t-s\|<\delta/r_n} \left| r_n \left[\theta_n(t) - \theta_0(t)\right] - r_n \left[\theta_n(s) - \theta_0(s)\right] \right| = o_P(1);$$

(B) there exists $K_1 < \infty$ such that $|\theta_0(t) - \theta_0(s)| \leq K_1 \|t - s\|$ for all $t, s \in \mathcal{T}$;

(C) there exists $K_0 > 0$ such that $K_0 \|t - s\| \leq |\theta_0(t) - \theta_0(s)|$ for all $t, s \in \mathcal{T}$.

Based on these conditions, we have the following result.

Theorem 2. *If (A)–(C) hold and $\omega_n = o_P(r_n^{-1})$, then $\|\theta_n^* - \theta_n\|_{\mathcal{T}} = o_P(r_n^{-1})$.*

The proof of Theorem 2 is presented in Appendix A.2. This result indicates that the projected estimator is uniformly asymptotically equivalent to the original estimator in supremum norm at the rate r_n.

Condition (A) is related to, but notably weaker than, uniform stochastic equicontinuity (van der Vaart and Wellner, 1996, p. 37). (A) follows if, in particular, the process $\{r_n[\theta_n(t) - \theta_0(t)] : t \in \mathcal{T}\}$ converges weakly to a tight limit in the space $\ell^\infty(\mathcal{T})$. However, the latter condition is sufficient but not necessary for (A) to hold. This is important for application of our results to kernel smoothing estimators, which typically do not converge weakly to a tight limit, but for which condition (A) nevertheless often holds. We discuss this at length in Section 4.2. The results of Daouia and Park (2013) (see in particular condition (C3) therein) and Chernozhukov, Fernández-Val and Galichon (2010) rely on uniform stochastic equicontinuity in demonstrating asymptotic equivalence of their correction procedures, which essentially limits the applicability of their procedures to estimators that converge weakly to a tight limit in $\ell^\infty(\mathcal{T})$.

Condition (B) constrains θ_0 to be Lipschitz. Condition (C) constrains the variation of θ_0 from below, and is slightly more restrictive than a requirement for strict monotonicity. If, for instance, θ_0 is differentiable, then (C) is satisfied if all first-order partial derivatives of θ_0 are bounded away from zero. Condition (C) excludes, for instance, situations in which θ_0 is differentiable with null derivative over an interval. In such cases, θ_n^* may have strictly smaller variance on these intervals than θ_n because θ_n^* will pool estimates across the flat region while θ_n may not. Hence, in such cases, θ_n^* may potentially asymptotically improve on θ_n, so that θ_n^* and θ_n are not asymptotically equivalent at the rate r_n. Theoretical results in these cases would be of interest, but are beyond the scope of this article.

In addition to conditions (A)–(C), Theorem 2 requires that the mesh ω_n of $\mathcal{T}_n$ tend to zero in probability faster than r_n^{-1}. Since $\mathcal{T}_n$ is chosen by the user, as long as r_n (or an upper bound thereof) is known, this is not a problem in practice. Furthermore, except in irregular problems, the rate of convergence is typically not faster than $n^{-1/2}$, and hence it is typically sufficient to set $\omega_n = c_n n^{-1/2}$ for some $c_n = o(1)$. We note, however, that the computational complexity of obtaining the isotonic regression of θ_n over $\mathcal{T}_n$ increases as ω_n decreases. Hence, in cases where the rate of convergence of the initial estimator is strictly slower than $n^{-1/2}$, it may be preferable to choose ω_n more carefully based on a precise determination of r_n. We expect this to be especially true in the context of large d and n.

We note that conditions (A)–(C) and $\omega_n = o_P(r_n^{-1})$ also imply that $\|\theta_n^* - \theta_n\|_{L_p(\mathcal{T})} = o_P(r_n^{-1})$ for any $p \in [1, \infty)$, where $L_p(\mathcal{T})$ is the L_p norm on $\mathcal{T}$ with respect to uniform measure on $\mathcal{T}$. However, it may be possible to relax conditions (A)–(C) for the purpose of demonstrating L_p asymptotic equivalence of θ_n^* and θ_n for $p < \infty$. It is not clear whether our method of proof of Theorem 2 is amenable to such weakening. We have chosen to focus on uniform asymptotic equivalence in part for its use in constructing uniform confidence bands for θ_0, as we discuss in the next section.

2.3. Construction of confidence bands

Suppose there exists a fixed function $\gamma_\alpha : \mathcal{T} \to \mathbb{R}$ such that ℓ_n and u_n satisfy:

(a) $\|r_n(\theta_n - \ell_n) - \gamma_\alpha\|_{\mathcal{T}} \to_{\mathrm{P}} 0$;
(b) $\|r_n(u_n - \theta_n) - \gamma_\alpha\|_{\mathcal{T}} \to_{\mathrm{P}} 0$;
(c) $P_0\left[r_n|\theta_n(t) - \theta_0(t)| \geq \gamma_\alpha(t) \text{ for all } t \in \mathcal{T}\right] \longrightarrow 1 - \alpha$.

As an example of a confidence band that satisfies conditions (a)–(c), suppose that $\sigma_0 : \mathcal{T} \to (0, +\infty)$ is a scaling function and c_α is a fixed constant such that, as n tends to infinity,

$$P_0 \left(r_n \left\| \frac{\theta_n - \theta_0}{\sigma_0} \right\|_{\mathcal{T}} \geq c_\alpha \right) \longrightarrow 1 - \alpha \ .$$

If σ_n is an estimator of σ_0 satisfying $\|\sigma_n - \sigma_0\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ and $c_{\alpha,n}$ is an estimator of c_α such that $c_{\alpha,n} \to_{\mathrm{P}} c_\alpha$, then the Wald-type band defined by lower and upper endpoints $\ell_n(t) := \theta_n(t) - c_{\alpha,n} r_n^{-1} \sigma_n(t)$ and $u_n(t) := \theta_n(t) + c_\alpha r_n^{-1} \sigma_n(t)$ satisfies (a)–(c) with $\gamma_\alpha = c_\alpha \sigma_0$. However, the latter conditions can also be satisfied by other types of bands, such as those constructed with a consistent bootstrap procedure.

Under conditions (a)–(c), the confidence band $[\ell_n, u_n]$ has asymptotic coverage $1 - \alpha$. When conditions (A) and (B) also hold, the corrected band $[\ell_n^*, u_n^*]$ has the same asymptotic coverage as the original band $[\ell_n, u_n]$, as stated in the following result.

Corollary 1. *If (A)–(B) and (a)–(c) hold, γ_α is uniformly continuous on $\mathcal{T}$, and $\omega_n = o_P(r_n^{-1})$, then the band $[\ell_n^*, u_n^*]$ has asymptotic coverage $1 - \alpha$.*

The proof of Corollary 1 is presented in Appendix A.3. We also note that Theorem 2 implies that Wald-type confidence bands constructed around θ_n have the same asymptotic coverage if they are constructed around θ_n^* instead.

3. Refined results under additional structure

In this section, we provide more detailed conditions that imply condition (A) in two special cases: when θ_n is asymptotically linear, and when θ_n is a kernel smoothing-type estimator.

3.1. Special case I: asymptotically linear estimators

Suppose that the initial estimator θ_n is uniformly asymptotically linear (UAL): for each $t \in \mathcal{T}$, there exists $\phi_{0,t} : \mathcal{X} \mapsto \mathbb{R}$ depending on P_0 such that $\int \phi_{0,t} \, dP_0 = 0$, $\int \phi_{0,t}^2 \, dP_0 < \infty$, and

$$\theta_n(t) = \theta_0(t) + \frac{1}{n} \sum_{i=1}^{n} \phi_{0,t}(X_i) + R_{n,t} \tag{3.1}$$

for a remainder term $R_{n,t}$ with $n^{1/2} \sup_{t \in \mathcal{T}} |R_{n,t}| = o_P(1)$. The function $\phi_{0,t}$ is the influence function of $\theta_n(t)$ under sampling from P_0. It is desirable for θ_n to have representation (3.1) because this implies its uniform weak consistency as well as the pointwise asymptotic normality of $n^{1/2} [\theta_n(t) - \theta_0(t)]$ for each $t \in \mathcal{T}$. If in addition the collection $\{\phi_{0,t} : t \in \mathcal{T}\}$ of influence functions forms a P_0-Donsker class, then $\{n^{1/2} [\theta_n(t) - \theta_0(t)] : t \in \mathcal{T}\}$ converges weakly in $\ell^\infty(\mathcal{T})$ to a

Gaussian process with covariance function $\Sigma_0 : (t, s) \mapsto \int \phi_{0,t}(x)\phi_{0,s}(x)dP_0(x)$. Uniform asymptotic confidence bands based on θ_n can then be formed by using appropriate quantiles from any suitable approximation of the distribution of the supremum of the limiting Gaussian process.

We introduce two additional conditions:

(A1) the collection $\{\phi_{0,t} : t \in \mathcal{T}\}$ of influence curves is a P_0-Donsker class;
(A2) Σ_0 is uniformly continuous in the sense that

$$\lim_{\|t-s\|\to 0} \sup |\Sigma_0(s, t) - \Sigma_0(t, t)| = 0 \ .$$

Whenever θ_n is uniformly asymptotically linear, Theorem 2 can be shown to hold under (A1), (A2) and (B), as implied by the theorem below. The validity of (A1) and (A2) can be assessed by scrutinizing the influence function $\phi_{0,t}$ of $\theta_n(t)$ for each $t \in \mathcal{T}$. This fact renders the verification of these conditions very simple once uniform asymptotic linearity has been established.

Theorem 3. *For any UAL estimator θ_n, (A1)–(A2) together imply (A).*

The proof of Theorem 3 is provided in Appendix A.4. In Section 4.1, we illustrate the use of Theorem 3 for the estimation of a G-computed distribution function.

We note that conditions (A1)–(A2) are actually sufficient to establish uniform asymptotic equicontinuity, which as discussed above is stronger than (A). Therefore, Theorem 3 can also be used to prove asymptotic equivalence of the majorization/minorization correction procedure studied in Daouia and Park (2013).

3.2. Special case II: kernel smoothed estimators

For certain parameters, asymptotically linear estimators are not available. In particular, this is the case when the parameter of interest is not sufficiently smooth as a mapping of P_0. For example, density functions, regression functions, and conditional quantile functions do not permit asymptotically linear estimators in a nonparametric model when the exposure is continuous. In these settings, a common approach to nonparametric estimation is kernel smoothing.

Recent results suggest that, as a process, the only possible weak limit of $\{r_n[\theta_n(t) - \theta_0(t)] : t \in \mathcal{T}\}$ in $\ell^\infty(\mathcal{T})$ is zero when θ_n is a kernel smoothed estimator. For example, in the case of the Parzen-Rosenblatt density estimator with bandwidth h_n, Theorem 3 of Stupfler (2016) implies that if

$$c_n := r_n \left(nh_n/|\log h_n|\right)^{-1/2} \to 0,$$

then $\{r_n[\theta_n(t) - \theta_0(t)] : t \in \mathcal{T}\}$ converges weakly to zero in $\ell^\infty(\mathcal{T})$, whereas if $c_n \to c \in (0, \infty]$, then it does not converge weakly to a tight limit in $\ell^\infty(\mathcal{T})$. As a result, $\{r_n[\theta_n(t) - \theta_0(t)] : t \in \mathcal{T}\}$ only satisfies uniform stochastic equicontinuity for r_n such that $c_n \to 0$. However, for any such rate r_n, r_n^{-1} is slower than the pointwise and uniform rates of convergence of $\theta_n - \theta_0$. As a result, θ_n and θ_n^* may not be asymptotically equivalent at the uniform rate of convergence of $\theta_n - \theta_0$, so that confidence intervals and regions based on the limit distribution of $\theta_n - \theta_0$, but centered around θ_n^*, may not have correct coverage. We note that, while Stupfler (2016) establishes formal results for the Parzen-Rosenblatt estimator, we expect that the results therein extend to a variety of kernel smoothed estimators.

As a result of the lack of uniform stochastic equicontinuity of $r_n(\theta_n - \theta_0)$ for useful rates r_n, establishing (A) is much more difficult for kernel smoothed estimators than for asymptotically linear estimators. However, since (A) is weaker than uniform stochastic equicontinuity, it may still be possible. Here, we provide alternative sufficient conditions that imply condition (A) and that we have found useful for studying a kernel smoothed estimator θ_n.

When the initial estimator θ_n is kernel smoothed, we can often show that

$$\sup_{t \in \mathcal{T}} \left| r_n \left[\theta_n(t) - \theta_0(t)\right] - a_n b_0(t) - R_n(t) \right| \xrightarrow{\mathrm{P}} 0 \ , \tag{3.2}$$

where $b_0 : \mathcal{T} \to \mathbb{R}$ is a deterministic bias, a_n is sequences of positive constants, and $R_n : \mathcal{T} \to \mathbb{R}$ is a random remainder term. We then have that

$$\begin{aligned} &\sup_{\|t-s\|<\delta/r_n} \left| r_n \left[\theta_n(t) - \theta_0(t)\right] - r_n\left[\theta_n(s) - \theta_0(s)\right] \right| \\ &\quad = \sup_{\|t-s\|<\delta/r_n} a_n \left| b_0(t) - b_0(s) \right| + \sup_{\|t-s\|<\delta/r_n} \left| R_n(t) - R_n(s) \right| + o_P(1) \ . \end{aligned}$$

If b_0 is uniformly continuous on $\mathcal{T}$ and $a_n = O(1)$, or b_0 is uniformly α-Hölder on $\mathcal{T}$ and $a_n = O\left(r_n^{\alpha}\right)$, then the first term on the right-hand side tends to zero in probability. Attention may then be turned to demonstrating that the second term vanishes in probability. It appears difficult to provide a general characterization of the form of R_n that encompasses kernel smoothed estimators. However, in our experience, it is frequently the case that $R_n(t)$ involves terms of the form $\mathbb{G}_n \nu_{n,t}$, where $\nu_{n,t} : \mathcal{X} \to \mathbb{R}$ is a deterministic function for each $n \in \{1, 2, \dots\}$ and $t \in \mathcal{T}$. In the course of demonstrating that

$$\sup_{\|t-s\|<\delta/r_n} \left| R_n(t) - R_n(s) \right| \xrightarrow{\mathrm{P}} 0 \ ,$$

a rate of convergence for

$$\sup_{\|t-s\|<\delta/r_n} \left| \mathbb{G}_n \left(\nu_{n,t} - \nu_{n,s}\right) \right|$$

is then required. Defining $\mathcal{F}_{n,\eta} := \{\nu_{n,t} - \nu_{n,s} : \|t - s\| < \eta\}$ for each $\eta > 0$, this is equivalent to establishing a rate of convergence for the local empirical process $\|\mathbb{G}_n\|_{\mathcal{F}_{n,\delta/r_n}} := \sup_{\xi \in \mathcal{F}_{n,\delta/r_n}} |\mathbb{G}_n \xi|$. Such rates can be established using tail bounds for empirical processes. We briefly comment on two approaches to obtaining such tail bounds.

We first define bracketing and covering numbers of a class of functions $\mathcal{F}$ — see van der Vaart and Wellner (1996) for a comprehensive treatment. We denote by $\|F\|_{P,2} = [P(F^2)]^{1/2}$ the $L_2(P)$ norm of a given P-square-integrable function $F : \mathcal{X} \to \mathbb{R}$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_2(P))$ of a class of functions $\mathcal{F}$ with respect to the $L_2(P)$ norm is the smallest number of ε-brackets needed to cover $\mathcal{F}$, where an ε-bracket is any set of functions $\{f : \ell \leq f \leq u\}$ with ℓ and u such that $\|\ell - u\|_{P,2} < \varepsilon$. The covering number $N(\varepsilon, \mathcal{F}, L_2(Q))$ of $\mathcal{F}$ with respect to the $L_2(Q)$ norm is the smallest number of ε-balls in $L_2(Q)$ required to cover $\mathcal{F}$. The uniform covering number is the supremum of $N(\varepsilon\|F\|_{2,Q}, \mathcal{F}, L_2(Q))$ over all discrete probability measures Q such that $\|F\|_{Q,2} > 0$, where F is an envelope function for $\mathcal{F}$. The bracketing and uniform entropy integrals for $\mathcal{F}$ with respect to F are then defined as

$$\begin{aligned} J_{[]}(\delta, \mathcal{F}) &:= \int_0^\delta \left[1 + \log N_{[]}\left(\varepsilon\|F\|_{P_0,2}, \mathcal{F}, L_2(P_0)\right)\right]^{1/2} d\varepsilon \\ J(\delta, \mathcal{F}) &:= \sup_Q \int_0^\delta \left[1 + \log N\left(\varepsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right)\right]^{1/2} d\varepsilon \ . \end{aligned}$$

We discuss two approaches to controlling $\|\mathbb{G}_n\|_{\mathcal{F}_{n,\delta/r_n}}$ using these integrals. Suppose that $\mathcal{F}_{n,\eta}$ has envelope function $F_{n,\eta}$ in the sense that $|\xi(x)| \leq F_{n,\eta}$ for all $\xi \in \mathcal{F}_{n,\eta}$ and $x \in \mathcal{X}$. The first approach is useful when $\|F_{n,\delta/r_n}\|_{P_0,2}$ can be adequately controlled. Specifically, if either $J(1, \mathcal{F}_{n,\delta/r_n})$ or $J_{[]}(1, \mathcal{F}_{n,\delta/r_n})$ is $O(1)$, then $\|\mathbb{G}_n\|_{\mathcal{F}_{n,\delta/r_n}} \leq M_\delta \|F_{n,\delta/r_n}\|_{P_0,2}$ for all n and some constant $M_\delta \in (0, \infty)$ not depending on n by Theorems 2.14.1 and 2.14.2 of van der Vaart and Wellner (1996).

The second approach we consider is useful when the envelope functions do not shrink in expectation, but the functions in $\mathcal{F}_{n,\eta}$ still get smaller in the sense that $\gamma_{n,\delta} := \sup_{\xi \in \mathcal{F}_{n,\delta/r_n}} \|\xi\|_{P_0,2}$ tends to zero. For example, if $\nu_{n,t}$ is defined as $\nu_{n,t}(x) := I(0 \leq x \leq t)$ for each $x \in \mathcal{X} \subseteq \mathbb{R}$, $t \in [0,1]$, and n, then $F_{n,\eta} : x \mapsto I(0 \leq x \leq 1)$ is the natural envelope function for $\mathcal{F}_{n,\eta}$ for all n and η, so that $\|F_{n,\delta/r_n}\|_{P_0,2}$ does not tend to zero. However, if the density p_0 corresponding to P_0 is bounded above by $\bar{p}_0$, then $\gamma_{n,\delta}^2 \leq \bar{p}_0 \delta / r_n$, which does tend to zero. In these cases, the basic tail bounds in Theorem 2.14.1 and 2.14.2 of van der Vaart and Wellner (1996) are too weak. Sharper, but slightly more complicated, bounds may be used instead. Specifically, if $F_{n,\delta/r_n} \leq C < \infty$ for all n large enough and either

$$J\left(\gamma_{n,\delta}, \mathcal{F}_{n,\delta/r_n}\right) + \frac{J\left(\gamma_{n,\delta}, \mathcal{F}_{n,\delta/r_n}\right)^2}{\gamma_{n,\delta}^2 n^{1/2}} \ \text{ or } \ J_{[]}\left(\gamma_{n,\delta}, \mathcal{F}_{n,\delta/r_n}\right) + \frac{J_{[]}\left(\gamma_{n,\delta}, \mathcal{F}_{n,\delta/r_n}\right)^2}{\gamma_{n,\delta}^2 n^{1/2}}$$

are $o(z_n^{-1})$, then $\|\mathbb{G}_n\|_{\mathcal{F}_{n,\delta/r_n}} = o_P(z_n^{-1})$ by Lemma 3.4.2 of van der Vaart and Wellner (1996) and Theorem 2.1 of van der Vaart and Wellner (2011). Analogous statements hold if these expressions are $O(z_n^{-1})$.

In some cases, both of these approaches must be used to control different terms arising within $R_n(t)$, as for the conditional distribution function discussed in Section 4.2.

4. Illustrative examples

4.1. Example 1: Estimation of a G-computed distribution function

We first demonstrate the use of Theorem 3 in the particular problem in which we wish to draw inference on a G-computed distribution function. Suppose that the data unit is the vector $X = (Y, A, W)$, where Y is an outcome, $A \in \{0, 1\}$ is an exposure, and W is a vector of baseline covariates. The observed data consist of independent draws $X_1, X_2, \ldots, X_n$ from $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a nonparametric model.

For $P \in \mathcal{M}$ and $a_0 \in \{0, 1\}$, we define the parameter value θ_{P,a_0} pointwise as $\theta_{P,a_0}(t) := E_P\{P(Y \leq t \mid A = a_0, W)\}$, the G-computed distribution function of Y evaluated at t, where the outer expectation is over the marginal distribution of W under P. We are interested in estimating $\theta_{0,a_0} := \theta_{P_0,a_0}$. This parameter is often of interest as an interpretable marginal summary of the relationship between Y and A accounting for the potential confounding induced by W. Under certain causal identification conditions, θ_{0,a_0} is the distribution function of the counterfactual outcome $Y(a_0)$ defined by the intervention that deterministically sets exposure to $A = a_0$ (Robins, 1986; Gill and Robins, 2001).

For each t, the parameter $P \mapsto \theta_{P,a_0}(t)$ is pathwise differentiable in a nonparametric model, and its nonparametric efficient influence function $\varphi_{P,a_0,t}$ at $P \in \mathcal{M}$ is given by

$$(y, a, w) \mapsto \frac{I(a = a_0)}{g_P(a_0 \mid w)}\left[I(y \leq t) - \bar{Q}_P(t \mid a_0, w)\right] + \bar{Q}_P(t \mid a_0, w) - \theta_{P,a_0}(t)\ ,$$

where $g_P(a_0 \mid w) := P(A = a_0 \mid W = w)$ is the propensity score and $\bar{Q}_P(t \mid a_0, w) := P(Y \leq t \mid A = a_0, W = w)$ is the conditional exposure-specific distribution function, as implied by P (van der Laan and Robins, 2003). Given estimators g_n and $\bar{Q}_n$ of $g_0 := g_{P_0}$ and $\bar{Q}_0 := \bar{Q}_{P_0}$, respectively, several approaches can be used to construct, for each t, an asymptotically linear estimator of $\theta_0(t)$ with influence function $\phi_{0,a_0,t} = \varphi_{P_0,a_0,t}$. For example, the use of either optimal estimating equations or the one-step correction procedure leads to the doubly-robust augmented inverse-probability-of-weighting estimator

$$\theta_{n,a_0}(t) := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i = a_0)}{g_n(a_0 \mid W_i)}\left[I(Y_i \leq t) - \bar{Q}_n(t \mid a_0, W_i)\right] + \bar{Q}_n(t \mid a_0, W_i)\right\},$$

as discussed in detail in van der Laan and Robins (2003). Under conditions on g_n and $\bar{Q}_n$, including consistency at fast enough rates, $\theta_{n,a_0}(t)$ is asymptotically efficient relative to $\mathcal{M}$. In this case, $\theta_{n,a_0}(t)$ satisfies (3.1) with influence function $\phi_{0,a_0,t}$. However, there is no guarantee that θ_{n,a_0} is monotone.

In the context of this example, we can identify simple sufficient conditions under which conditions (A)–(B), and hence the asymptotic equivalence of the initial and isotonized estimators of the G-computed distribution function, are guaranteed. Specifically, we find this to be the case when both:

(i) there exists $\eta > 0$ such that $g_0(a_0 \mid W) \geq \eta$ almost surely under P_0;

(ii) there exist non-negative real-valued functions K_1, K_2 such that

$$K_1(w)|t-s| \leq |\bar{Q}_0(t \mid a_0, w) - \bar{Q}_0(s \mid a_0, w)| \leq K_2(w)|t-s|$$

for all $t, s \in \mathcal{T}$, and such that, under P_0, $K_1(W)$ is strictly positive with non-zero probability and $K_2(W)$ has finite second moment.

We conducted a simulation study to validate our theoretical results in the context of this particular example. For samples sizes $100, 250, 500, 750$, and 1000, we generated 1000 random datasets as follows. We first simulated a bivariate covariate W with independent components W_1 and W_2, respectively distributed as a Bernoulli variate with success probability 0.5 and a uniform variate on $(-1, 1)$. Given $W = (w_1, w_2)$, exposure A was simulated from a logistic regression model with

$$P_0(A = 1 \mid W_1 = w_1, W_2 = w_2) = \text{expit}(0.5 + w_1 - 2w_2) \ .$$

Given $W = (w_1, w_2)$ and $A = a$, Y was simulated as the inverse-logistic transformation of a normal variate with mean $0.2 - 0.3a - 4w_2$ and variance 0.3.

For each simulated dataset, we estimated $\theta_{0,0}(t)$ and $\theta_{0,1}(t)$ for t equal to each outcome value observed between 0.1 and 0.9. To do so, we used the estimator described above, with propensity score and conditional exposure-specific distribution function estimated using correctly-specified parametric models. We employed two correction procedures for the estimators $\theta_{n,0}$ and $\theta_{n,1}$. First, we projected $\theta_{n,0}$ and $\theta_{n,1}$ onto the space of monotone functions separately. Second, noting that $\theta_{0,0}(t) \leq \theta_{0,1}(t)$ for all t, so that $(a, t) \mapsto \theta_{0,a}(t)$ is component-wise monotone for this particular data-generating distribution, we considered the projection of $(a, t) \mapsto \theta_{n,a}(t)$ onto the space of bivariate monotone functions on $\{0, 1\} \times \mathcal{T}$. For each simulation and each projection procedure, we recorded the maximal absolute differences between (i) the initial and and projected estimates, (ii) the initial estimate and the truth, and (iii) the projected estimate and the truth. We also recorded the maximal widths of the initial and projected confidence bands.

Figure 1 displays the results of this simulation study, with output from the univariate and bivariate projection approaches summarized in the top and bottom rows, respectively. The left column displays the empirical distribution of the scaled maximum absolute discrepancy between θ_n and θ_n^* for all sample sizes studied. This plot confirms that the discrepancy between these two estimators indeed decreases faster than $n^{-1/2}$, as our theory suggests. Furthermore, for each n, the discrepancy is larger for the two-dimensional projection.

The middle column of Figure 1 displays the empirical distribution function of the ratio between the maximum discrepancy between θ_n and θ_0 and that of θ_n^* and θ_0. This plot confirms that θ_n^* is always at least as close to θ_0 than is θ_n over $\mathcal{T}_n$. The maximum discrepancy between θ_n and θ_0 can be more than 25% larger than that between θ_n^* and θ_0 in the univariate case, and up to 50% larger in the bivariate case.

The right column of Figure 1 displays the empirical distribution function of the ratio between the maximum size of the initial uniform 95% influence

TABLE 1
Coverage of 95% confidence bands for the true counterfactual distribution function.

	n	100	250	500	750	1000
d=1	Initial band	92.5	94.1	96.0	94.5	95.5
	Monotone band	92.5	94.1	96.0	94.5	95.5
d=2	Initial band	93.9	94.0	95.0	94.6	94.9
	Monotone band	95.7	95.9	95.5	95.3	95.1

function-based confidence band and that of the isotonic band. For large samples, the maximal widths are often close, but for smaller samples, the initial confidence bands can be up to 50% larger than the isotonic bands, especially for the bivariate case. The empirical coverage of both bands is provided in Table 1. The coverage of the isotonic band is essentially the same as the initial band for the univariate case, whereas it is slightly larger than that of the initial band in the bivariate case.

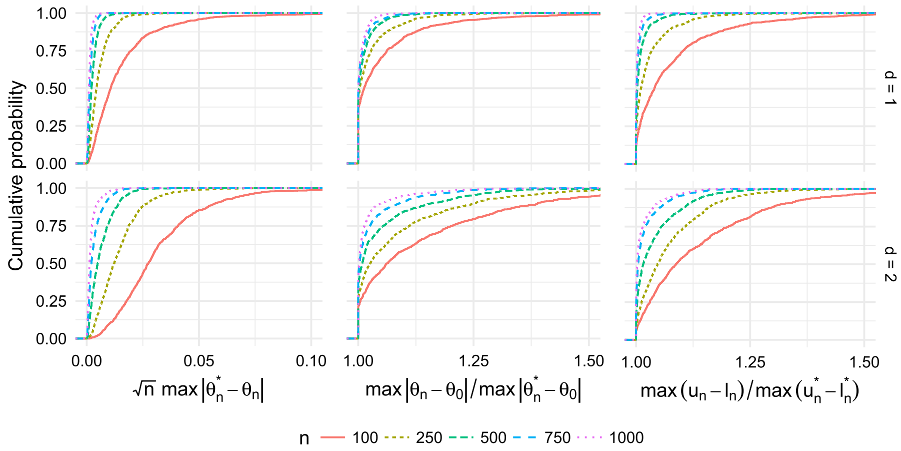

FIG 1. *Summary of simulation results for G-computed distribution function. Each plot shows cumulative distributions of a particular discrepancy over 1000 simulated datasets for different values of n. Left panel: maximal absolute difference between the initial and isotonic estimators over the grid used for projecting, scaled up by root-n. Middle panel: ratio of the maximal absolute difference between the initial estimator and the truth and the maximal absolute difference between the isotonic estimator and the truth. Right panel: ratio of the maximal width of the initial confidence band and the maximal width of the isotonic confidence band. The top row shows the results for the univariate projection, and the bottom row shows the results for the bivariate projection.*

4.2. Example 2: Estimation of a conditional distribution function

We next demonstrate the use of Theorem 2 with dimension $d = 2$ for drawing inference on a conditional distribution function. Suppose that the data unit is the vector $X = (A, Y)$, where Y is an outcome and A is now a continuous exposure. The observed data consist of independent draws $(A_1, Y_1), (A_2, Y_2), \ldots, (A_n, Y_n)$ from $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a nonparametric model. We define the parameter value θ_P pointwise as $\theta_P(t_1, t_2) := P(Y \leq t_1 \mid A = t_2)$. Thus, θ_P is the conditional distribution function of Y at t_1 given $A = t_2$. The map $(t_1, t_2) \mapsto \theta_P(t_1, t_2)$ is necessarily monotone in t_1 for each fixed t_2, and in some settings, it may be known that it is also monotone in t_2 for each fixed t_1. This parameter completely describes the conditional distribution of Y given A, and can be used to obtain the conditional mean, conditional quantiles, or any other conditional parameter of interest.

For each t_1, the true function $\theta_0(t_1, t_2) = \theta_{P_0}(t_1, t_2)$ may be written as the conditional mean of $I(Y \leq t_1)$ given $A = t_2$. Hence, any method of nonparametric regression can be used to estimate $t_2 \mapsto \theta_0(t_1, t_2)$ for fixed t_1, and repeating such a method over a grid of values of t_1 yields an estimator of the entire function. We expect that our results would apply to many of these methods. Here, we consider the local linear estimator (Fan and Gijbels, 1996), which may be expressed as

$$\theta_n(t_1, t_2) := \frac{1}{nh_n} \sum_{i=1}^{n} I(Y_i \leq t_1) \left[\frac{s_{2,n}(t_2) - s_{1,n}(t_2)\,(A_i - t_2)}{s_{0,n}(t_2) s_{2,n}(t_2) - s_{1,n}(t_2)^2} \right] K\left(\frac{A_i - t_2}{h_n} \right) ,$$

where $K : \mathbb{R} \to \mathbb{R}$ is a symmetric and bounded kernel function, $h_n \to 0$ is a sequence of bandwidths, and

$$s_{j,n}(t_2) := \frac{1}{nh_n} \sum_{i=1}^{n} (A_i - t_2)^j \, K\left(\frac{A_i - t_2}{h_n} \right)$$

for $j \in \{0, 1, 2\}$. Under regularity conditions on the true distribution function θ_0, the marginal density f_0 of A, the bandwidth sequence h_n, and the kernel function K, for any fixed (t_1, t_2), θ_n satisfies

$$(nh_n)^{1/2} \left[\theta_n(t_1, t_2) - \theta_0(t_1, t_2) - h_n^2 V_K b_0(t_1, t_2) \right] \xrightarrow{\mathrm{d}} N\left(0, S_K v_0(t_1, t_2)\right),$$

where $V_K := \int x^2 K(x) dx$ is the variance of K, $S_K := \int K(x)^2 dx$, and $b_0(t_1, t_2)$ and $v_0(t_1, t_2)$ depend on the derivatives of θ_0 and on f_0. If h_n is chosen to be of order $n^{-1/5}$, the rate that minimizes the asymptotic mean integrated squared error of θ_n relative to θ_0, then $n^{2/5} \left[\theta_n(t_1, t_2) - \theta_0(t_1, t_2)\right]$ converges in law to a normal random variate with mean $V_K b_0(t_1, t_2)$ and variance $S_K v_0(t_1, t_2)$. Under stronger regularity conditions, the rate of convergence of the uniform norm $\|\theta_n - \theta_0\|_{\mathcal{T}}$ can be shown to be $(nh_n / \log n)^{1/2}$ (Hardle, Janssen and Serfling, 1988).

Theorem 3 cannot be used to establish (A) in this problem, since θ_n is not an asymptotically linear estimator. Furthermore, as discussed above, recent results suggest that $\{r_n[\theta_n(t) - \theta_0(t)] : t \in \mathcal{T}\}$ does not converge weakly to a tight limit in $\ell^\infty(\mathcal{T})$ for any useful rate r_n. Despite this lack of weak convergence, condition (A) can be verified directly in the context of this example under smoothness conditions on θ_0 and f_0 using the tail bounds for empirical processes outlined in Section 3.2. Denoting by θ'_{0,t_2} and θ''_{0,t_2} the first and second derivatives of θ_0 with respect to its second argument, we define

$$R_\theta^{(2)}(t,\delta) := \theta_0(t_1, t_2+\delta) - \theta_0(t_1,t_2) - \delta\theta'_{0,t_2}(t_1,t_2) - \tfrac{1}{2}\delta^2\theta''_{0,t_2}(t_1,t_2)$$

and $R_f^{(1)}(t,\delta) := f_0(t_2+\delta) - f_0(t_2) - \delta f_0'(t_2)$, where f_0' is the derivative of f_0. We then introduce the following conditions on θ_0, f_0, and K:

(d) θ''_{0,t_2} exists and is continuous on $\mathcal{T}$, and as $\delta \to 0$, $\sup_{t\in\mathcal{T}} |R_\theta^{(2)}(t,\delta)| = o(\delta^2)$;

(e) $\inf_{t\in\mathcal{T}} f_0(t) > 0$, f_0' exists and is continuous on $\mathcal{T}$, and $\sup_{t\in\mathcal{T}} |R_f^{(1)}(t,\delta)| = o(\delta)$;

(f) K is a Lipschitz function supported on $[-1,1]$ satisfying condition (M) of Stupfler (2016).

We also define

$$\begin{aligned} \nu_{n,t}(y,a) &:= [I(y \le t_1) - \theta_0(t_1,a)]\, K\left(\frac{a-t_2}{h_n}\right) \ ; \\ g_n(t_2) &:= s_{0,n}(t_2)s_{2,n}(t_2) - s_{1,n}(t_2)^2 \ ; \\ R_n(t) &:= h_n^{-1/2}\left[\frac{s_{2,n}(t_2)}{g_n(t_2)}\mathbb{G}_n\nu_{n,t} - \frac{s_{1,n}(t_2)}{g_n(t_2)}\mathbb{G}_n\left(\ell_t\nu_{n,t}\right)\right] \ . \end{aligned}$$

We then have the following result.

Proposition 1. *If (d)–(f) hold, $nh_n^4/\log h_n^{-1} \to \infty$ and $nh_n^5 = O(1)$, then*

$$\sup_{t\in\mathcal{T}}\left|(nh_n)^{1/2}[\theta_n(t_1,t_2) - \theta_0(t_1,t_2)] - \left(nh_n^5\right)^{1/2}\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)K_2 - R_n(t)\right| \xrightarrow{\mathrm{P}} 0 \ .$$

Proposition 1 aids in establishing the following result, which formally establishes asymptotic equivalence of the local linear estimator of a conditional distribution function and its correction obtained via isotonic regression at the rate $r_n = (nh_n)^{1/2}$.

Proposition 2. *If (d)–(f) hold and $nh_n^5 \to c \in (0,\infty)$, then (A) holds for the local linear estimator with $r_n = (nh_n)^{1/2}$.*

The proofs of Propositions 1 and 2 are provided in Appendix A.5. These results may also be of interest in their own right for establishing other properties of the local linear estimator.

As with the first example, we conducted a simulation study to validate our theoretical results. For samples sizes $n \in \{100, 250, 500, 750, 1000\}$, we generated 1000 random datasets as follows. We first simulated A as a Beta$(2,3)$ variate. Given $A = a$, Y was simulated as the inverse-logistic transformation of a normal variate with mean $0.5 \times [1 + (a-1.2)^2]$ and variance one.

For each simulated dataset, we estimated $\theta_0(y,a)$ for each (y,a) in an equally spaced square grid of mesh $\omega_n = n^{-4/5}$. For each unique y in this grid, we estimated the function $a \mapsto \theta_0(y, a)$ using the local linear estimator, as implemented in the `R` package `KernSmooth` (Wand, 2015; Wand and Jones, 1995). For each value of y in the grid, we computed the optimal bandwidth based on the direct plug-in methodology of Ruppert, Sheather and Wand (1995) as implemented by the `dpill` function, and we then set our bandwidth as the average of these y-specific bandwidths. We constructed initial confidence bands using a variable-width nonparametric bootstrap (Hall and Kang, 2001).

We first note that, for all sample sizes considered, over 99% of simulations had monotonicity violations in both the y- and a-directions. Figure 2 displays the results of this simulation study. The left exhibit of Figure 2 confirms that the discrepancy between θ_n and θ_n^* decreases faster than $r_n^{-1} = n^{-2/5}$, as our theory suggests. The middle exhibit indicates that in roughly 50% of simulations, there is less than 5% difference between $\|\theta_n^* - \theta_0\|_{\mathcal{T}_n}$ and $\|\theta_n - \theta_0\|_{\mathcal{T}_n}$, but even for $n = 1000$, in roughly 25% of simulations, θ_n^* offers at least a 25% improvement in estimation error. In smaller samples, the estimation error of θ_n^* is less than half that of θ_n in 5–10% of simulations. The rightmost exhibit indicates that the projected confidence bands regularly reduce the uniform size of the initial bands by 10–20%. Finally, the empirical coverage of uniform 95% bootstrap-based bands and their projected versions is provided in Table 2. As before, the projected band is always more conservative than the initial band, and the difference in coverage diminishes as n grows. However, the initial bands in this example are anti-conservative, even at $n = 1000$, likely due to the slower rate of convergence, and the corrected bands offer a much more substantial improvement in this example than in the first.

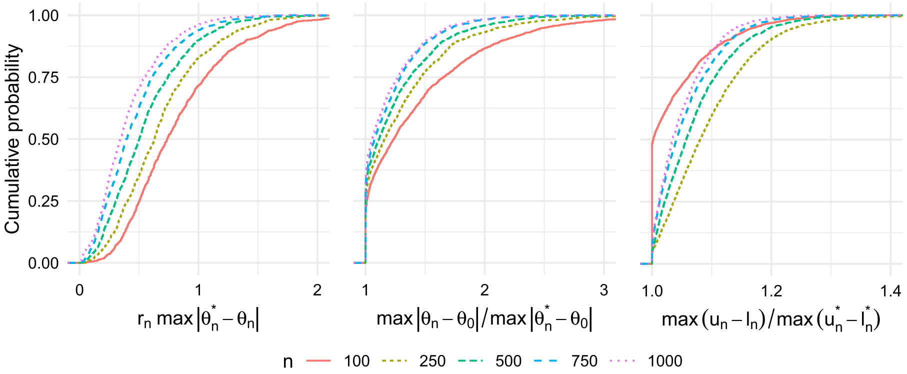

FIG 2. *Summary of simulation results for conditional distribution function. The three columns display the same results as those in Figure 1.*

5. Discussion

Many estimators of function-valued parameters in nonparametric and semiparametric models are not guaranteed to respect shape constraints on the true func-

Table 2
Coverage of 95% confidence bands for the true conditional distribution function.

n	100	250	500	750	1000
Initial band	37.6	64.9	83.2	86.3	89.7
Monotone band	60.8	80.4	90.3	92.3	93.9

tion. A simple and general solution to this problem is to project the initial estimator onto the constrained parameter space over a grid whose mesh goes to zero fast enough with sample size. However, this introduces the possibility that the projected estimator has different properties than the original estimator. In this paper, we studied the important shape constraint of multivariate componentwise monotonicity. We provided results indicating that the projected estimator is generically no worse than the initial estimator, and that if the true function is strictly increasing and the initial estimator possesses a relatively weak type of stochastic equicontinuity, the projected estimator is uniformly asymptotically equivalent to the initial estimator. We provided especially simple sufficient conditions for this latter result when the initial estimator is uniformly asymptotically linear, and provided guidance on establishing the key condition for kernel smoothed estimators.

We studied the application of our results in two examples: estimation of a G-computed distribution function, for use in understanding the effect of a binary exposure on an outcome when the exposure-outcome relationship is confounded by recorded covariates, and of a conditional distribution function, for use in characterizing the marginal dependence of an outcome on a continuous exposure. In numerical studies, we found that the projected estimator yielded improvements over the initial estimator. The improvements were especially strong in the latter example.

In our examples, we only studied corrections in dimensions $d = 1$ and $d = 2$. In future work, it would be interesting to consider corrections in dimensions higher than 2. For example, for the conditional distribution function, it would be of interest to study multivariate local linear estimators for a continuous exposure A taking values in $\mathbb{R}^{d-1}$ for $d > 2$. Since tailored algorithms for computing the isotonic regression do not yet exist for $d > 2$, it would also be of interest to determine whether a version of Theorem 2 could be established for the relaxed isotonic estimator proposed by Fokianos, Leucht and Neumann (2017). Alternatively, it is possible that the uniform stochastic equicontinuity currently required by Chernozhukov, Fernández-Val and Galichon (2010) and Daouia and Park (2013) for asymptotic equivalence of the rearrangement- and envelope-based corrections, respectively, could be relaxed along the lines of our condition (A). Finally, our theoretical results do not give the exact asymptotic behavior of the projected estimator or projected confidence band when the true function possesses flat regions. This is also an interesting topic for future research.

Appendix A: Technical proofs

A.1. Proof of Theorem 1

Part (i) follows from Corollary B to Theorem 1.6.1 of Robertson, Wright and Dykstra (1988). For parts (ii) and (iii), we note that by assumption

$$|\theta_n^*(t) - \theta_0(t)| \;\leq\; \sum_k \lambda_{k,n}(t)|\theta_n^*(s_k) - \theta_0(s_k)| + \sum_k \lambda_{k,n}(t)|\theta_0(s_k) - \theta_0(t)|$$

for every $t \in \mathcal{T}$, where $\sum_k \lambda_{k,n}(t) = 1$, and for each k, $s_k \in \mathcal{T}_n$ and $\|s_k - t\| \leq 2\omega_n$. By part (i), the first term is bounded above by $\sup_{s\in\mathcal{T}_n} |\theta_n(s) - \theta_0(s)|$. The second term is bounded above by $\gamma(2\omega_n)$, where we define

$$\gamma(\delta) := \sup\{|\theta_0(t) - \theta_0(s)| : t, s \in \mathcal{T}, \|t - s\| \leq \delta\} \ .$$

If θ_0 is continuous on $\mathcal{T}$, then it is also uniformly continuous since $\mathcal{T}$ is compact. Therefore, $\gamma(\delta) \to \gamma(0) = 0$ as $\delta \to 0$, so that $\gamma(2\omega_n) \to_{\mathrm{P}} 0$ if $\omega_n \to_{\mathrm{P}} 0$. If $\gamma(\delta) = o(\delta^\alpha)$ as $\delta \to 0$, then $\gamma(2\omega_n) = o_P(\omega_n^\alpha)$.

Part (iv) follows from the proof of Proposition 3 of Chernozhukov, Fernández-Val and Galichon (2009), which applies to any order-preserving monotonization procedure. For the first statement of (v), by their definition as minimizers of the least-squares criterion function, we note that $\sum_{t\in\mathcal{T}_n} u_n^*(t) = \sum_{t\in\mathcal{T}_n} u_n(t)$, and similarly for ℓ_n^*. The second statement of (v) follows from a slight modification of Theorem 1.6.1 of Robertson, Wright and Dykstra (1988). As stated, the result says that $\sum_{t\in\mathcal{T}_n} G(\theta^*(t) - \theta(t)) \leq \sum_{t\in\mathcal{T}_n} G(\theta(t) - \psi(t))$ for any convex function $G : \mathbb{R} \to \mathbb{R}$ and monotone function ψ, where θ^* is the isotonic regression of θ over $\mathcal{T}_n$. A straightforward adaptation of the proof indicates that $\sum_{t\in\mathcal{T}_n} G(\theta_1^*(t) - \theta_2^*(t)) \leq \sum_{t\in\mathcal{T}_n} G(\theta_1(t) - \theta_2(t))$, where now θ_1^* and θ_2^* are the isotonic regressions of θ_1 and θ_2 over $\mathcal{T}_n$, respectively. As in Corollary B, taking $G(x) = |x|^p$ and letting $p \to \infty$ yields that $\|\theta_1^* - \theta_2^*\|_{\mathcal{T}_n} \leq \|\theta_1 - \theta_2\|_{\mathcal{T}_n}$. Applying this with $\theta_1 = u_n$ and $\theta_2 = \ell_n$ establishes the second portion of (v). □

A.2. Proof of Theorem 2

We prove Theorem 2 via three lemmas, which may be of interest in their own right. The first lemma controls the size of deviations in θ_n over small neighborhoods, and does not hinge on condition (C) holding.

Lemma 1. *If (A)–(B) hold and $b_n = o_P(r_n^{-1})$, then*

$$\sup_{\|t-s\|\leq b_n} |\theta_n(t) - \theta_n(s)| = o_P(r_n^{-1}) \ .$$

Proof of Lemma 1. In view of the triangle inequality,

$$|\theta_n(t) - \theta_n(s)| \leq |\{\theta_n(t) - \theta_0(t)\} - \{\theta_n(s) - \theta_0(s)\}| + |\theta_0(t) - \theta_0(s)| \ .$$

The first term is $o_P(r_n^{-1})$ by (A), whereas the second term is $o_P(r_n^{-1})$ by (B). □

The second lemma controls the size of neighborhoods over which violations in monotonicity can occur. Henceforth, we define

$$\kappa_n := \sup\left\{\|t-s\| : s,t\in\mathcal{T}, s\leq t, \theta_n(t)\leq\theta_n(s)\right\} \ .$$

In this lemma we again require (A) but now require (C) rather than (B).

Lemma 2. *If (A) and (C) hold, then $\kappa_n = o_P(r_n^{-1})$.*

Proof of Lemma 2. Let $\epsilon > 0$ and $\eta_n := \epsilon/r_n$. Suppose that $\kappa_n > \eta_n$. Then, there exist $s,t\in\mathcal{T}$ with $s<t$ and $\|t-s\| > \eta_n$ such that $\theta_n(s)\geq\theta_n(t)$. We claim that there must also exist $s^*, t^*\in\mathcal{T}$ with $s^* < t^*$ and $\|t^*-s^*\|\in[\eta_n/2,\eta_n]$ such that $\theta_n(s^*)\geq\theta_n(t^*)$. To see this, let $J = \lfloor\|t-s\|/(\eta_n/2)\rfloor - 1$, and note that $J\geq 1$. Define $t_j := s+(j\eta_n/2)(t-s)/\|t-s\|$ for $j=0,1,\dots,J$, and set $t_{J+1} := t$. Thus, $t_j < t_{j+1}$ and $\|t_{j+1}-t_j\|\in[\eta_n/2,\eta_n]$ for each $j=0,1,\dots,J$. Since then $\sum_{j=0}^{J}[\theta_n(t_{j+1})-\theta_n(t_j)] = \theta_n(t)-\theta_n(s)\leq 0$, it must be that $\theta_n(t_{j+1})\leq\theta_n(t_j)$ for at least one j. This proves the claim.

We now have that $\kappa_n > \eta_n$ implies that there exist $s,t\in\mathcal{T}$ with $s<t$ and $\|t-s\|\in[\eta_n/2,\eta_n]$ such that $\theta_n(s)\geq\theta_n(t)$. This further implies that

$$\{\theta_n(t)-\theta_0(t)\}-\{\theta_n(s)-\theta_0(s)\}\leq -\{\theta_0(t)-\theta_0(s)\}\leq -K_0\|t-s\|\leq -K_0\eta_n/2$$

by condition (B). Finally, this allows us to write

$$P_0\left(\kappa_n > \frac{\epsilon}{r_n}\right)\leq P_0\left\{\sup_{\|t-s\|\leq\epsilon/r_n} r_n\left|[\theta_n(t)-\theta_0(t)]-[\theta_n(s)-\theta_0(s)]\right|\geq\frac{K_0\epsilon}{2}\right\}.$$

By condition (A), this probability tends to zero for every $\epsilon > 0$, which completes the proof. □

Our final lemma bounds the maximal absolute deviation between θ_n^* and θ_n over the grid $\mathcal{T}_n$ in terms of the supremal deviations of θ_n over neighborhoods smaller than κ_n. This lemma does not depend on any of the conditions (A)–(C).

Lemma 3. *It holds that* $\max_{t\in\mathcal{T}_n}|\theta_n^*(t)-\theta_n(t)|\leq\sup_{\|s-t\|\leq\kappa_n}|\theta_n(s)-\theta_n(t)|$.

Proof of Lemma 3. By Theorem 1.4.4 of Robertson, Wright and Dykstra (1988), for any $t\in\mathcal{T}_n$,

$$\theta_n^*(t) = \max_{U\in\mathcal{U}_t}\min_{L\in\mathcal{L}_t}\theta_n(U\cap L) = \min_{L\in\mathcal{L}_t}\max_{U\in\mathcal{U}_t}\theta_n(U\cap L),$$

where, for any finite set $S\subseteq\mathcal{T}_n$, $\theta_n(S)$ is defined as $|S|^{-1}\sum_{s\in S}\theta_n(s)$. The sets U range over the collection $\mathcal{U}_t$ of upper sets of $\mathcal{T}_n$ containing t, where $U\subseteq\mathcal{T}_n$ is called an upper set if $t_1\in U, t_2\in\mathcal{T}_n$ and $t_1\leq t_2$ implies $t_2\in U$. The sets L range over the collection $\mathcal{L}_t$ of lower sets of $\mathcal{T}_n$ containing t, where $L\subseteq\mathcal{T}_n$ is called a lower set if $t_1\in L, t_2\in\mathcal{T}_n$ and $t_2\leq t_1$ implies $t_2\in L$.

Let $U_t := \{s : s\geq t\}$ and $L_t := \{s : s\leq t\}$. First, suppose there exists $L_0\in\mathcal{L}_t$ and $s_0\in L_0$ with $s_0 > t$ and $\|t-s_0\| > \kappa_n$. Then, we claim that there exists another lower set $L_0'\in\mathcal{L}_t$ such that $\theta_n(U_t\cap L_0) > \theta_n(U_t\cap L_0')$. If

$\theta_n(U_t \cap L_0) > \theta_n(t) = \theta_n(U_t \cap L_t)$, then $L_0' = L_t$ satisfies the claim. Otherwise, if $\theta_n(U_t \cap L_0) \leq \theta_n(t)$, let

$$L_0' := L_0 \setminus \{s : s > t, \|t - s\| > \kappa_n\} \ .$$

One can verify that $L_0' \in \mathcal{L}_t$, and since $s_0 \in L_0 \setminus L_0'$, L_0' is a strict subset of L_0. Furthermore, by definition of κ_n, $\theta_n(s) > \theta_n(t)$ for all $s > t$ such that $\|t - s\| > \kappa_n$, and since $\theta_n(U_t \cap L_0) \leq \theta_n(t)$, removing these elements from L_0 can only reduce the average, so that $\theta_n(U_t \cap L_0') < \theta_n(U_t \cap L_0)$. This establishes the claim. By an analogous argument, we can show that if there exists $U_0 \in \mathfrak{U}_t$ and $s_0 \in U_0$ with $s_0 < t$ and $\|t - s_0\| > \kappa_n$, then there exists another upper set $U_0' \in \mathfrak{U}_t$ such that $\theta_n(U_0 \cap L_t) < \theta_n(U_0' \cap L_t)$.

Let $L^* \in \operatorname{argmin}_{L \in \mathcal{L}_t} \theta_n(U_t \cap L)$ and $U^* \in \operatorname{argmax}_{U \in \mathfrak{U}_t} \theta_n(U \cap L_t)$. Then,

$$\theta_n^*(t) = \max_{U \in \mathfrak{U}_t} \min_{L \in \mathcal{L}_t} \theta_n(U \cap L) \geq \min_{L \in \mathcal{L}_t} \theta_n(U_t \cap L) = \theta_n(U_t \cap L^*) \quad \text{and}$$
$$\theta_n^*(t) = \min_{L \in \mathcal{L}_t} \max_{U \in \mathfrak{U}_t} \theta_n(U \cap L) \leq \max_{U \in \mathfrak{U}_t} \theta_n(U \cap L_t) = \theta_n(U^* \cap L_t) \ .$$

Hence, $\theta_n(U_t \cap L^*) \leq \theta_n^*(t) \leq \theta_n(U^* \cap L_t)$. By the above argument, we have that both

$$\theta_n(U_t \cap L^*) \geq \inf\{\theta_n(s) : s \geq t, \|t - s\| \leq \kappa_n\} \quad \text{and}$$
$$\theta_n(U^* \cap L_t) \leq \sup\{\theta_n(s) : s \leq t, \|t - s\| \leq \kappa_n\} \ .$$

Therefore, we find that

$$\begin{aligned} \inf\{\theta_n(s) - \theta_n(t) : \|t - s\| \leq \kappa_n\} \ &\leq \ \theta_n^*(t) - \theta_n(t) \\ &\leq \ \sup\{\theta_n(s) - \theta_n(t) : \|t - s\| \leq \kappa_n\} \ , \end{aligned}$$

and thus, $|\theta_n^*(t) - \theta_n(t)| \leq \sup\{|\theta_n(s) - \theta_n(t)| : \|t - s\| \leq \kappa_n\}$. Taking the maximum over $t \in \mathcal{T}_n$ yields the claim. □

The proof of Theorem 2 follows easily from Lemmas 1, 2 and 3.

Proof of Theorem 2. By construction, for each $t \in \mathcal{T}$, we can write

$$|\theta_n^*(t) - \theta_n(t)| \leq \Sigma_{j=1}^{2^d} \lambda_{j,n}(t) |\theta_n^*(s_j) - \theta_n(s_j)| + \Sigma_{j=1}^{2^d} \lambda_{j,n}(t) |\theta_n(s_j) - \theta_n(t)| \ ,$$

where $s_j \in \mathcal{T}_n$ and $\|s_j - t\| \leq 2\omega_n$ for all t, s_j by definition. Thus, since $\sum_j \lambda_{j,n}(t) = 1$, it follows that

$$\sup_{t \in \mathcal{T}} |\theta_n^*(t) - \theta_n(t)| \ \leq \ \max_{t \in \mathcal{T}_n} |\theta_n^*(t) - \theta_n(t)| + \sup_{\|s - t\| \leq 2\omega_n} |\theta_n(s) - \theta_n(t)| \ .$$

By Lemma 3, the first summand is bounded above by $\sup_{\|s-t\| \leq \kappa_n} |\theta_n(s) - \theta_n(t)|$, which is $o_P(r_n^{-1})$ by Lemmas 1 and 2. The second summand is $o_P(r_n^{-1})$ by Lemma 1. □

A.3. Proof of Corollary 1

We note that $\ell_n(t) \leq \theta_0(t) \leq u_n(t)$ if and only if

$$\begin{aligned}\{r_n[\theta_n(t) - \ell_n(t)] - \gamma_\alpha(t)\} + \gamma_\alpha(t) &\geq r_n[\theta_n(t) - \theta_0(t)] \\ &\geq -\gamma_\alpha(t) - \{r_n[u_n(t) - \theta_n(t)] - \gamma_\alpha(t)\} \ .\end{aligned}$$

Therefore, by conditions (a)–(c), $P_0\left[\ell_n(t) \leq \theta_0(t) \leq u_n(t) \text{ for all } t \in \mathcal{T}\right] \to 1-\alpha$. Next, we let $\delta > 0$ and note that

$$\begin{aligned}&\sup_{\|t-s\|\leq\delta/r_n} |r_n\{\ell_n(t) - \theta_0(t)\} - r_n\{\ell_n(s) - \theta_0(s)\}| \\ &\quad\leq \sup_{\|t-s\|\leq\delta/r_n} |r_n\{\theta_n(t) - \theta_0(t)\} - r_n\{\theta_n(s) - \theta_0(s)\}| \\ &\qquad\quad + \sup_{\|t-s\|\leq\delta/r_n} |\gamma_\alpha(t) - \gamma_\alpha(s)| + 2\|r_n(\theta_n - \ell_n) - \gamma_\alpha\|_{\mathcal{T}} \ .\end{aligned}$$

The first term tends to zero in probability by (A), the second by conditions (a)–(c), and the third by the assumed uniform continuity of γ_α. An analogous decomposition holds for u_n. Therefore, we can apply Theorem 2 with u_n and ℓ_n in place of θ_n to find that $\|\ell_n^* - \ell_n\|_{\mathcal{T}} = o_P(r_n^{-1})$ and $\|u_n^* - u_n\|_{\mathcal{T}} = o_P(r_n^{-1})$. Finally, applying an analogous argument to the event $\ell_n^* \leq \theta_0 \leq u_n^*$ as we applied to $\ell_n \leq \theta_0 \leq u_n$ above yields the result. □

A.4. Proof of Theorem 3

Let $\epsilon, \delta, \eta > 0$. By (3.1) and since $\sup_{t\in\mathcal{T}} |R_{n,t}| = o_P(n^{-1/2})$,

$$n^{1/2}\left|\{\theta_n(t) - \theta_0(t)\} - \{\theta_n(s) - \theta_0(s)\}\right| \leq |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| + o_P(1) \ .$$

Condition (A2) implies that $\{\phi_{0,t} : t \in \mathcal{T}\}$ is uniformly mean-square continuous, in the sense that

$$\lim_{h\to 0} \sup_{\|t-s\|\leq h} \int \{\phi_{0,s}(x) - \phi_{0,t}(x)\}^2 \, dP_0(x) = 0 \ .$$

Since $\mathcal{T}$ is totally bounded in $\|\cdot\|$, this also implies that $\{\phi_{0,t} : t \in \mathcal{T}\}$ is totally bounded in the $L_2(P_0)$ metric. This, in addition to (A1), implies that $\{\mathbb{G}_n\phi_{0,t} : t \in \mathcal{T}\}$ converges weakly in $\ell^\infty(\mathcal{T})$ to a Gaussian process $\mathbb{G}$ with covariance function Σ_0. Furthermore, (A2) implies that this limit process is a tight element of $\ell^\infty(\mathcal{T})$. By Theorem 1.5.4 of van der Vaart and Wellner (1996), $\{\mathbb{G}_n\phi_{0,t} : t \in \mathcal{T}\}$ is asymptotically tight. By Theorem 1.5.7 of van der Vaart and Wellner (1996), $\{\mathbb{G}_n\phi_{0,t} : t \in \mathcal{T}\}$ is thus asymptotically uniformly mean-square equicontinuous in probability, in the sense that there exists some $\delta_0 = \delta_0(\epsilon, \eta) > 0$ such that

$$\limsup_{n\to\infty} P_0\left[\sup_{\rho(s,t)<\delta_0} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon\right] < \eta$$

with $\rho(s,t) := [\int\{\phi_{0,t}(x) - \phi_{0,s}(x)\}^2 dP_0(x)]^{1/2}$. By (A2), $\sup_{\|t-s\|\leq h}\rho(t,s) < \delta_0$ for some $h > 0$. Hence, for all n large, both $\delta n^{-1/2} \leq h$ and

$$P_0\left[\sup_{\rho(s,t)<\delta_0} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon\right] < \eta \ ,$$

so that

$$P_0\left[\sup_{\|t-s\|\leq\delta/n^{1/2}} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon\right] \leq P_0\left[\sup_{\rho(t,s)<\delta_0} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon\right] < \eta$$

and the proof is complete. □

A.5. Proof of Propositions 1 and 2

Below, we refer to van der Vaart and Wellner (1996) as VW. Throughout, the symbol $\lesssim$ should be interpreted to mean 'bounded above, up to a multiplicative constant not depending on n, t, y or a.'

We first note that condition (M) of Stupfler (2016) guarantees that the class

$$\left\{x \mapsto K\left(\frac{x-t}{h}\right) : h > 0, t \in \mathbb{R}\right\}$$

is Vapnik–Chervonenkis (henceforth VC) with index 2. In addition, we define $K_j := \int u^j K(u)\, du$ and

$$\begin{aligned} w_n(a,t_2) &:= s_{2,n}(t_2) - s_{1,n}(t_2)(a - t_2) \ , \\ w_0(a,t_2) &:= f_0(t_2) - f_0'(t_2)(a - t_2) \ . \end{aligned}$$

Before proving Propositions 1 and 2, we state and prove a lemma we will use.

Lemma 4. *If (d)–(f) hold, $nh_n^4 \to \infty$ and $nh_n^5 = O(1)$, then*

$$\begin{aligned} &(nh_n^5)^{1/2} \sup_{t\in\mathcal{T}} \left|\frac{s_{1,n}(t_2)}{g_n(t_2)} - \frac{f_0'(t_2)}{f_0(t_2)^2}\right| \xrightarrow{\mathrm{P}} 0 \ , \\ &(nh_n^5)^{1/2} \sup_{t\in\mathcal{T}} \left|\frac{s_{2,n}(t_2)}{g_n(t_2)} - \frac{1}{f_0(t_2)}\right| \xrightarrow{\mathrm{P}} 0 \ , \\ &(nh_n^5)^{1/2} \sup_{t\in\mathcal{T}} \sup_{|a-t_2|\leq h_n} \left|\frac{w_n(a,t_2)}{g_n(t_2)} - \frac{w_0(a,t_2)}{f_0(t_2)^2}\right| \xrightarrow{\mathrm{P}} 0 \ , \end{aligned}$$

and for any $\delta > 0$,

$$\begin{aligned} &(nh_n^5)^{1/2} \sup_{\|t-s\|\leq\delta/(nh_n)^{1/2}} \left|\frac{s_{1,n}(t_2)}{g_n(t_2)} - \frac{s_{1,n}(s_2)}{g_n(s_2)}\right| \xrightarrow{\mathrm{P}} 0 \ , \\ &(nh_n^4) \sup_{\|t-s\|\leq\delta/(nh_n)^{1/2}} \left|\frac{s_{2,n}(t_2)}{g_n(t_2)} - \frac{s_{2,n}(s_2)}{g_n(s_2)}\right| \xrightarrow{\mathrm{P}} 0 \ . \end{aligned}$$

Proof of Lemma 4. We first show that $\sup_{t\in\mathcal{T}} |s_{0,n}(t_2) - f_0(t_2)| = o_P(h_n)$. We have that

$$\begin{aligned} s_{0,n}(t_2) - f_0(t_2) &= h_n^{-1} \int K\left(\frac{a-t_2}{h_n}\right) f_0(a)\, da - f_0(t_2) \\ &\qquad + n^{-1/2} h_n^{-1} \mathbb{G}_n K\left(\frac{\cdot - t_2}{h_n}\right) . \end{aligned}$$

By the change of variables $u = (a - t_2)/h_n$, we have that

$$\begin{aligned} &h_n^{-1} \int K\left(\frac{a-t_2}{h_n}\right) f_0(a)\, da - f_0(t_2) \\ &\qquad = \int K(u) \left[f_0(t_2 + h_n u) - f_0(t_2)\right] du \\ &\qquad = h_n \int u K(u) (h_n u)^{-1} R_f^{(1)}\left((t_1, t_2), h_n u\right) du , \end{aligned}$$

which, in view of the assumed uniform negligibility of $R_f^{(1)}$, tends to zero uniformly over t_2 faster than h_n. For the second term, since K is uniformly bounded and the class

$$\left\{ a \mapsto K\left(\frac{a-t_2}{h_n}\right) : t_2 \in [0,1] \right\}$$

is P_0-Donsker, as implied by condition (M) of Stupfler (2016), Theorem 2.14.1 of VW implies that

$$\sup_{t_2} \left| \mathbb{G}_n K\left(\frac{\cdot - t_2}{h_n}\right) \right| = O_P(1) .$$

Then, since $n^{-1/2}h_n^{-1} = h_n \left(nh_n^4\right)^{-1/2} = o_P(h_n)$, this term is also $o_P(h_n)$.

We next show that $\left(nh_n^5\right)^{1/2} \sup_{t\in\mathcal{T}} \left|h_n^{-2} s_{1,n}(t_2) - f_0'(t_2) K_2\right| = o_P(1)$. We have that

$$\begin{aligned} (nh_n)^{1/2} s_{1,n}(t_2) &= \left(nh_n^{-1}\right)^{1/2} \int (a - t_2) K\left(\frac{a-t_2}{h_n}\right) f_0(a)\, da \\ &\qquad + h_n^{-1/2} \iint (a-t_2) K\left(\frac{a-t_2}{h_n}\right) \mathbb{G}_n(dy, da) . \end{aligned}$$

By the change of variables $u = (a - t_2)/h_n$, the first term equals

$$\begin{aligned} &(nh_n^3)^{1/2} \int u K(u) f_0(t_2 + h_n u)\, du \\ &\quad = (nh_n^3)^{1/2} \int u K(u) \left[f_0(t_2 + h_n u) - f_0(t) - (h_n u) f_0'(t_2)\right] du \\ &\qquad + \left(nh_n^5\right)^{1/2} f_0'(t_2) K_2 \\ &\quad = (nh_n^5)^{1/2} \int u K(u) h_n^{-1} R_f^{(1)}\left((t_1, t_2), h_n u\right) du + (nh_n^5)^{1/2} f_0'(t_2) K_2 . \end{aligned}$$

By the assumed uniform negligibility of $R_f^{(1)}$ and since $h_n = O(n^{-1/5})$, the first term tends to zero in probability uniformly over $t \in \mathcal{T}$.

Turning to the second term in $s_{1,n}(t_2)$, we will apply Theorem 2.14.1 of VW to obtain a tail bound for the supremum of this empirical process over the one-dimensional class indexed by t_2. We note that, since K is bounded by some $\bar{K}$ and supported on $[-1, 1]$,

$$\left|(a - t_2)K\left(\frac{a - t_2}{h_n}\right)\right| \leq \bar{K}|a - t_2| I\left(|a - t_2| \leq h_n\right) \leq \bar{K}h_n \ .$$

Therefore, the class of functions

$$\left\{(y, a) \mapsto (a - t_2)K\left(\frac{a - t_2}{h_n}\right) : (t_1, t_2) \in \mathcal{T}\right\}$$

has envelope $\bar{K}h_n$. Furthermore, since $(y, a) \mapsto (a - t_2)$ and K are both uniformly bounded VC classes of functions, and K is bounded, the class of functions possesses finite entropy integral. Hence, we have that

$$E_0\left[\sup_{(t_1,t_2)\in\mathcal{T}}\left|h_n^{-1/2}\iint(a - t_2)K\left(\frac{a - t_2}{h_n}\right)\mathbb{G}_n(dy, da)\right|\right] \ \leq\ C'h_n^{1/2} \longrightarrow 0 \ .$$

We now have that $(nh_n)^{1/2}\sup_{t\in\mathcal{T}}\left|s_{1,n}(t_2) - h_n^2 f_0'(t_2)K_2\right| = o_P(1)$, which implies in particular that

$$\sup_{t\in\mathcal{T}}|s_{1,n}(t_2)| = (nh_n)^{-1/2}o_P(1) + h_n^2 O_P(1) = O_P\left((nh_n)^{-1/2}\right) \ .$$

Next, we show that $\left(nh_n^5\right)^{1/2}\sup_{t\in\mathcal{T}}\left|h_n^{-2}s_{2,n}(t_2) - f_0(t_2)K_2\right| = o_P(h_n)$. The proof of this is nearly identical to the preceding proof. We have that

$$\begin{aligned}(nh_n)^{1/2}s_{2,n}(t_2) \ &=\ \left(nh_n^{-1}\right)^{1/2}\int(a - t_2)^2K\left(\frac{a - t_2}{h_n}\right)f_0(a)\,da \\ &\quad + h_n^{-1/2}\iint(a - t_2)^2K\left(\frac{a - t_2}{h_n}\right)\mathbb{G}_n(dy, da) \ .\end{aligned}$$

By the change of variables $u = (a - t_2)/h_n$, the first term equals

$$\begin{aligned}(nh_n^5)^{1/2}\int u^2K(u)\,f_0(t_2 + h_nu)\,du \ &=\ (nh_n^5)^{1/2}h_n\int u^3K(u)\,\frac{R_f^{(1)}(t, h_nu)}{h_nu}\,du \\ &\quad + \left(nh_n^5\right)^{1/2}f_0(t_2)K_2 \ .\end{aligned}$$

By the uniform negligibility of $R_f^{(1)}$, the first term is $o_P(h_n)$ uniformly in t.

Analysis of the second term in $s_{2,n}$ is analogous to that of $s_{1,n}$, except that the envelope function is now $\bar{K}h_n^2$, so that the empirical process term is $O_P(h_n^{3/2})$. We also note that $\sup_{t_2}|s_{2,n}(t_2)| = O_P\left((nh_n)^{-1/2}\right)$.

The above derivations imply that

$$
\begin{aligned}
&\left(nh_n^5\right)^{1/2} \sup_{t\in\mathfrak{T}} \left|h_n^{-2} g_n(t_2) - f_0(t_2)^2 K_2\right| \\
&\leq\ \left(nh_n^5\right)^{1/2} \sup_{t\in\mathfrak{T}} \left|\left[h_n^{-2} s_{2,n}(t_2) - f_0(t_2) K_2\right] s_{0,n}(t_2)\right| + (nh_n)^{1/2} \left[\sup_{t\in\mathfrak{T}} |s_{1,n}(t_2)|\right]^2 \\
&\quad + \left(nh_n^5\right)^{1/2} \sup_{t\in\mathfrak{T}} \left|\left[s_{0,n}(t_2) - f_0(t_2)\right] f_0(t_2) K_2\right| \\
&=\ o_P(1) O_P(1) + \left(nh_n^5\right)^{1/2} o_P(h_n) + (nh_n)^{1/2} O_P\left(\left(nh_n\right)^{-1}\right)\ =\ o_P(1)\ .
\end{aligned}
$$

We now proceed to the statements in the lemma. We write that

$$
\begin{aligned}
\left|\frac{s_{n,1}(t_2)}{g_n(t_2)} - \frac{f_0'(t_2)}{f_0(t_2)^2}\right|\ &=\ \left|\frac{h_n^{-2} s_{n,1}(t_2)}{h_n^{-2} g_n(t_2)} - \frac{f_0'(t_2) K_2}{f_0(t_2)^2 K_2}\right| \\
=\ &\left|\frac{h_n^{-2} s_{n,1}(t_2) - f_0'(t_2) K_2}{h_n^{-2} g_n(t_2)} - f_0'(t_2) K_2 \frac{h_n^{-2} g_n(t_2) - f_0(t_2)^2 K_2}{h_n^{-2} g_n(t_2) f_0(t_2)^2 K_2}\right| \\
\leq\ &\frac{\left|h_n^{-2} s_{n,1}(t_2) - f_0'(t_2) K_2\right|}{h_n^{-2} g_n(t_2)} + f_0'(t_2) K_2 \frac{\left|h_n^{-2} g_n(t_2) - f_0(t_2)^2 K_2\right|}{h_n^{-2} g_n(t_2) f_0(t_2)^2 K_2}\ .
\end{aligned}
$$

Since $\inf_{t\in\mathfrak{T}} |f_0(t_2)| > 0$, we have that

$$
\sup_{t\in\mathfrak{T}} \left[h_n^{-2} g_n(t_2)\right]^{-1} = O_P(1) \ \text{ and } \ \sup_{t\in\mathfrak{T}} \left[h_n^{-2} g_n(t_2) f_0(t_2)^2\right]^{-1} = O_P(1)\ ,
$$

and the result follows.

We omit the proof of the statement about $s_{n,2}$, since it is almost identical to the above. For the statement about w_n, by the above calculations, we have that

$$
\left(nh_n^5\right)^{1/2} \sup_{(t_1,t_2)\in\mathfrak{T}} \sup_{|a-t_2|\leq h_n} \left|h_n^{-2} w_n(a,t_2) - w_0(a,t_2) K_2\right| \xrightarrow{\mathrm{P}} 0\ .
$$

We write that

$$
\begin{aligned}
&\left|\frac{w_n(a,t_2)}{g_n(t_2)} - \frac{w_0(a,t_2)}{f_0(t_2)^2}\right| = \left|\frac{h_n^{-2} w_n(a,t_2)}{h_n^{-2} g_n(t_2)} - \frac{w_0(a,t_2) K_2}{f_0(t_2)^2 K_2}\right| \\
&\quad = \left|\frac{h_n^{-2} w_n(a,t_2) - w_0(a,t_2) K_2}{h_n^{-2} g_n(t_2)} - w_0(a,t_2) \frac{h_n^{-2} g_n(t_2) - f_0(t_2)^2 K_2}{h_n^{-2} g_n(t_2) f_0(t_2)^2}\right| \\
&\quad \leq \left[h_n^{-2} g_n(t_2)\right]^{-1} \left|h_n^{-2} w_n(a,t_2) - w_0(a,t_2) K_2\right| \\
&\qquad + |w_0(a,t_2)| \left[h_n^{-2} g_n(t_2) f_0(t_2)^2\right]^{-1} \left|h_n^{-2} g_n(t_2) - f_0(t_2)^2 K_2\right|
\end{aligned}
$$

and the result follows.

We note that the above implies that

$$\begin{aligned}
&\sup_{|t_2-s_2|\leq\eta} |s_{1,n}(t_2) - s_{1,n}(s_2)| \\
&\quad\leq\ 2\sup_{t_2} \left|s_{1,n}(t_2) - h_n^2 f_0'(t_2)K_2\right| + h_n^2 \sup_{|t_2-s_2|\leq\eta} |f_0'(t_2) - f_0'(s_2)|K_2 \\
&\quad\lesssim\ o_P\left((nh_n)^{-1/2}\right) + h_n^2\eta\ ,
\end{aligned}$$

so that

$$\sup_{|t_2-s_2|\leq\delta/(nh_n)^{-1/2}} |s_{1,n}(t_2) - s_{1,n}(s_2)| = o_P\left((nh_n)^{-1/2}\right).$$

Similarly, we have that $\sup_{|t_2-s_2|\leq\eta} |s_{0,n}(t_2) - s_{0,n}(s_2)| = o_P\left(h_n\right)$ and

$$\sup_{|t_2-s_2|\leq\eta} |s_{2,n}(t_2) - s_{2,n}(s_2)| = o_P\left(h_n\left(nh_n\right)^{-1/2}\right).$$

Therefore, we find that

$$\begin{aligned}
&\sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}} |g_n(t_2) - g_n(s_2)| \\
&\quad\leq\ \sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}} \left|\left[s_{0,n}(t_2) - s_{0,n}(s_2)\right] s_{2,n}(s_2)\right| \\
&\qquad\qquad + \sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}} \left|s_{0,n}(t_2)\left[s_{2,n}(t_2) - s_{2,n}(s_2)\right]\right| \\
&\qquad\qquad + \sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}} \left|\left[s_{1,n}(t_2) - s_{1,n}(s_2)\right]\left[s_{1,n}(t_2) + s_{1,n}(s_2)\right]\right| \\
&\quad\lesssim\ o_P(h_n)O_P\left((nh_n)^{-1/2}\right) + O_P(1)o_P\left(h_n\left(nh_n\right)^{-1/2}\right) \\
&\qquad + o_P\left((nh_n)^{-1/2}\right)O_P\left((nh_n)^{-1/2}\right) \\
&\quad=\ o_P\left(h_n\left(nh_n\right)^{-1/2}\right)\ .
\end{aligned}$$

We can now write that

$$\begin{aligned}
&\sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}} \left|\frac{s_{1,n}(t_2)}{g_n(t_2)} - \frac{s_{1,n}(s_2)}{g_n(s_2)}\right| \\
&\qquad\leq\ h_n^{-2} \sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}} \left|\frac{s_{1,n}(t_2) - s_{1,n}(s_2)}{h_n^{-2}g_n(t_2)}\right| \\
&\qquad\quad + h_n^{-4} \sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}} \left|s_{1,n}(s_2)\frac{g_n(t_2) - g_n(s_2)}{h_n^{-2}g_n(t_2)h_n^{-2}g_n(s_2)}\right| \\
&\qquad=\ h_n^{-2}o_P\left((nh_n)^{-1/2}\right) + h_n^{-4}O_P\left((nh_n)^{-1/2}\right)o_P\left(h_n\left(nh_n\right)^{-1/2}\right) \\
&\qquad=\ o_P\left(\left(nh_n^5\right)^{-1/2}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}}\left|\frac{s_{2,n}(t_2)}{g_n(t_2)}-\frac{s_{2,n}(s_2)}{g_n(s_2)}\right| \\
&\quad\leq\ h_n^{-2}\sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}}\left|\frac{s_{2,n}(t_2)-s_{2,n}(s_2)}{h_n^{-2}g_n(t_2)}\right| \\
&\qquad+h_n^{-4}\sup_{\|t-s\|\leq\delta/(nh_n)^{-1/2}}\left|s_{2,n}(s_2)\frac{g_n(t_2)-g_n(s_2)}{h_n^{-2}g_n(t_2)h_n^{-2}g_n(s_2)}\right| \\
&\quad=\ h_n^{-2}o_P\left(h_n\left(nh_n\right)^{-1/2}\right)+h_n^{-4}O_P\left(\left(nh_n\right)^{-1/2}\right)o_P\left(h_n\left(nh_n\right)^{-1/2}\right) \\
&\quad=\ o_P\left(\left(nh_n^4\right)^{-1}\right)\ . \qquad\square
\end{aligned}$$

We can now prove Proposition 1.

Proof of Proposition 1. We define

$$\begin{aligned}
m_{1,n}(t_1,t_2) &:= h_n^{-1}\iint\left[\theta_0(t_1,a)-\theta_0(t_1,t_2)\right]\frac{w_n(a,t_2)}{g_n(t_2)}K\left(\frac{a-t_2}{h_n}\right)\mathbb{P}_n(dy,da) \\
m_{2,n}(t_1,t_2) &:= h_n^{-1}\iint\left[I(y\leq t_1)-\theta_0(t_1,a)\right]\frac{w_n(a,t_2)}{g_n(t_2)}K\left(\frac{a-t_2}{h_n}\right)\mathbb{P}_n(dy,da)\ .
\end{aligned}$$

Then, we have that $\theta_n(t_1,t_2)-\theta_0(t_1,t_2)=m_{1,n}(t_1,t_2)+m_{2,n}(t_1,t_2)$. We note that, since $E_0\left[I(Y\leq t_1)\mid A=a\right]=\theta_0(t_1,a)$, $(nh_n)^{1/2}m_{2,n}(t_1,t_2)$ equals

$$\begin{aligned}
&h_n^{-1/2}\iint\left[I(y\leq t_1)-\theta_0(t_1,a)\right]\frac{w_n(a,t_2)}{g_n(t_2)}K\left(\frac{a-t_2}{h_n}\right)\mathbb{G}_n(dy,da) \\
&\quad=\ h_n^{-1/2}\left[\frac{s_{2,n}(t_2)}{g_n(t_2)}\mathbb{G}_n v_{n,t}-\frac{s_{1,n}(t_2)}{g_n(t_2)}\mathbb{G}_n\left(\ell_t v_{n,t}\right)\right].
\end{aligned}$$

Therefore, we can write that

$$\begin{aligned}
&(nh_n)^{1/2}\left[\theta_n(t_1,t_2)-\theta_0(t_1,t_2)\right]-\left(nh_n^5\right)^{1/2}\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)K_2-R_n(t_1,t_2) \\
&\quad=\ (nh_n)^{1/2}m_{1,n}(t_1,t_2)-\left(nh_n^5\right)^{1/2}\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)K_2\ .
\end{aligned}$$

We now proceed to analyze $m_{1,n}$. We have that $(nh_n)^{1/2}m_{1,n}(t_1,t_2)$ equals

$$\begin{aligned}
&\left(nh_n^{-1}\right)^{1/2}\int\left[\theta_0(t_1,a)-\theta_0(t_1,t_2)\right]\frac{w_n(a,t_2)}{g_n(t_2)}K\left(\frac{a-t_2}{h_n}\right)f_0(a)\,da \\
&\quad+\ h_n^{-1/2}\iint\left[\theta_0(t_1,a)-\theta_0(t_1,t_2)\right]\frac{w_n(a,t_2)}{g_n(t_2)}K\left(\frac{a-t_2}{h_n}\right)\mathbb{G}_n(dy,da)\ .
\end{aligned}$$

The second term in $m_{1,n}$ may be further decomposed as

$$h_n^{-1/2}\frac{s_{2,n}(t_2)}{g_n(t_2)}\mathbb{G}_n\gamma_{t,n}-h_n^{-1/2}\frac{s_{1,n}(t_2)}{g_n(t_2)}\mathbb{G}_n\left(\ell_t\gamma_{t,n}\right)$$

with $\gamma_{t,n}(y,a) := [\theta_0(t_1,a)-\theta_0(t_1,t_2)]\, K\left(\frac{a-t_2}{h_n}\right)$ and $\ell_t(y,a) := a - t_2$. By Lemma 4, we have that

$$\sup_{t\in\mathcal{T}}\left|\frac{s_{2,n}(t_2)}{g_n(t_2)}\right| = O_P\left(\left(nh_n^5\right)^{-1/2}\right),$$

and similarly for $s_{1,n}$. We will use Theorem 2.14.2 of VW to obtain bounds for $\sup_{t\in\mathcal{T}}|\mathbb{G}_n\gamma_{t,n}|$ and $\sup_{t\in\mathcal{T}}|\mathbb{G}_n\left(\ell_t\gamma_{t,n}\right)|$. We first note that, since K is bounded and supported on $[-1,1]$ and θ_0 is Lipschitz on $\mathcal{T}$, $\sup_{t\in\mathcal{T}}|\gamma_{t,n}| \lesssim h_n$ and $\sup_{t\in\mathcal{T}}|\ell_t\gamma_{t,n}| \lesssim h_n^2$. These will be our envelope functions for these classes. Next, since K is Lipschitz, we have that

$$\begin{aligned}
|\gamma_{t,n}-\gamma_{s,n}| \;&\leq\; |[\theta_0(t_1,a)-\theta_0(s_1,a)]-[\theta_0(t_1,t_2)-\theta_0(s_1,s_2)]|\, K\left(\frac{a-s_2}{h_n}\right)\\
&\qquad + |\theta_0(t_1,a)-\theta_0(t_1,t_2)|\left|K\left(\frac{a-t_2}{h_n}\right)-K\left(\frac{a-s_2}{h_n}\right)\right|\\
&\lesssim\; |t_1-s_1| + \|t-s\| + |t_2-s_2|h_n^{-1} \;\lesssim\; \|t-s\|h_n^{-1}\ .
\end{aligned}$$

Therefore, by VW Theorem 2.7.11, we have

$$N_{[]}\left(2\varepsilon h_n^{-1}, \mathcal{G}_n, L_2(P_0)\right) \;\lesssim\; N(\varepsilon, \mathcal{T}, \|\cdot\|) \;\lesssim\; \varepsilon^{-2}\ ,$$

where $\mathcal{G}_n := \{\gamma_{n,t} : t \in \mathcal{T}\}$. Thus, by Theorem 2.14.2 of VW,

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}|\mathbb{G}_n\gamma_{n,t}| \lesssim \int_0^1 \left[N_{[]}\left(\varepsilon h_n, \mathcal{G}_n, L_2(P_0)\right)\right]^{1/2} d\varepsilon\, h_n \lesssim \int_0^1 \left[-\log(\varepsilon h_n^2)\right]^{1/2} d\varepsilon\, h_n\\
&= \;h_n^{-1}\int_0^{h_n^2} \left(-\log\varepsilon\right)^{1/2}\, d\varepsilon \;\lesssim\; h_n^{-1}\{h_n^2\left[\log\left(h_n^{-2}\right)\right]^{1/2}\} \;\lesssim\; h_n\left(\log h_n^{-1}\right)\ ,
\end{aligned}$$

where we have used the fact that $\int_0^z (\log x^{-1})^{1/2}\,dx \lesssim z(\log z^{-1})^{1/2}$ for all z small enough. A similar argument applies to $\sup_{t\in\mathcal{T}}|\mathbb{G}_n(\ell_t\gamma_{n,t})|$. We thus have that the second summand in $m_{1,n}$ is bounded above up to a constant not depending on n and uniformly in t by

$$h_n^{-1/2}O_P\left(\left(nh_n^5\right)^{-1/2}\right)h_n\left(\log h_n^{-1}\right)^{1/2} = O_P\left(\left(\frac{nh_n^4}{\log h_n^{-1}}\right)^{-1/2}\right),$$

which is $o_P(1)$ since $nh_n^4/(\log h_n^{-1}) \to \infty$.

By the change of variables $u = (a-t_2)/h_n$, the first term in $m_{1,n}$ equals

$$\begin{aligned}
&(nh_n)^{1/2}\int [\theta_0(t_1, t_2+h_nu)-\theta_0(t_1,t_2)]\,\frac{w_n(t_2+h_nu,t_2)}{g_n(t_2)}K\left(u\right)f_0(t_2+h_nu)\,du\\
&= (nh_n)^{1/2}\int \left[R_\theta^{(2)}(t,h_nu)+\theta'_{0,t_2}(t_1,t_2)(h_nu)+\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)(h_nu)^2\right]\\
&\qquad\qquad \cdot\left[R_f^{(1)}(t_2,h_nu)+f_0(t_2)+f_0'(t_2)(h_nu)\right]\frac{w_n(t_2+h_nu,t_2)}{g_n(t_2)}K(u)\,du\ .
\end{aligned}$$

Expanding the product, this is equal to

$$
\begin{aligned}
&(nh_n)^{1/2}\int (h_n u)\left[\theta'_{0,t_2}(t_1,t_2)+\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)(h_n u)\right]\left[f_0(t_2)+f_0'(t_2)h_n u\right]\\
&\qquad\qquad\cdot \frac{w_n(t_2+h_n u,t_2)}{g_n(t_2)}K(u)\,du\\
&+(nh_n)^{1/2}\int R_\theta^{(2)}(t,h_n u)\left[f_0(t_2)+f_0'(t_2)h_n u\right]\frac{w_n(t_2+h_n u,t_2)}{g_n(t_2)}K(u)\,du\\
&+(nh_n)^{1/2}\int R_f^{(1)}(t_2,h_n u)(h_n u)\left[\theta'_{0,t_2}(t_1,t_2)+\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)h_n u\right]\\
&\qquad\qquad\cdot \frac{w_n(t_2+h_n u,t_2)}{g_n(t_2)}K(u)\,du\\
&+(nh_n)^{1/2}\int R_\theta^{(2)}(t,h_n u)R_f^{(1)}(t_2,h_n u)\frac{w_n(t_2+h_n u,t_2)}{g_n(t_2)}K(u)\,du\ .
\end{aligned}
$$

By the assumed negligibility of $R_\theta^{(2)}$ and $R_f^{(1)}$ as well as Lemma 4, the second through fourth summands tend to zero in probability uniformly over $\mathcal{T}$. The first term equals

$$
\begin{aligned}
&\int f_0'(t_2)\left[\theta'_{0,t_2}(t_1,t_2)+\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)(h_n u)\right]\\
&\qquad\cdot\left(nh_n^5\right)^{1/2}\left[\frac{w_n(t_2+h_n u,t_2)}{g_n(t_2)}-\frac{w_0(t_2+h_n u,t_2)}{f_0(t_2)^2}\right]u^2K(u)\,du\\
&+\left(nh_n^5\right)^{1/2}\int f_0'(t_2)\left[\theta'_{0,t_2}(t_1,t_2)+\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)(h_n u)\right]\\
&\qquad\qquad\cdot\frac{w_0(t_2+h_n u,t_2)}{f_0(t_2)^2}u^2K(u)\,du\\
&+\left(nh_n^3\right)^{1/2}\int f_0'(t_2)\left[\theta'_{0,t_2}(t_1,t_2)+\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)(h_n u)\right]\\
&\qquad\qquad\cdot\frac{w_n(t_2+h_n u,t_2)}{g_n(t_2)}uK(u)\,du\ .
\end{aligned}
$$

By Lemma 4, the first summand tends to zero uniformly over $\mathcal{T}$. By symmetry of K, the second plus third summands simplifies to

$$
\begin{aligned}
&\left(nh_n^5\right)^{1/2}\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)K_2+\left(nh_n^5\right)^{1/2}\left[\frac{s_{2,n}(t_2)}{g_n(t_2)}-\frac{1}{f_0(t_2)}\right]f_0(t_2)\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)K_2\\
&-\left(nh_n^5\right)^{1/2}\left[\frac{s_{1,n}(t_2)}{g_n(t_2)}-\frac{f_0'(t_2)}{f_0(t_2)^2}\right]\theta'_{0,t_2}(t_1,t_2)f_0(t_2)K_2\ .
\end{aligned}
$$

Once again, the second and third summands tend to zero uniformly over $\mathcal{T}$ by Lemma 4. We have now shown that

$$
\sup_{t\in\mathcal{T}}\left|(nh_n)^{1/2}m_{1,n}(t_1,t_2)-\left(nh_n^5\right)^{1/2}\tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)K_2\right|\xrightarrow{\mathrm{P}}0\ ,
$$

which completes the proof. □

Finally, we prove Proposition 2.

***Proof of Proposition** 2.* Since θ''_{0,t_2} is uniformly continuous and $nh_n^5 = O(1)$,

$$\sup_{\|t-s\|\leq\delta/r_n} \left| \left(nh_n^5\right)^{1/2} \tfrac{1}{2}\theta''_{0,t_2}(t_1,t_2)K_2 - \left(nh_n^5\right)^{1/2} \tfrac{1}{2}\theta''_{0,t_2}(s_1,s_2)K_2 \right| \longrightarrow 0 \ .$$

Therefore, it only remains to show that $\sup_{\|t-s\|\leq\delta/r_n} |R_n(t) - R_n(s)| \xrightarrow{\mathrm{P}} 0$. Recalling that $\ell_t(y,a) := a - t_2$ and

$$\nu_{n,t}(y,a) := \left[I(y \leq t_1) - \theta_0(t_1,a)\right] K\left(\frac{a-t_2}{h_n}\right),$$

we have that $R_n(t) - R_n(s)$ equals

$$\left[\frac{s_{2,n}(t_2)}{g_n(t_2)} - \frac{s_{2,n}(s_2)}{g_n(s_2)}\right] \mathbb{G}_n\nu_{n,t} - \left[\frac{s_{1,n}(t_2)}{g_n(t_2)} - \frac{s_{1,n}(s_2)}{g_n(s_2)}\right] \mathbb{G}_n\left(\ell_t\nu_{n,t}\right)$$
$$+ \frac{s_{2,n}(s_2)}{g_n(s_2)}\mathbb{G}_n\left(\nu_{n,t} - \nu_{n,s}\right) - \frac{s_{1,n}(s_2)}{g_n(s_2)}\mathbb{G}_n\left(\ell_t\nu_{n,t} - \ell_s\nu_{n,s}\right) \ .$$

Focusing first on $\mathbb{G}_n\nu_{n,t}$, we have $\mathbb{G}_n\nu_{n,t} = \mathbb{G}_n\nu_{n,t,1} - \mathbb{G}_n\nu_{n,t,2}$ for

$$\nu_{n,t,1}(y,a) = I(y \leq t_1)K\left(\frac{a-t_2}{h_n}\right) \quad \text{and} \quad \nu_{n,t,2}(y,a) = \theta_0(t_1,a)K\left(\frac{a-t_2}{h_n}\right).$$

The classes

$$\{I(y \leq t_1) : t \in \mathcal{T}\} \quad \text{and} \quad \left\{K\left(\frac{a-t_2}{h_n}\right) : t \in \mathcal{T}\right\}$$

are both uniformly bounded above and VC. Therefore, the uniform covering numbers of the class

$$\left\{I(y \leq t_1)K\left(\frac{a-t_2}{h_n}\right) : t \in \mathcal{T}\right\}$$

are bounded up to a constant by ε^{-V} for some $V < \infty$, so that the uniform entropy integral satisfies

$$J(\eta, \mathcal{G}_{n,1}) \ \lesssim \ \eta\left(\log\eta^{-1}\right)^{1/2}$$

for all η small enough, where $\mathcal{G}_{n,1} := \{\nu_{n,t,1} : t \in \mathcal{T}\}$. We also have $P_0\left(\nu_{n,t,1}\right)^2 \lesssim h_n$ for all $t \in \mathcal{T}$ and all n large enough. Thus, Theorem 2.1 of van der Vaart and Wellner (2011) implies that

$$\sup_{t\in\mathcal{T}} |\mathbb{G}_n\nu_{n,t,1}| \ \lesssim \ h_n^{1/2}\left(\log h_n^{-1}\right)^{1/2} + n^{-1/2}\log h_n^{-1} \ .$$

For $\mathbb{G}_n\nu_{n,t,2}$, we have that

$$|\nu_{n,t,2}(y,a) - \nu_{n,s,2}(y,a)| \ \lesssim \ \|t-s\|(1+h_n^{-1}) \ \lesssim \ h_n^{-1}\|t-s\|$$

for all n large enough and all (y, a). We can therefore apply Theorem 2.7.11 of VW to conclude that $N_{[]}\left(2\varepsilon h_n^{-1}, \mathcal{G}_{n,2}, L_2(P_0)\right) \lesssim \varepsilon^{-2}$ for all ε small enough, where $\mathcal{G}_{n,2} = \{\nu_{n,t,2} : t \in \mathcal{T}\}$, which implies that $N_{[]}\left(\varepsilon, \mathcal{G}_{n,2}, L_2(P_0)\right) \lesssim (\varepsilon h_n)^{-2}$. Thus, we have that

$$J_{[]}(\eta, \mathcal{G}_{n,2}) \ \lesssim \ \eta\left[\log(\eta h_n)^{-1}\right]^{1/2} \ .$$

Since $P_0\left(\nu_{n,t,2}\right)^2 \lesssim h_n$ as well, by Lemma 3.4.2 of VW, we then have

$$E_{P_0} \sup_{t\in\mathcal{T}} \left|\mathbb{G}_n \nu_{n,t,2}\right| \ \lesssim \ h_n^{1/2}\left(\log h_n^{-1}\right)^{1/2} + n^{-1/2}\log h_n^{-1} \ .$$

Combining these two bounds with the last statement of Lemma 4 yields

$$\begin{aligned} & h_n^{-1/2} \sup_{t\in\mathcal{T}} \left|\left[\frac{s_{2,n}(t_2)}{g_n(t_2)} - \frac{s_{2,n}(s_2)}{g_n(s_2)}\right]\mathbb{G}_n\nu_{n,t}\right| \\ & \lesssim \ h_n^{-1/2} o_P\left(\left(nh_n^4\right)^{-1}\right) O_P\left(h_n^{1/2}\left(\log h_n^{-1}\right)^{1/2} + n^{-1/2}\log h_n^{-1}\right) \\ & = \ o_P(1)\left[\frac{nh_n^4}{\left(\log h_n^{-1}\right)^{1/2}}\right]^{-1} + o_P(1)\left(nh_n^{11/3}\right)^{-3/2} h_n \log h_n^{-1} \ . \end{aligned}$$

Both terms tend to zero.

The analysis for $\mathbb{G}_n\left(\ell_t \nu_{n,t}\right)$ is very similar. In this case, we have $P_0\left(\ell_t\nu_{n,t}\right)^2 \lesssim h_n^3$, so that, using the same approach as above, we get

$$E_{P_0} \sup_{t\in\mathcal{T}} \left|\mathbb{G}_n\left(\ell_t\nu_{n,t}\right)\right| \ \lesssim \ h_n^{3/2}\left(\log h_n^{-1}\right)^{1/2} + n^{-1/2}\left(\log h_n^{-1}\right)^{1/2}$$

and therefore, in view of Lemma 4,

$$\begin{aligned} & h_n^{-1/2} \sup_{t\in\mathcal{T}} \left|\left[\frac{s_{1,n}(t_2)}{g_n(t_2)} - \frac{s_{1,n}(s_2)}{g_n(s_2)}\right]\mathbb{G}_n\left(\ell_t\nu_{n,t}\right)\right| \\ & \lesssim \ h_n^{-1/2} o_P\left(\left(nh_n^5\right)^{-1/2}\right) O_P\left(h_n^{3/2}\left(\log h_n^{-1}\right)^{1/2} + n^{-1/2}\left(\log h_n^{-1}\right)^{1/2}\right) \\ & = \ o_P(1)\left(nh_n^4\right)^{-1/2}\left(h_n \log h_n^{-1}\right)^{1/2} + o_P(1)\left(nh_n^3\right)^{-1/2}\left(h_n\log h_n^{-1}\right)^{1/2} \ , \end{aligned}$$

which goes to zero in probability.

It remains to bound

$$\sup_{\|t-s\|<\delta/r_n} \left|\mathbb{G}_n\left(\nu_{n,t} - \nu_{n,s}\right)\right| \quad \text{and} \quad \sup_{\|t-s\|<\delta/r_n} \left|\mathbb{G}_n\left(\ell_t\nu_{n,t} - \ell_s\nu_{n,s}\right)\right| \ .$$

For the former, we work on the terms $\mathbb{G}_n\left(\nu_{n,t,1} - \nu_{n,s,1}\right)$ and $\mathbb{G}_n\left(\nu_{n,t,2} - \nu_{n,s,2}\right)$ separately. For the first of these, we let $\mathcal{F}_{n,\delta,2} := \{\nu_{n,t,1} - \nu_{n,s,1} : \|t - s\| \leq \delta/r_n\}$. We have that $\left\|\nu_{n,t,1} - \nu_{n,s,1}\right\|_{P_0,2}$ is bounded above by

$$\left[E_{P_0}\left\{\left[I(Y \leq t_1) - I(Y \leq s_1)\right]^2 K\left(\frac{A - s_2}{h_n}\right)^2\right\}\right]^{1/2}$$

$$
\begin{aligned}
&+ \left[E_{P_0}\left\{I(Y \leq t_1)\left[K\left(\frac{A-t_2}{h_n}\right) - K\left(\frac{A-s_2}{h_n}\right)\right]^2\right\}\right]^{1/2} \\
&\lesssim \left(E_{P_0}\left\{I(s_1 < Y \leq t_1) I\left(|A - s_2| \leq h_n\right)\right\}\right)^{1/2} + h_n^{-1}|t_2 - s_2| \\
&\lesssim h_n^{1/2}\left|t_1 - s_1\right|^{1/2} + h_n^{-1}\left|t_2 - s_2\right| \ .
\end{aligned}
$$

Therefore, it follows that

$$
\sup_{f \in \mathcal{F}_{n,\delta,1}} \left(P_0 f^2\right)^{1/2} \lesssim \left(n h_n^{-1}\right)^{-1/4} + \left(n h_n^3\right)^{-1/2} \lesssim \left(n h_n^3\right)^{-1/2}
$$

for all n large enough. In addition, $\mathcal{F}_{n,\delta,1}$ has uniform covering numbers bounded up to a constant by ε^{-V} for all n and δ because the classes $\{I(y \leq t_1) : t \in \mathcal{T}\}$ and

$$
\left\{K\left(\frac{a - t_2}{h_n}\right) : t \in \mathcal{T}\right\}
$$

are VC. Therefore, $J\left(\eta, \mathcal{F}_{n,\delta,1}\right) \lesssim \eta\left(\log \eta^{-1}\right)^{1/2}$ for all η small enough. Thus, Theorem 2.1 of VW implies that

$$
E_{P_0} \sup_{\|t-s\| \leq \delta/r_n} \left|\mathbb{G}_n\left(\nu_{n,t,1} - \nu_{n,s,1}\right)\right| \lesssim \left[\frac{\log\left(n h_n^3\right)}{n h_n^3}\right]^{1/2} + \frac{\log\left(n h_n^3\right)}{n^{1/2}} \ .
$$

Turning to $\mathbb{G}_n\left(\nu_{n,t,2} - \nu_{n,s,2}\right)$, we analogously define

$$
\mathcal{F}_{n,\delta,2} := \left\{\nu_{n,t,2} - \nu_{n,s,2} : \|t - s\| \leq \delta/r_n\right\}.
$$

By the Lipschitz property of θ_0 and K, we have that

$$
\left|\theta_0(t_1, a) K\left(\frac{a - t_2}{h_n}\right) - \theta_0(s_1, a) K\left(\frac{a - s_2}{h_n}\right)\right| \lesssim h_n^{-1}\|t - s\| \ .
$$

Therefore, up to a constant, an envelope function $F_{n,\delta,2}$ for $\mathcal{F}_{n,\delta,2}$ is given by $h_n^{-1}\delta/r_n \lesssim \left(n h_n^3\right)^{-1/2}$. Next, we have, for any (t, s) and (t', s') in $\mathcal{T}^2$,

$$
\begin{aligned}
&\Bigg|\left[\theta_0(t_1, a) K\left(\frac{a - t_2}{h_n}\right) - \theta_0(s_1, a) K\left(\frac{a - s_2}{h_n}\right)\right] \\
&\qquad - \left[\theta_0(t_1', a) K\left(\frac{a - t_2'}{h_n}\right) - \theta_0(s_1', a) K\left(\frac{a - s_2'}{h_n}\right)\right]\Bigg| \\
&\leq \left|\theta_0(t_1, a) - \theta_0(t_1', a)\right| K\left(\frac{a - t_2'}{h_n}\right) \\
&\qquad + \left|\theta_0(t_1, a)\right| \left|K\left(\frac{a - t_2}{h_n}\right) - K\left(\frac{a - t_2'}{h_n}\right)\right| \\
&\qquad + \left|\theta_0(s_1, a) - \theta_0(s_1', a)\right| K\left(\frac{a - s_2'}{h_n}\right)
\end{aligned}
$$

$$\begin{aligned}
&\quad + |\theta_0(s_1, a)| \left| K\left(\frac{a - s_2}{h_n}\right) - K\left(\frac{a - s_2'}{h_n}\right)\right| \\
&\lesssim\ |t_1 - t_1'| + h_n^{-1}|t_2 - t_2'| + |s_1 - s_1'| + h_n^{-1}|s_2 - s_2'|\ \lesssim\ h_n^{-1}\left\|(t,s) - (t',s')\right\|_{\mathcal{T}^2}
\end{aligned}$$

with $\|(t,s) - (t',s')\|_{\mathcal{T}^2} := \max\{\|t-s\|, \|t'-s'\|\}$. Therefore, by Theorem 2.7.11 of VW, we have that

$$N_{[]}\left(2\varepsilon h_n^{-1}, \mathcal{F}_{n,\delta,2}, L_2(P_0)\right)\ \leq\ N\left(\varepsilon, \mathcal{U}_{\delta/r_n}, \|\cdot\|_{\mathcal{T}^2}\right),$$

where $\mathcal{U}_{\delta/r_n} := \{(t,s) \in \mathcal{T}^2 : \|t-s\| \leq \delta/r_n\}$. Since $\mathcal{U}_{\delta/r_n} \subseteq \mathcal{T}^2$, we trivially have that $N\left(\varepsilon, \mathcal{U}_{\delta/r_n}, \|\cdot\|_{\mathcal{T}^2}\right) \lesssim \varepsilon^{-4}$. Thus, it follows that

$$N_{[]}\left(\varepsilon\left(nh_n^3\right)^{-1/2}, \mathcal{F}_{n,\delta,2}, L_2(P_0)\right)\ \lesssim\ \left[\varepsilon\left(nh_n\right)^{-1/2}\right]^{-4}.$$

Therefore, Theorem 2.14.2 of VW implies that

$$\begin{aligned}
&E_{P_0} \sup_{\|t-s\|\leq\delta/r_n} \left|\mathbb{G}_n\left(\nu_{n,t,2} - \nu_{n,s,2}\right)\right| \\
&\lesssim\ \left(nh_n^3\right)^{-1/2} \int_0^1 \left\{\log\left[\varepsilon(nh_n)^{-1/2}\right]^{-1}\right\}^{1/2} d\varepsilon \\
&=\ \left(nh_n^3\right)^{-1/2}\left(nh_n\right)^{1/2} \int_0^{(nh_n)^{-1/2}} \left(\log u^{-1}\right)^{1/2} du\ \lesssim\ \left(nh_n^3\right)^{-1/2}\left[\log(nh_n)\right]^{1/2}.
\end{aligned}$$

We now have that

$$\begin{aligned}
&h_n^{-1/2} \sup_{\|t-s\|\leq\delta/r_n} \left| \frac{s_{2,n}(s_2)}{g_n(s_2)} \mathbb{G}_n\left(\nu_{n,t} - \nu_{n,s}\right)\right| \\
&\lesssim\ h_n^{-1/2} O_P\left(\left(nh_n^5\right)^{-1/2}\right) O_P\left(\left(nh_n^3\right)^{-1/2}\left[\log\left(nh_n^3\right)\right]^{1/2} + n^{-1/2}\log\left(nh_n^3\right)\right) \\
&=\ O_P(1)\left\{(nh_n^{9/2})^{-1}\left[\log\left(nh_n^3\right)\right]^{1/2} + (nh_n^3)^{-1}\log\left(nh_n^3\right)\right\}.
\end{aligned}$$

Both terms tend to zero in probability.

We can address $\sup_{\|t-s\|<\delta/r_n}\left|\mathbb{G}_n\left(\ell_t\nu_{n,t} - \ell_s\nu_{n,s}\right)\right|$ in a very similar manner. As before, we work on terms $\mathbb{G}_n\left(\ell_t\nu_{n,t,1} - \ell_s\nu_{n,s,1}\right)$ and $\mathbb{G}_n\left(\ell_t\nu_{n,t,2} - \ell_s\nu_{n,s,2}\right)$ separately. It is straightforward to see that the same line of reasoning as used above applies to each of these terms as well, yielding the same negligibility. □

Acknowledgements

The authors gratefully acknowledge the constructive comments of the editors and anonymous reviewers as well as grant support from the National Institute of Allergy and Infectious Diseases (NIAID) and the National Heart, Lung and Blood Institute (NHLBI) of the National Institutes of Health.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. (1955). An Empirical Distribution Function for Sampling with Incomplete Information. *Ann. Math. Statist.* **26** 641–647. MR0073895

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley New York. MR0326887

Bril, G., Dykstra, R., Pillers, C. and Robertson, T. (1984). Algorithm AS 206: Isotonic Regression in Two Independent Variables. *J. R. Stat. Soc. Ser. C. Appl. Stat* **33** 352–357.

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010). Quantile and Probability Curves Without Crossing. *Econometrica* **78** 1093–1125. MR2667913

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96** 559–575. MR2538757

Daouia, A. and Park, B. U. (2013). On Projection-type Estimators of Multivariate Isotonic Functions. *Scandinavian Journal of Statistics* **40** 363–386. MR3066419

Dette, H., Neumeyer, N. and Pilz, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli* **12** 469–490. MR2232727

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. CRC Press, Boca Raton. MR1383587

Fokianos, K., Leucht, A. and Neumann, M. H. (2017). On Integrated L^1 Convergence Rate of an Isotonic Regression Estimator for Multivariate Observations. *arXiv e-prints* arXiv:1710.04813.

Gill, R. D. and Robins, J. M. (2001). Causal Inference for Complex Longitudinal Data: The Continuous Case. *Ann. Statist.* **29** 1785–1811. MR1891746

Hall, P. and Kang, K.-H. (2001). Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *Ann. Statist.* **29** 1443–1468. MR1873338

Hardle, W., Janssen, P. and Serfling, R. (1988). Strong Uniform Consistency Rates for Estimators of Conditional Functionals. *Ann. Statist.* **16** 1428–1449. MR0964932

Kyng, R., Rao, A. and Sachdeva, S. (2015). Fast, Provable Algorithms for Isotonic Regression in all ℓ_p-norms. In *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds.) 2719–2727. Curran Associates, Inc.

Liao, X. and Meyer, M. C. (2014). coneproj: An R Package for the Primal or Dual Cone Projections with Routines for Constrained Regression. *Journal of Statistical Software* **61** 1–22.

Meyer, M. C. (1999). An extension of the mixed primal–dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference* **81** 13–31. MR1718405

Mukarjee, H. and Stern, S. (1994). Feasible Nonparametric Estimation of Multiargument Monotone Functions. *Journal of the American Statistical Association* **89** 77–80. MR1266288

Patra, R. K. and Sen, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 869–893. MR3534354

Robertson, T., Wright, F. and Dykstra, R. (1988). *Order Restricted Statistical Inference*. Wiley, New York. MR0961262

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling* **7** 1393–1512. MR0877758

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90** 1257–1270. MR1379468

Stupfler, G. (2016). On the weak convergence of the kernel density estimator in the uniform topology. *Electron. Commun. Probab.* **21** 13 pp. MR3485386

R Core Team (2018). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Turner, R. (2015). Iso: Functions to Perform Isotonic Regression R package version 0.0-17.

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media. MR1958123

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York. MR1385671

van der Vaart, A. and Wellner, J. A. (2011). A local maximal inequality under uniform entropy. *Electron. J. Statist.* **5** 192–203. MR2792551

Wand, M. (2015). KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995) R package version 2.23-15. MR1319818

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London. MR1319818