Chernozhukov, Victor; Lee, Sokbae; Rosen, Adam M.

Working Paper

Intersection bounds: Estimation and inference

cemmap working paper, No. CWP19/09

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Lee, Sokbae; Rosen, Adam M. (2009) : Intersection bounds: Estimation and inference, cemmap working paper, No. CWP19/09, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2009.1909

This Version is available at:
https://hdl.handle.net/10419/64682

cemmap
centre for microdata methods and practice

Intersection bounds: estimation and inference

Victor Chernozhukov
Sokbae Lee
Adam M. Rosen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP19/09

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

INTERSECTION BOUNDS: ESTIMATION AND INFERENCE

VICTOR CHERNOZHUKOV, SOKBAE LEE, ADAM M. ROSEN

ABSTRACT. We develop a practical and novel method for inference on intersection bounds, namely bounds defined by either the infimum or supremum of a parametric or nonparametric function, or equivalently, the value of a linear programming problem with a potentially infinite constraint set. Our approach is especially convenient in models comprised of a continuum of inequalities that are separable in parameters, and also applies to models with inequalities that are non-separable in parameters. Since analog estimators for intersection bounds can be severely biased in finite samples, routinely underestimating the length of the identified set, we also offer a (downward/upward) median unbiased estimator of these (upper/lower) bounds as a natural by-product of our inferential procedure. Furthermore, our method appears to be the first and currently only method for inference in nonparametric models with a continuum of inequalities. We develop asymptotic theory for our method based on the strong approximation of a sequence of studentized empirical processes by a sequence of Gaussian or other pivotal processes. We provide conditions for the use of nonparametric kernel and series estimators, including a novel result that establishes strong approximation for general series estimators, which may be of independent interest. We illustrate the usefulness of our method with Monte Carlo experiments and an empirical example.

KEY WORDS. Bound analysis, conditional moments, partial identification, strong approximation, infinite dimensional constraints, linear programming, concentration inequalities, anti-concentration inequalities.

JEL SUBJECT CLASSIFICATION. C12, C13, C14.
AMS SUBJECT CLASSIFICATION. 62G05, 62G15, 62G32.

Date: 20 July 2009.

We thank R. Blundell, A. Chesher, F. Molinari, W. Newey, and J. Stoye for detailed discussion and suggestions, and participants at numerous seminars and conferences for their comments. We thank Nicolas Roys for providing excellent research assistance. This paper is a revised version of "Inference on Intersection Bounds" initially presented at the University of Virginia and the Harvard/MIT econometrics seminars in December 2007, as well as the March 2008 CEMMAP/Northwestern conference on "Inference in Partially Identified Models with Applications". Financial support from the Economic and Social Research Council for the ESRC Centre for Microdata Methods and Practice (RES-589-28-0001) and the small research grant (RES-000-22-2761) is gratefully acknowledged.

1. Introduction

This paper develops a practical and novel method for estimation and inference on parameters restricted by intersection bounds. These are settings where the true parameter value, say θ^*, is known to lie within the bounds $\left[\theta^l(v), \theta^u(v)\right]$ for each value v in a possibly infinite set $\mathcal{V}$. The identification region for θ^* is then

$$\Theta_I = \cap_{v\in\mathcal{V}} \left[\theta^l(v), \theta^u(v)\right] = \left[\sup_{v\in\mathcal{V}} \theta^l(v), \inf_{v\in\mathcal{V}} \theta^u(v)\right]. \tag{1.1}$$

Intersection bounds arise naturally from exclusion restrictions (Manski (2003)) and appear in numerous applied and theoretical examples.[1] This paper covers both parametric and non-parametric estimators of the bound-generating functions $v \mapsto \theta_u(v)$ and $v \mapsto \theta_l(v)$, and also covers cases where the constraint set $\mathcal{V}$ is a continuum. Thus, this paper improves upon prior approaches, which only treat finite constraint sets and parametric estimation of bound-generating functions. More generally, the methods of this paper apply to any estimator for the value of a linear programming problem with an infinite dimensional constraint set.

This paper overcomes significant complications for estimation and inference in such contexts. First, since sample analogs of the lower and upper bounds of Θ_I are the suprema and infima of estimated bound-generating functions, they have substantial finite sample bias, and the estimated bounds tend to be much tighter than the population bounds. This has been noted by Manski and Pepper (2000, 2008), and some heuristic bias adjustments have been proposed by Haile and Tamer (2003) and Kreider and Pepper (2007). Second, the fact that the boundary estimates are suprema and infima of parametric or nonparametric empirical processes typically renders closed-form characterization of their asymptotic distributions unavailable or difficult to establish. As a consequence, researchers have typically used the canonical bootstrap for inference. Yet results from the recent literature indicate that the canonical bootstrap is not generally consistent in such settings, see e.g. Andrews and Han (2009), Bugni (2009), and Canay (2009).

[1] Examples include monotone instrumental variables and the returns to schooling (Manski and Pepper (2000)), English auctions (Haile and Tamer (2003)), the returns to language skills (Gonzalez (2005)), set identification with Tobin regressors (Chernozhukov, Rigobon, and Stoker (2007)), endogeneity with discrete outcomes (Chesher (2007)), changes in the distribution of wages (Blundell, Gosling, Ichimura, and Meghir (2007)), the study of disability and employment (Kreider and Pepper (2007)), estimation of income poverty measures (Nicoletti, Foliano, and Peracchi (2007)), unemployment compensation reform (Lee and Wilke (2009)), bounds on the distribution of treatment effects under strong ignorability (Fan (2009)), and set identification with imperfect instruments (Nevo and Rosen (2008)).

We solve the problem of estimation and inference for intersection bounds by proposing (downward or upward) median unbiased estimators of the upper and lower bounds, as well as confidence intervals. Specifically, our approach employs a precision-correction to the estimated bound-generating functions $v \mapsto \widehat{\theta}^l(v)$ and $v \mapsto \widehat{\theta}^u(v)$ before applying the supremum and infimum operators. Indeed, we adjust the estimated bound-generating functions for their precision by adding to each of them an appropriate critical value times their pointwise standard error. Then, depending on the choice of the critical value, the intersection of these precision-adjusted bounds provides (i) a downward median unbiased estimator for the upper bound $\inf_{v\in\mathcal{V}} \theta_u(v)$ and an upward median unbiased estimator for the lower bound $\sup_{v\in\mathcal{V}} \theta_l(v)$ and (ii) confidence sets for either the identified set Θ_I or the true parameter value θ^*.[2] We select the critical value either analytically or via simulation of an approximating Gaussian process. Our method applies in both parametric and non-parametric settings. For both cases we provide formal justification via asymptotic theory based on the strong approximation of a sequence of studentized empirical processes by a sequence of Gaussian or other pivotal processes. This includes an important new result on strong approximation for series estimators that applies to any estimator that admits a linear approximation, essentially providing a functional central limit theorem for series estimators for the first time in the literature. In principle this functional central limit theorem covers linear and non-linear series estimators, both with and without endogeneity.

This paper contributes to a growing literature on inference on set-identified parameters bounded by inequality restrictions. The prior literature has focused primarily on models with a finite number of unconditional inequality restrictions. Some examples include Andrews and Jia (2008), Beresteanu and Molinari (2008), Chernozhukov, Hong, and Tamer (2007), Galichon and Henry (2009), Romano and Shaikh (2008), Romano and Shaikh (2009), and Rosen (2008), among others. To the best of our knowledge, our paper is the first to consider inference with a continuum of inequalities, which includes conditional moment inequalities as a particular case but also covers other examples such as conditional quantile inequalities. Recent papers (some in progress) on conditional moment inequalities, written independently and contemporaneously, include Andrews and Shi (2009), Fan (2009), Kim (2009), and Menzel (2009), and all employ different

[2]We say an estimator is downward (upward) median unbiased if the probability it lies below (above) its target value is less (greater) than or equal to one half asymptotically. Achieving exact median unbiasedness is not possible in full generality.

approaches.[3] Our approach is especially convenient for performing inference in parametric and non-parametric models with a continuum of inequalities that are separable in parameters, and it also applies to inference in models with inequalities that are non-separable in parameters. Furthermore, our method appears to be the first and currently only method available for performing inference with fully nonparametric inequality restrictions. An attractive feature of our approach is that in addition to providing a valid method of inference, we provide a novel construction for (downward or upward) median unbiased estimators for (upper or lower) intersection bounds. In fact, the only difference in the construction of our estimators and confidence intervals is the choice of critical value, which is a quantile of an appropriate approximating distribution. Thus, practitioners need not implement two entirely different methods to construct estimators and confidence bands with desirable properties.

We organize the paper as follows. In section 2, we motivate the analysis with examples and provide an informal overview of our results. In section 3 we provide a formal treatment of our method, providing conditions and theorems for validity in both parametric and nonparametric contexts. In 4 we provide a Monte Carlo study, and in section 5 we give an empirical example. In section 6 we conclude. In the Appendix we provide proofs, establish strong approximation results for both series and kernel estimators, and describe the steps required to implement our method in practice.

2. Motivating Examples and Informal Overview of Results

In this section we briefly describe four examples of intersection bounds from the literature and provide an informal overview of our results.

Example 1: Treatment Effects and Instrumental Variables. In the analysis of treatment response, the ability to uniquely identify the distribution of potential outcomes is typically lacking without either experimental data or strong assumptions. This owes to the fact that for each individual unit of observation, only the outcome from the received treatment is observed; the counterfactual outcome that would have occurred given a

[3]Some approaches, such as Andrews and Shi (2009), rely on Bierens type integrated moment tests and some, such as Menzel (2009), on standard tests with finite inequalities, using an increasing number of inequalities, both of which differ from the approach pursued here. Using goodness-of-fit tests as a simple analogy, our approach is most similar to Kolmogorov-Smirnov type tests, whereas the approach in Andrews and Shi (2009) appears similar to Bierens type tests, and Menzel (2009)'s approach appears similar to Pearson type tests. Just as in the goodness-of-fit literature, none of the approaches are likely to universally dominate others since there are no uniformly most powerful tests in complex settings such as the one considered here.

different treatment is not known. Though we focus here on treatment effects, similar issues are present in other areas of economics. In the analysis of markets, for example, observed equilibrium outcomes reveal quantity demanded at the observed price, but do not reveal what demand would have been given other prices.

To illustrate how bounds on treatment effects fit into our framework, first suppose only that the support of the outcome space is known, but no other assumptions are made regarding the distribution of counterfactual outcomes. Then Manski (1989) and Manski (1990) provide worst-case bounds on mean treatment outcomes for any treatment t conditional on covariates w, $LB_{wc}(w,t) \leq E[Y(t)|w] \leq UB_{wc}(w,t)$. These bounds are conditional expectations of observed random variables, and are thus trivially intersection bounds where the intersection set is singleton. If $w = (x,v)$ and v is an instrumental variable satisfying $E[Y(t)|x,v] = E[Y(t)|x]$, then the sharp bounds on $E[Y(t)|x]$ are $LB_{iv}(x,t) \leq E[Y(t)|x] \leq UB_{iv}(x,t)$, where $LB_{iv}(x,t) = \sup_{v \in V} LB_{wc}((x,v),t)$ and $UB_{iv}(x,t) = \inf_{v \in V} UB_{wc}((x,v),t)$. In this case the identified set is the intersection over the support of the instrument v of the worst-case bounds at $w = (x,v)$. Similarly, bounds implied by restrictions such as monotone treatment response, monotone treatment selection, and monotone instrumental variables, as in Manski (1997) and Manski and Pepper (2000), also take the form of intersection bounds. In particular, the returns to schooling application of section 5 considers estimation and inference on intersection bounds implied by joint monotone treatment selection and monotone instrumental variable restrictions. □

Example 2: Bounding Distributions to Account for Selection. Similar analysis to that of Manski (1994) and Manski and Pepper (2000) can be applied generally to inference on distributions whose observations are censored due to selection. Such an approach is employed by Blundell, Gosling, Ichimura, and Meghir (2007) to study changes in male and female wages, while accounting for the censoring of the wage distribution incurred by selection into employment. The starting point of their analysis is that the cumulative distribution of wages at any point w, conditional on covariates x must satisfy the worst case bounds

$$F(w|x, E=1)P(x) \leq F(w|x) \leq F(w|x, E=1)P(x) + 1 - P(x)$$

where E is an indicator of employment, and $P(x) \equiv \Pr(E=1|x)$. This relation is then used to bound quantiles of the distribution of wages conditional on covariates. The

worst-case bounds are often not very informative, so additional restrictions motivated by economic theory are used to tighten the bounds.

One such restriction is an exclusion restriction of the continuous variable out-of-work income, z, see Blundell, Gosling, Ichimura, and Meghir (2007, pp. 331-333). Two such possibilities are considered: the use of z as an excluded instrument, and the use of z as a monotone instrument. The former restriction implies

$$\begin{aligned}\max_{z}\left\{F\left(w|x,z,E=1\right)P\left(x,z\right)\right\} &\leq F\left(w|x\right)\\ &\leq \min_{z}\left\{F\left(w|x,z,E=1\right)P\left(x,z\right)+1-P\left(x,z\right)\right\},\end{aligned}$$

while the weaker monotonicity restriction implies that for any z_0 on the support of Z,

$$\begin{aligned}\max_{z\geq z_0}\left\{F\left(w|x,z,E=1\right)P\left(x,z\right)\right\} &\leq F\left(w|x,z_0\right)\\ &\leq \min_{z\leq z_0}\left\{F\left(w|x,z,E=1\right)P\left(x,z\right)+1-P\left(x,z\right)\right\}.\end{aligned}$$

□

Example 3: English Auctions. Invoking two weak assumptions on bidder behavior in an independent private values paradigm, Haile and Tamer (2003) use the distribution of observed bids to formulate bounds on the distribution of bidders' valuations. The two assumptions on bidder behavior, which nest various equilibria, are that each bidder's bid is no greater than her valuation, and that bidders who did not win would not have been willing to pay more than the winning bid. Theorems 1 and 2 of Haile and Tamer (2003, pp. 7-10) give the following implied bounds on the cumulative distribution of valuations at any point v,

$$\max_{2\leq n\leq \bar{M}}\phi\left(G^{\Delta}_{n:n}\left(v\right);n-1,n\right)\leq F\left(v\right)\leq \min_{2\leq n\leq \bar{M},\ 1\leq i\leq n}\phi\left(G_{i:n}\left(v\right);i,n\right),$$

where $\bar{M}$ is the number of potential bidders in an auction, and n is the number who actually submit bids. Here, $G_{i:n}$ denotes the distribution of the i^{th} order statistic of bids, and $\phi\left(\cdot;i,n\right)$ is a monotone transformation relating any parent distribution F to the distribution of its i^{th} order statistic, i.e.

$$F\left(v\right)=\phi\left(F_{i:n}\left(v\right);i,n\right).$$

$G^{\Delta}_{n:n}$ denotes the distribution of the n^{th} order statistic of bids, plus minimum bid increment Δ, in an auction of n bidders. The derived bounds fall into the present framework, as the distributions $G^{\Delta}_{n:n}\left(\cdot\right)$ and $G_{i:n}\left(\cdot\right)$ are identified and consistently estimable.

Example 4: Conditional Moment Inequalities. Our inferential method can also be used to conduct pointwise inference on parameters in models comprised of conditional moment inequalities. This can be done whether the conditioning variables are discrete or continuous. Such restrictions arise naturally in empirical work in industrial organization and in particular in models of oligopoly entry, see for example Pakes, Porter, Ho, and Ishii (2005) and Berry and Tamer (2007).

To illustrate, consider the restriction

$$E\left[m\left(x,\gamma_0\right)|v\right] \geq 0 \text{ for every } v \in \mathcal{V}, \tag{2.1}$$

where $m\left(\cdot,\cdot\right)$ is a real-valued function, (x,v) are random variables observable by the econometrician, and γ_0 is the parameter of interest. For example, in a model of oligopoly entry γ_0 could measure the effect of one firm's entry decision on a rival's profit. It may be of interest to test whether γ_0 is equal to some conjectured value γ, e.g. $\gamma = 0$. To see how our framework can be used to test this hypothesis, define $\theta\left(\gamma,v\right) := E\left[m\left(x,\gamma\right)|v\right]$ and $\widehat{\theta}\left(\gamma,v\right)$ a consistent estimator. Suppose that we would like to test (2.1) at level α for the conjectured parameter value $\gamma_0 = \gamma$ against an unrestricted alternative. Under some continuity conditions this is equivalent to the test of

$$\inf_{v\in\mathcal{V}} \theta\left(\gamma,v\right) \geq 0 \text{ against } \inf_{v\in\mathcal{V}} \theta\left(\gamma,v\right) < 0.$$

Let $\theta_0\left(\gamma\right) := \inf_{v\in\mathcal{V}} \theta\left(\gamma,v\right)$. Our method for inference delivers a statistic

$$\widehat{\theta}_\alpha(\gamma) = \inf_{v\in\mathcal{V}} \left[\widehat{\theta}\left(\gamma,v\right) + k\cdot s\left(\gamma,v\right)\right]$$

such that $\lim_{n\to\infty} P(\theta_0\left(\gamma\right) \geq \widehat{\theta}_\alpha(\gamma)) \leq \alpha$. Here, $s(\gamma,v)$ is the standard error of $\theta(\gamma,v)$ and k is a critical value, which will be described below. If $\widehat{\theta}_\alpha(\gamma) < 0$, then we reject the null hypothesis, while if $\widehat{\theta}_\alpha(\gamma) \geq 0$, then we do not reject. This provides a method for pointwise inference on γ_0. □

Informal Overview of Results. We now provide an informal description of our method for estimation and inference. Let θ^* denote the parameter of interest. Consider an upper bound θ_0 on θ^* of the form

$$\theta^* \leq \theta_0 := \inf_{v\in\mathcal{V}} \theta(v), \tag{2.2}$$

where $v \mapsto \theta(v)$ is a bound-generating function, and $\mathcal{V}$ is the set over which the minimum is taken. Likewise, there could be lower bounds defined symmetrically. Since our method covers lower bounds in an analogous way, we focus on describing our method for (2.2).

We base estimation and inference on a uniformly consistent estimator $\{\widehat{\theta}(v), \ \ v \in \mathcal{V}\}$ of the bound-generating function, which could be parametric or nonparametric.

What are good *estimators* and *confidence regions* for the bound θ_0? The first and perhaps simplest idea is to base estimation and inference on the sample analog: $\inf_{v\in\mathcal{V}} \widehat{\theta}(v)$. However, this estimator does not perform well in practice. First, the sample analog estimator tends to be downward (optimistically) biased in finite samples. Second, and perhaps more importantly, unequal sampling error of the estimator $\widehat{\theta}(v)$ across v can overwhelm inference in finite samples. Indeed, different levels of precision of $\widehat{\theta}(v)$ at different points can severely distort the perception of the minimum of the bound-generating function $\theta(v)$. Figure 1 illustrates these problems geometrically. The solid curve is the true bound-generating function $v \mapsto \theta(v)$, and the dash-dotted thick curve is its estimate $v \mapsto \widehat{\theta}(v)$. The remaining dashed curves represent eight additional potential realizations of the estimator, illustrating the precision of the estimator. In particular, we see that the precision of the estimator is much lower on the right side than on the left. A naïve sample analog estimate for θ_0 is provided by the minimum of the dash-dotted curve, but this estimate can in fact be quite far away from θ_0. This large deviation from the true value arises from both the lower precision of the estimated curve on the right side of the figure and from the downward bias created by taking the minimum of the estimated curve.

To overcome these problems, we propose a *precision-corrected* estimate of θ_0:

$$\widehat{\theta} := \min_{v\in\widehat{V}} [\widehat{\theta}(v) + k \cdot s(v)], \tag{2.3}$$

where $s(v)$ is the standard error of $\widehat{\theta}(v)$, $\widehat{V}$ is a data-dependent set that converges in probability to a non-stochastic set V that contains $V_0 := \arg\min_{v\in\mathcal{V}} \theta(v)$, and k is a critical value, whose construction we describe below. That is, our estimator $\widehat{\theta}$ minimizes the *precision-corrected curve* given by $\widehat{\theta}(v)$ plus critical value k times the pointwise standard error $s(v)$. Figure 2 shows a precision-corrected curve as a dashed curve with a particular choice of critical value k. In this figure, we see that the minimizer of the precision-corrected curve can indeed be much closer to θ_0 than the sample analog $\inf_{v\in\mathcal{V}} \widehat{\theta}(v)$. Although this illustration is schematic in nature, it conveys geometrically why our approach can remove the downward bias. In what follows, we provide both theoretical and Monte-Carlo evidence that further supports this point.

Let us now discuss the choice of the critical value k. Ideally, we would choose k in (2.3) as a quantile of the supremum of the normalized stochastic process

$$Z_n(v) := \left(\frac{\theta(v) - \widehat{\theta}(v)}{s(v)} \right), \quad v \in V \subset \mathbb{R}^d. \tag{2.4}$$

In particular, for the purpose of estimation of θ_0, we would like to set

$$k = \text{Median} \left[\sup_{v \in \widehat{V}} \frac{\theta(v) - \widehat{\theta}(v)}{s(v)} \right], \tag{2.5}$$

which gives us a downward median-unbiased estimate $\widehat{\theta}$ of θ_0. For the purpose of inference on θ_0, we would like to set

$$k = (1-\alpha)\text{-Quantile} \left[\sup_{v \in \widehat{V}} \frac{\theta(v) - \widehat{\theta}(v)}{s(v)} \right], \tag{2.6}$$

which gives us a one-sided $(1-\alpha)$ confidence region $(-\infty, \widehat{\theta}]$ for θ_0. Of course, these values of k are unknown in practice, and we have to replace them with suitable estimates.

We estimate critical values as follows. Generally, the finite-sample distribution of the process $Z_n = \{Z_n(v) : v \in \mathcal{V}\}$ is unknown, but we can approximate it uniformly by a sequence of processes with a known (or at least estimable) distribution. Indeed, we can approximate Z_n uniformly by a sequence of processes Z_n', which are zero-mean Gaussian or other pivotal processes with a known distribution, that is,

$$a_n \sup_{v \in \mathcal{V}} |Z_n(v) - Z_n'(v)| = o_p(1), \tag{2.7}$$

for some sequence of constants a_n. Once we have Z_n', we consider the variable

$$\mathcal{E}_n(V) = a_n [\sup_{v \in V} Z_n'(v) - b_n] \tag{2.8}$$

for some sequences of constants a_n and b_n. Then we obtain the estimates of the p-th quantile of $\mathcal{E}_n(V)$, denoted by $\widehat{c}(p)$, by one of two methods:

1. Simulation Method, where we simulate the Gaussian process $Z_n'(v)$ and compute its quantiles numerically.
2. Analytical Method, where we use limit quantiles or approximate quantiles of $\mathcal{E}_n(V)$, which we derive by limit arguments or Hotelling's tube method for the suprema of Gaussian processes.

Finally, we then set the critical value $k := \widehat{b}_n + \widehat{c}(p)/\widehat{a}_n$, where $\widehat{a}_n$ and $\widehat{b}_n$ consistently estimate a_n and b_n, respectively, and where $p = 1/2$ for estimation and $p = 1 - \alpha$ for inference.

At an abstract level our method does not distinguish parametric estimators of $\theta(v)$ from nonparametric estimators; however, details of the analysis and regularity conditions are quite distinct. Specifically, in section 3, we divide the analysis into Donsker and non-Donsker cases, corresponding approximately to parametric and non-parametric cases. In both cases, we employ strong approximation analysis to approximate the quantiles of $\mathcal{E}_n(V)$, and we verify our main conditions separately for each case.

An important input into our procedure is the choice of the estimator $\widehat{V}$ of V_0, the argmin set of the true bound-generating function. We describe a specific choice of such an estimator in Section 3. At a general level we require $\widehat{V}$ to be bigger than (to include) V_0, with probability approaching one; at the same time, we require this estimate not to be too much bigger than V_0.[4] The first requirement guarantees that we are not performing overly optimistic inference, and the second requirement guarantees that were are not performing overly pessimistic inference. Indeed, from (2.5) and (2.6) we see that the critical value k is decreasing in the size of the set $\widehat{V}$, so that smaller $\widehat{V}$ leads to a lower (less conservative) k. Lower k in turn leads to point estimates with a less conservative bias-correction, and less conservative confidence intervals. A good estimator $\widehat{V}$ is therefore essential. We illustrate the gains that can be made from estimating the argmin set V_0 in Figures 2 and 3. In Figure 2, we depict a precision-corrected curve (dashed curve) that adjusts the boundary estimate $\widehat{\theta}(v)$ (dotted curve) by an amount proportional to its point-wise standard error using the conservative choice $\widehat{V} = \mathcal{V} = [0, 1]$. In Figure 3, we depict the same initial precision-corrected curve and also a two-step precision-corrected curve (dash-dotted cruve) that adjusts the boundary estimate $\widehat{\theta}(v)$ (dotted curve) by an amount proportional to its point-wise standard error using a critical value that was computed using an estimate $\widehat{V}$ of V_0, which is much less conservative than using the entire set $\mathcal{V} = [0, 1]$. The gain from estimating the argmin set V_0 here is that the minimum of this precision-corrected curve is now much closer to the true minimum θ_0 of the bound-generating function $\theta(v)$ than the minimum of the initial precision-corrected curve.

[4]Of course, an ideal but infeasible choice of $\widehat{V}$ would be to simply use V_0.

3. Theory of Inference on Intersection Bounds

3.1. **Theory under High-Level Conditions.** We begin by presenting a set of simple high-level conditions, under which we demonstrate validity and general applicability of our inferential approach. In subsequent sections we verify these conditions for parametric and nonparametric estimators of the bound-generating function $\theta(v)$.

In the conditions that follow, the studentized stochastic process defined in (2.4) plays a particularly important role. Moreover, we also employ a general superset estimate $\widehat{V}$ consistent for the argmin superset V, which is a set that contains the argmin set

$$V_0 = \arg\inf_{v\in\mathcal{V}} \theta(v),$$

that is $V_0 \subseteq V$. We require that the superset estimate $\widehat{V}$ be consistent for the superset V with respect to the Hausdorff distance, i.e.

$$d_H(\widehat{V}, V) := \max\{\sup_{v\in\widehat{V}} d(v, V), \sup_{v\in V} d(v, \widehat{V})\} \to_p 0,$$

where $d(v, V) = \inf_{v'\in V} \|v - v'\|$. While it is generally desirable for the set V to be small, we shall see later that working with supersets V of the argmin set V_0, rather than with the argmin set itself, turns out to be essential in non-parametric settings.

We are now prepared to state the following conditions on the studentized stochastic process and estimators of the superset.[5]

Condition C. 1. *Let V be a superset of V_0, that is, $V_0 \subseteq V$. For some sequence of nonnegative normalizing constants a_n and b_n, we have that the normalized supremum of the studentized process $a_n \cdot (\sup_{v\in V} Z_n(v) - b_n)$ can either be* ***(a)*** *approximated in distribution by a variable $\mathcal{E}_\infty(V)$, namely*

$$a_n \cdot (\sup_{v\in V} Z_n(v) - b_n) =_d \mathcal{E}_\infty(V) + o_p(1),$$

or ***(b)*** *approximately majorized in distribution by a variable $\mathcal{E}_\infty(V)$, namely*

$$a_n \cdot (\sup_{v\in V} Z_n(v) - b_n) \leq_d \mathcal{E}_\infty(V) + o_p(1),$$

[5]The notation used in Condition C.1 is as follows: for a sequence of random variables X_n and a random variable X, we use $X_n =_d X + o_p(1)$ to denote that there exist a sequence of random variables $\tilde{X}_n$ and a random variable $\tilde{X}$ on the same probability space satisfying $X_n =_d \tilde{X}_n$ for each n, $X =_d \tilde{X}$, and $\tilde{X}_n \to_p \tilde{X}$, where $X =_d \tilde{X}$ denotes that the distribution of X is the same as that of that of $\tilde{X}$. Similarly, we use $X_n \leq_d Y_n + o_p(1)$ to mean that there exist $\tilde{X}_n$ and $\tilde{Y}_n$ on the same probability space satisfying $X_n \leq_d \tilde{X}_n$, $Y_n =_d \tilde{Y}_n$ for each n, and $\tilde{X}_n - \tilde{Y}_n \to_p 0$, where $X \leq_d \tilde{X}$ denotes that the distribution of X is first-order stochastically dominated by that of $\tilde{X}$.

where $\mathcal{E}_\infty(V)$ has a known continuous distribution function.

This is a basic asymptotic condition, which either requires standard convergence in distribution or majorization in distribution by a limit random variable with a continuous distribution function. We also consider the following generalization of C.1 which is useful for our purposes.

Condition C*. 1. *Let V be a superset of V_0, that is, $V_0 \subseteq V$. For some sequence of nonnegative normalizing constants a_n and b_n, we have that the normalized supremum of the studentized process $a_n \cdot (\sup_{v \in V} Z_n(v) - b_n)$ can either be* ***(a)*** *approximated in distribution by a variable $\mathcal{E}_n(V)$, namely*

$$a_n \cdot (\sup_{v \in V} Z_n(v) - b_n) =_d \mathcal{E}_n(V) + o_p(1),$$

or ***(b)*** *approximately majorized in distribution by a variable $\mathcal{E}_n(V)$, namely*

$$a_n \cdot (\sup_{v \in V} Z_n(v) - b_n) \leq_d \mathcal{E}_n(V) + o_p(1),$$

where $\mathcal{E}_n(V) = O_p(1)$ has a known distribution and satisfies a sequential continuity or anti-concentration property, specifically that for any sequence $\epsilon_n \searrow 0$,

$$\sup_{x \in \mathbb{R}} P[|\mathcal{E}_n(V) - x| \leq \epsilon_n] \to 0. \tag{3.1}$$

Conditions C.1 or C*.1 justify the use of quantiles of $\mathcal{E}_\infty(V)$ or $\mathcal{E}_n(V)$, respectively, for inference. Condition C.1(a) requires that the supremum of the normalized process $Z_n(v)$, appropriately studentized, converges in distribution to the random variable $\mathcal{E}_\infty(V)$. As shown in section 3, it applies with either parametric or nonparametric kernel estimation of the bound-generating function $\theta(\cdot)$. Condition C.1(b) is a weaker condition that does not require the studentized supremum of $Z_n(v)$ to have an asymptotic distribution, but only requires that its distribution can be majorized by that of $\mathcal{E}_\infty(V)$. Section 3.4 establishes its validity for nonparametric series estimation of the bound-generating function. Note that by the term "known distribution," in reference to $\mathcal{E}_\infty(V_0)$ and $\mathcal{E}_n(V)$, we mean a distribution whose parameters can be estimated consistently. Also, instead of using standard convergence in distribution notation, we employ strong approximation, which is without loss of generality relative to the former due to the Skorohod-Dudley-Wichura construction. In general, the normalizing constants a_n and b_n may depend on V and can be different depending on which of C.1(a) and C.1(b) hold.

Condition C*.1 is a generalization of C.1, which allows for the use of some intermediate or penultimate approximations for inference. For example, in the case of series approximation we can approximate the supremum of the process $Z_n(v)$ by the supremum $\mathcal{E}_n(V)$ of a Gausssian process, which does not in general converge to a fixed random variable, but can instead be majorized in distribution by an exponential random variable $\mathcal{E}_\infty(V)$. However, this majorization can be conservative. We can instead use the quantiles of $\mathcal{E}_n(V)$ for inference, which in our experience provides a more accurate, less conservative approximation. In order for the penultimate approach to be valid, we require the sequential continuity, or anti-concentration, property (3.1) for the sequence of random variables $\mathcal{E}_n(V)$. This property is needed for the disappearance of the effect of approximation errors in critical values on the coverage probabilities. If $\mathcal{E}_n(V)$ has a continuous limit distribution the anti-concentration property follows automatically. If $\mathcal{E}_n(V)$ does not have a limit distribution, verification of this property is a harder problem, which can be achieved either numerically or, in some limited cases, analytically using exact versions of Hotelling's tubing method. Analytical limitations arise because little is known about the anti-concentration properties of the suprema of a sequence of Gaussian processes, in contrast to a vast knowledge on the concentration properties of such processes (see however Rudelson and Vershynin (2007), Rudelson and Vershynin (2008) and Tao and Vu (2009) for a discussion of anti-concentration inequalities for some "simpler" related problems).

The next condition deals with the effect of estimating the approximate argmin sets.

Condition C. 2. *Let $\widehat{V}$ denote any sequence of sets, possibly data-dependent, that contain a superset V of V_0, with probability approaching one, and that converge to V at the rate r_n, i.e., $d_H(\widehat{V}, V) \leq O_p(r_n)$, where r_n is a sequence of constants converging to zero. Also, let $\widehat{a}_n$ and $\widehat{b}_n$ denote corresponding, possibly data-dependent, normalizing constants. Then the normalized supremum of the studentized stochastic process is insensitive to the replacement of the superset V and normalizing constants (a_n, b_n) with the estimates $\widehat{V}$ and $(\widehat{a}_n, \widehat{b}_n)$, namely*

$$\widehat{a}_n \cdot (\sup_{v \in \widehat{V}} Z_n(v) - \widehat{b}_n) - a_n \cdot (\sup_{v \in V} Z_n(v) - b_n) \to_p 0.$$

This assumption allows for a data-dependent choice of $\widehat{V}$, but requires that $\widehat{V}$ should eventually settle down at V, without affecting the supremum of the studentized stochastic process. In Section 3.6, we construct such estimators from the level sets of the estimated bound-generating function $v \mapsto \widehat{\theta}(v)$ and show that these estimators converge

to the level sets of $v \mapsto \theta(v)$ at a rate sufficiently fast not to affect the behavior of the supremum of the estimated process. In nonparametric settings, this may sometimes require that level sets are strictly larger than the argmin set V_0.

We now state our first main result under the above conditions.

Theorem 1 (**Main Result Under C.1-C.2.**). *Let*

$$\widehat{\theta}_p = \inf_{v \in \widehat{V}} [\widehat{\theta}(v) + [\widehat{b}_n + \widehat{c}(p)/\widehat{a}_n] s(v)],$$

where $\widehat{c}(p)$ is defined below.

1. Suppose that conditions C.1(b) or C.1(b) and C.2 hold and that $\widehat{c}(p)$ is a consistent upper bound on $c_n(p) :=$ the $p-th$ quantile of $\mathcal{E}_n(V)$, where $n = \infty$ under C.1(b), namely*

$$\widehat{c}(p) \geq c_n(p) + o_p(1).$$

Then we have that the estimator $\widehat{\theta}_p$ is downward p-quantile unbiased, namely

$$\liminf_{n \to \infty} P[\theta_0 \leq \widehat{\theta}_p] \geq p.$$

2. Suppose conditions C.1(a) or C.1(a) and C.2 hold with $V = V_0$ and that $\widehat{c}(p)$ is a consistent estimate of $c_n(p) :=$ p-th quantile of $\mathcal{E}_n(V_0)$, where $n = \infty$ under C.1(a), namely*

$$\widehat{c}(p) = c_n(p) + o_p(1).$$

Then we have that the estimator $\widehat{\theta}_p$ is p-quantile unbiased, namely

$$\lim_{n \to \infty} P[\theta_0 \leq \widehat{\theta}_p] = p.$$

Thus, the quantity $\widehat{\theta}_p$ can be used to provide a one-sided confidence interval for θ_0, since $\lim_{n\to\infty} P[\theta_0 \leq \widehat{\theta}_p] \geq p$, with equality under C.1(a) or C*.1(a). Moreover, $\widehat{\theta}_{1/2}$ is a median downward-unbiased estimator for θ_0 in the sense that

$$\lim_{n \to \infty} P[\theta_0 \leq \widehat{\theta}_{1/2}] \geq \frac{1}{2}.$$

In words, the asymptotic probability that the estimator $\widehat{\theta}_{1/2}$ lies above the true θ_0 is at least a half.

3.2. **Donsker and Non-Donsker Cases.** We specialize the high-level conditions developed above into two general cases:

(1) The Donsker case, where the studentized process converges to a fixed continuous Gaussian process. This immediately implies the convergence of suprema as well as

the insensitivity of the supremum to replacement of the argmin sets with consistent estimates. This case primarily covers parametric estimation and includes a great variety of practical procedures, including the "finite-support case", where $\mathcal{V}$ is a finite set.

(2) The non-Donsker case, where the studentized process does not converge to a fixed continuous Gaussian process, but may instead be approximated by a sequence of Gaussian processes or other pivotal processes. This case is harder, but it also leads to a majorization of the supremum by tractable random variables as well as insensitivity to replacement of the argmin supersets with consistent estimates. This case primarily covers nonparametric estimation of the boundary and includes a rich variety of procedures, ranging from kernel to series methods.

Formally we define the Donsker case as follows.

Condition D. 1. *The normalized stochastic process Z_n converges to a continuous Gaussian process Z_∞ with a known distribution and a non-degenerate covariance function, in the space of bounded functions on $\mathcal{V}$, namely*

$$Z_n(\cdot) =_d Z_\infty(\cdot) + o_p(1), \quad \text{in } \ell^\infty(\mathcal{V}).$$

It is worth noting here that given weak convergence, convergence in probability is without loss of generality due to the Skorohod-Dudley-Wichura construction. According to the latter, given weak convergence, we can always find a suitably enriched probability space on which convergence in probability takes place.

The Donsker condition is widely applicable in parametric and semi-parametric estimation problems. It leads to immediate verification of the high-level conditions C.1 and C.2.

Lemma 1. *The Donsker condition D.1 implies conditions C.1 (a) with normalizing constants $a_n = \widehat{a}_n = 1$ and $b_n = \widehat{b}_n = 0$ and the limit variable $\mathcal{E}_\infty(V_0) = \sup_{v \in V_0} Z_\infty(v)$ with a continuous distribution and condition C.2, including the ideal case $V = V_0$, with any vanishing sequence of positive constants $r_n = o(1)$.*

Next, we formally define the non-Donsker cases as follows.

Condition N. *The normalized stochastic process Z_n can be approximated uniformly by a sequence of penultimate processes Z'_n, which is a sequence of either Gaussian processes or some other pivotal processes Z'_n, with a known distribution, namely*

$$a_n \sup_{v \in \mathcal{V}} |Z_n(v) - Z'_n(v)| = o_p(1),$$

for some sequence of constants a_n. Conditions C.1 and C.2 or C.1 and C.2 hold with $Z_n(v)$ replaced by the sequence of penultimate processes $Z'_n(v)$. The resulting conditions are referred to as Conditions* ***N.1****,* ***N*.1*** *and* ***N.2****, respectively.*

This condition requires the studentized stochastic process to be approximated by a sequence of pivotal processes, whose behavior is sufficiently regular to allow verification of the high-level conditions C.1 and C.2. Below, we show how this condition is fulfilled for series and kernel estimators.

Lemma 2. *Condition N implies C.1 (or C*.1) and C.2.*

3.3. **Parametric Estimation of $v \mapsto \theta(v)$.** In this subsection, we show that the conditions developed above apply to various parametric estimation methods of $v \mapsto \theta(v)$. Parametric estimation is an important practical case, and it turns out to be quite tractable. In particular, it includes the case where the set $\mathcal{V}$ is finite. We formally state the conditions required for parametric estimation in the following:

Condition P. *$\theta(v) = \theta(v, \gamma_0)$, where $\theta(v, \gamma)$ is a known function of finite-dimensional parameter $\gamma \in \mathbb{R}^k$, and $\partial\theta(v, \gamma)/\partial\gamma$ is uniformly continuous in (γ, v) for all γ in a neighborhood of $\gamma_0, v \in \mathcal{V}$.*

P.1 *An estimate $\widehat{\gamma}$ is available such that*

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) =_d \Omega^{1/2}\mathcal{N} + o_p(1), \quad \mathcal{N} =_d N(0, I),$$

where for

$$g(v)' = \frac{\partial\theta(v, \gamma_0)}{\partial\gamma}'\Omega^{1/2}$$

the norm $\|g(v)\|$ is bounded uniformly in v above and away from zero.

P.2 *There is an estimator for the standard deviation of $\theta(v, \widehat{\gamma})$ that satisfies*

$$s(v) = \frac{1}{\sqrt{n}}\|g(v)\|(1 + o_p(1)),$$

uniformly in $v \in \mathcal{V}$. For example, if there is an estimate $\widehat{\Omega}$ such that $\widehat{\Omega} = \Omega + o_p(1)$, then such an estimate of precision is given by $s(v) = \|\hat{g}(v)\|/\sqrt{n}$ with $\hat{g}(v)' = \frac{\partial\theta(v,\widehat{\gamma})}{\partial\gamma}'\widehat{\Omega}^{1/2}$.

For the case of a finite number of support points, one can set $\theta(v) = \sum_{j=1}^{J}\gamma_j 1(v = v_j)$, where $(v_1, \ldots, v_J)$ are the support points and $1(\cdot)$ is the usual indicator function. The following lemma shows that Condition D follows under the conditions stated above.

Lemma 3. *Condition P implies Condition D with the limit process*

$$Z_\infty(v) = \frac{g(v)'\mathcal{N}}{\|g(v)\|}, \quad g(v)' = \frac{\partial\theta(v,\gamma_0)}{\partial\gamma}'\Omega^{1/2}.$$

3.4. **Nonparametric Estimation of $\theta(v)$ via Series.** Series estimation is effectively like parametric estimation, but the dimension of the estimated parameter tends to infinity and bias arises due to approximation based on a finite number of basis functions. If we select the number of terms in the series expansion so that the estimation error is of larger magnitude than the approximation error, then the analysis closely mimics that of the parametric case.

Condition S. *Suppose that the series estimator $\widehat{\theta}(v)$ for the function $\theta(v)$ has the form*

$$\widehat{\theta}(v) = p(v)'\widehat{\beta},$$

where $p_n(v) := (p_1(v), \ldots, p_K(v))$ is a collection of K-dimensional approximating functions, $K \to \infty$, $K = o(n)$, and $\widehat{\beta}$ is a K-vector of series regression estimates. Furthermore, assume the following conditions hold.

S.1 *The estimator satisfies the following linearization and strong approximation condition in $\ell^\infty(\mathcal{V})$*

$$\frac{\sqrt{n}(\widehat{\theta}(v) - \theta(v))}{\|g_n(v)\|} =_d \frac{g_n(v)'\mathcal{N}_n}{\|g_n(v)\|} + R_n(v),$$

where

$$g_n(v)' = p_n(v)'\Omega_n^{1/2}, \ \mathcal{N}_n =_d N(0, I_K), \quad \sup_{v\in\mathcal{V}} |R_n(v)| = o_p(1/\sqrt{\log n}),$$

where Ω_n are positive definite matrices, and $\|\nabla_v g_n(v)/\|g_n(v)\|\|$ is of polynomial growth in K uniformly in $v \in \mathcal{V}$.

S.2 *There exists an estimate $s(v)$ of precision such that*

$$s(v) = \frac{\|g_n(v)\|}{\sqrt{n}}(1 + o_p(1)),$$

uniformly in $v \in \mathcal{V}$. For example, if there is an estimate $\widehat{\Omega}$ such that $\|\widehat{\Omega} - \Omega\| = o_p(1)$, then such estimate of precision is given by $s(v) = \|\hat{g}_n(v)\|/\sqrt{n}$ with $\hat{g}_n(v)' = p_n(v)'\widehat{\Omega}^{1/2}$.

Assumption S.1 embeds a number of requirements. The first is that the series estimator admits a linear approximation in terms of a zero-mean vector $\tilde{\mathcal{N}} \sim (0, I)$, which

is typically a rescaled sum of vectors. The second is undersmoothing, namely that the approximation bias is asymptotically negligible. The third is the approximation of the vector $\tilde{\mathcal{N}}$ by a normal vector $\mathcal{N} =_d N(0, I)$. This approximation is immediate, for example, in series regression with normal errors, but it also applies considerably more generally. Indeed, using the coupling of Yurinskii (1977), we provide sufficient primitive conditions for this approximation in Appendix B.1.

Lemma 4. *Assume that mes$(V) > 0$, where mes(V) denotes the Lebesgue measure of V. Condition S implies condition N.1*(b) with the penultimate process*

$$Z_n'(v) = \alpha_n(v)'\mathcal{N}_n,\ \mathcal{N}_n = N(0, I_K),\ \alpha_n(v) = \frac{g_n(v)}{\|g_n(v)\|},$$
$$\mathcal{E}_n(V) = a_n[\sup_{v\in V} Z_n'(v) - b_n],\ a_n = b_n \sim \sqrt{2d\log(2L/\sqrt{2\pi})},$$
$$L = \sup_{v\in V} \|\nabla\alpha_n(v)\| \cdot \ diam(V).$$

When $d = 1$, we can also use a sharper constant $a_n = \sqrt{2\log\frac{\kappa_n(V)}{2\pi}}$ where $\kappa_n(V) = \int_V \|\nabla\alpha_n(v)\|dv$. Furthermore, condition S implies condition N.1(b), as the sequence of random variables $\mathcal{E}_n(V)$ is stochastically dominated in distribution by the standard exponential random variable

$$\mathcal{E}_n(V) \leq_d \mathcal{E}_\infty + o_p(1), \quad P[\mathcal{E}_\infty > p] = \exp(-p).$$

Lemma 4 provides a majorizing limiting variable $\mathcal{E}_\infty$ for the normalized supremum of the studentized empirical process Z_n. It also provides a penultimate approximation $\mathcal{E}_n(V)$ for this supremum. We can use these results for construction of critical values. The p-th quantile of $\mathcal{E}_\infty(V)$ is given by

$$c_\infty(p) = -\log(1-p).$$

Therefore, we can set

$$k_{1-\alpha} = a_n(V) + \frac{c_\infty(1-\alpha)}{a_n(V)}. \tag{3.2}$$

Alternatively, we can base inference on quantiles of $\mathcal{E}_n(V)$ and estimate them numerically. We describe the practical details of simulation of critical values in Appendix C. It is not restrictive to assume that V has strictly positive measure. Even if V_0 is singleton, we can select V to be a superset of V_0 of positive measure, in which case our method for inference is valid but conservative.

Lemma 5. *Assume that mes*$(V) > 0$. *Let* $a_n = b_n$ *and* $\widehat{a}_n = \widehat{b}_n := a_n(\widehat{V})$. *Then, condition N.2 holds if* $a_n(\widehat{V})^2 - a_n(V)^2 \to_p 0$ *and the following growth condition holds*

$$a_n \cdot r_n \cdot \sup_{v \in \mathcal{V}} \|\nabla \alpha_n(v)\| \to 0. \tag{3.3}$$

The requirement that $a_n(\widehat{V})^2 - a_n(V)^2 \to_p 0$ is a weak assumption. For example, consider $a_n = \sqrt{2 \log \frac{\kappa_n(V)}{2\pi}}$ in one-dimensional settings. In this case, $a_n(\widehat{V})^2 - a_n(V)^2 \to_p 0$ is satisfied if $\log \frac{\kappa_n(\widehat{V})}{\kappa_n(V)} \to_p 0$. If $1 \lesssim \kappa_n(V)$, which is the case when V has non-zero Lebesgue measure, then $|\frac{\kappa_n(\widehat{V})}{\kappa_n(V)} - 1| \lesssim r_n \cdot \sup_{v \in \mathcal{V}} \|\nabla \alpha_n(v)\| \to 0$. For typical series the upper bound on $\|\nabla \alpha_n(v)\|$ is of order $\sqrt{K}$. If also $r_n = (\log n)^c (K/n)^{1/2\rho}$ for some $c > 0$, then the growth condition (3.3) reduces to

$$(\log n)^{c+1/2} (K/n)^{1/2\rho} K^{1/2} \to 0,$$

When the parameter $\rho = 1$, this amounts to a rather mild condition $K^2 (\log n)^{c'}/n \to 0$, for some $c' > 0$, on growth on the number of series terms. The value $\rho = 1$ is plausible when the superset V is the ϵ-argmin of the bound-generating function for some $\epsilon > 0$, as we discuss in Section 3.6.

3.5. **Nonparametric Estimation of $\theta(v)$ via local methods.** In this section we provide conditions under which a kernel-type estimator of the bound-generating function satisfies Conditions N.1 and N.2 and we also describe how to obtain critical values k. Kernel-type estimators include standard kernel estimators as well as local polynomial estimators.

For any positive integer d and a d-dimensional vector $u = (u_1, \ldots, u_d)$, let $\mathbf{K}(u) = \prod_{i=1}^d K(u_i)$, where K is a kernel function on $\mathbb{R}$. We assume that a kernel-type estimator $\widehat{\theta}(v)$ of a bound-generating function $\theta(v)$ satisfies the following conditions. These conditions cover local estimation of bound-generating functions defined as conditional expectation functions, in which case given i.i.d. random variables (Y_i, V_i, U_i) we have $\theta(v) = E[Y_i | V_i = v]$, and $\sigma^2(v) = \mathrm{Var}(Y_i | V_i = v)$ in the expression given below. These conditions also cover local estimation of bound-generating functions defined as conditional quantile functions, although in this case the underlying interpretation of parameters differs.

Condition K. 1. *Assume that the estimator satisfies the following linearization and strong approximation condition in $\ell^\infty(\mathcal{V})$:*

$$\frac{(nh_n^d)^{1/2}[\widehat{\theta}(v) - \theta(v)]}{\|w_n(v)\|} =_d \frac{w_n(v)'\mathbf{U_n}}{\|w_n(v)\|} + R_n(v),$$

where

$$\mathbf{U_n} =_d N_n(0, I) \text{ conditional on } (V_1, \ldots, V_n), \quad \sup_{v\in\mathcal{V}} |R_n(v)| = o_p(a_n^{-1}),$$

$w_n(v)$ is typically an n-dimensional vector of the form

$$w_n(v) = \left(\frac{\sigma(V_i)\mathbf{K}\left[h_n^{-1}(v - V_i)\right]}{(nh_n^d)^{1/2} f_V(v)}, i = 1, \ldots, n \right), \tag{3.4}$$

K is a kernel function that is bounded and is continuously differentiable with a bounded derivative, h_n is a bandwidth that satisfies $h_n \to 0$ and $\log n/(nh_n^d)^{1/2} \to 0$, $\sigma^2(v)$ is uniformly continuous, bounded and also bounded below from zero, $f_V(v)$ is the probability density function for V_i, which is bounded away from zero and has a bounded derivative, and $N_n(0, I)$ denotes the n-dimensional multivariate normal distribution with variance the identity matrix, and V_i are i.i.d.

Condition K. 2. *There exists an estimate of precision such that*

$$s(v) = \frac{\|w_n(v)\|}{\sqrt{nh_n^d}}(1 + o_p(1)),$$

uniformly in $v \in \mathcal{V}$. For example, a consistent estimate of precision is given by $s(v) = \|\hat{w}_n(v)\|/\sqrt{nh_n^d}$, where

$$\hat{w}_n(v) = \left(\frac{\hat{\sigma}(V_i)\mathbf{K}\left[h_n^{-1}(v - V_i)\right]}{(nh_n^d)^{1/2} \hat{f}_V(v)}, i = 1, \ldots, n \right), \tag{3.5}$$

where $\sup_{v\in\mathcal{V}} |\hat{\sigma}(v) - \sigma(v)| = o_p(1)$ and $\sup_{v\in\mathcal{V}} |\hat{f}_V(v) - f_V(v)| = o_p(1)$.

Conditions K.1 and K.2 embed a number of requirements. As was the case for series estimators, a simple immediate case is nonparametric mean regression with normal errors that are mean independent of regressors with known conditional variance $\sigma^2(v)$. It is not difficult to extend conditions K.1 and K.2 to more general cases with non-normal errors, an unknown conditional variance function, and additional covariates other than v. In Appendix B.2, we give sufficient conditions for strong approximation of kernel-type estimators of conditional expectation functions.

In order to provide an analytic approximation for the asymptotic distribution of the supremum of the studentized estimation process Z_n, let $\rho_d(s) = \prod_{j=1}^{d} \rho(s_j)$, where $s \equiv (s_1, \ldots, s_d)$ is a d-dimensional vector and

$$\rho(s_j) = \frac{\int K(u)K(u - s_j)du}{\int K^2(u)du} \tag{3.6}$$

for each j.

Lemma 6. *Let $a_n(V) = b_n(V)$ be the largest solution to the following equation:*

$$mes(V){h_n}^{-d}\lambda^{d/2}(2\pi)^{-(d+1)/2}a_n^{d-1}\exp(-a_n^2/2) = 1, \tag{3.7}$$

where

$$\lambda = \frac{-\int K(u)K''(u)du}{\int K^2(u)du}.$$

Assume that K.1 and K.2 hold and $mes(V) > 0$. Then, condition N.1(a) holds with the penultimate process

$$Z_n'(v) := \frac{w_n(v)'\mathbf{U_n}}{\|w_n(v)\|}.$$

Furthermore, we have that

$$Z_n'(v) =_d Z_n''\left(h_n^{-1}v\right) + R_n'(v), \quad \sup_{v\in V}|R_n'(v)| = o_p(a_n(V)^{-1}).$$

where $\{Z_n''(h_n^{-1}v) : v \in V\}$ is a sequence of Gaussian processes with continuous sample paths such that

$$E[Z_n''(s)] = 0,$$
$$E[Z_n''(s_1)Z_n''(s_2)] = \rho_d(s_1 - s_2) \ \ \textit{for } s, s_1, s_2 \in \mathcal{V}_n := h_n^{-1}V,$$

Finally, we have that

$$\mathcal{E}_n(V) := a_n(V)\left[\sup_{v\in V} Z_n''\left(h_n^{-1}v\right) - a_n(V)\right] =_d \mathcal{E}_\infty + o_p(1), \tag{3.8}$$

where $\mathcal{E}_\infty$ has the type I extreme-value distribution.

Lemma 6 provides a majorizing limiting variable $\mathcal{E}_\infty$ for the normalized supremum of the studentized empirical process Z_n. It also provides a penultimate approximation $\mathcal{E}_n(V)$ for this supremum. We can use these results for construction of critical values.

For example, when $d = 1$ and $V = [a, b]$,

$$a_n(V) = \left(2 \log(h_n^{-1}(b-a)) + 2 \log \frac{\lambda^{1/2}}{2\pi} \right)^{1/2}. \tag{3.9}$$

The $1-\alpha$ quantiles of $\mathcal{E}_\infty$ is given by

$$c_\infty(1-\alpha) = -\log\log(1-\alpha)^{-1}$$

Then we set

$$k_{1-\alpha} = a_n(V) + \frac{c_\infty(1-\alpha)}{a_n(V)}, \tag{3.10}$$

which consistently estimates the $1-\alpha$ quantile of $\mathcal{E}_\infty(V)$. Alternatively, we can base inference on quantiles of $\mathcal{E}_n(V)$ and estimate them numerically. We describe the practical details for the simulation of critical values in Appendix C. Note that it is not restrictive to assume that V has strictly positive measure. Even if V_0 is singleton, we can select V to be a superset of V_0 of positive measure, in which case our method for inference is valid but conservative.

It is possible to construct an asymptotically valid, alternative critical value. Equation (A.6) in the proof of Theorem 6 suggests that we might construct an alternative critical value by using the leading term in equation (A.6). In other words, instead of using quantiles of $\mathcal{E}_\infty(V)$, we can use quantiles of the following distribution-like function:

$$\exp\left\{ -\exp\left(-x - \frac{x^2}{2a_n^2} \right) \left[1 + \frac{x}{a_n^2} \right]^{d-1} \right\}.$$

For example, when $d = 1$ and $V = [a, b]$, instead of using (3.10), we can use the alternative critical value:

$$k'_{1-\alpha} = (a_n(V)^2 - 2\log\log(1-\alpha)^{-1})^{1/2}. \tag{3.11}$$

In some contexts this approximation may behave better in finite samples than an approximation using the extreme value distribution (see, e.g. Piterbarg (1996) and Lee, Linton, and Whang (2009)). In addition, we can consider the following form:

$$k_{1-\alpha} = b_n(V) + \frac{c_\infty(1-\alpha)}{a_n(V)}, \tag{3.12}$$

where $V = [a, b]$, $a_n(V) = \sqrt{2\log(h_n^{-1}(b-a))}$, and $b_n(V) = a_n(V) + \log\sqrt{(\lambda/2\pi)/a_n(V)}$. The critical value in (3.12) is a one-sided version of equation (29) of Härdle and Linton (1994), which seems to behave better in finite samples, compared to (3.10). The critical values in (3.10), (3.11), and (3.12) are asymptotically equivalent.

The following lemma provides sufficient conditions for Condition N.2.

Lemma 7. *Assume that mes$(V) > 0$. Let $\alpha_n(v) := w_n(v)/\|w_n(v)\|$. Also, let $a_n = b_n$ and $\widehat{a}_n = \widehat{b}_n := a_n(\widehat{V})$. Then, condition N.2 holds if $a_n(\widehat{V})^2 - a_n(V)^2 \to_p 0$ and the following growth condition holds:*

$$a_n(V) \cdot r_n \cdot \sup_{v \in \mathcal{V}} \|\nabla \alpha_n(v)\| \to_p 0. \tag{3.13}$$

Furthermore, under Condition K, $\sup_{v\in\mathcal{V}} \|\nabla \alpha_n(v)\| = O_p(h_n^{-1})$.

As was the case for series estimation, the requirement that $a_n(\widehat{V})^2 - a_n(V)^2 \to_p 0$ is a weak assumption. For example, consider $a_n(V)$ in (3.9). In this case, $a_n(\widehat{V})^2 - a_n(V)^2 \to_p 0$ is satisfied if $\text{mes}(V) > 0$ and $\text{mes}(\widehat{V})/\text{mes}(V) \to_p 0$. In Theorem 2 below we state sufficient conditions for this. If $r_n = (\log n)^c (nh_n^d)^{-1/2\rho}$, then the growth condition (3.13) holds if

$$(\log n)^{c+1/2} (nh_n^{d+2\rho})^{-1/2\rho} \to 0.$$

When the parameter $\rho = 1$, this amounts to a rather mild condition $nh_n^{d+2}/(\log n)^{c'} \to \infty$, for some $c' > 0$, on the growth of bandwidth.

3.6. **Estimation of V.** Next we consider the choice and estimation of V, which we choose to be the $\epsilon-$argmin of the function $\theta(v)$. In parametric cases, we can take $\epsilon = 0$, that is $V = V_0$. In nonparametric cases, it may not always be feasible to take $\epsilon = 0$ and attain both conditions C.1 and C.2. The reason is that the degree of identifiability of V_0 is decreasing in the number of smooth derivatives that the bound-generating function $\theta(v)$ has on the boundary of V_0, while the rate of convergence of $\widehat{\theta}(v) - \theta(v)$ is increasing in this number. These two effects work to offset each other. However, we can use $V = V_\epsilon$, the $\epsilon-$ argmin, whose degree of identifiability for $\epsilon > 0$, under some reasonable conditions, does not depend on the number of smooth derivatives.

Condition V. *There are two parts:*

V.1 *The estimator $\widehat{\theta}(v)$ satisfies*

$$\sup_{v\in\mathcal{V}} |\widehat{\theta}(v) - \theta(v)|/s(v) = O_p(c_n), \text{ where } c_n \gtrsim 1,$$

for example, $c_n = a_n^{-1} + b_n$ under the conditions C.1 and C.2. Also

$$\ell_n := 2\sqrt{\log n} \cdot \sup_{v\in\mathcal{V}} s(v)$$

satisfies $\gamma_n := \ell_n \cdot c_n \to 0$.

V.2 *The function $\theta(v)$ is separated away from $\theta_0 + \epsilon$ on the complement of the set*

$$V_\epsilon := \ \epsilon\text{-argmin of } \theta(v) \ = \{v \in \mathcal{V} : \theta(v) \leq \theta_0 + \epsilon\}$$

by a polynomial minorant in the distance from this set , namely

$$\theta(v) - \theta_0 - \epsilon \geq (cd(v, V_\epsilon))^{\rho(\epsilon)} \wedge \delta$$

for any $v_\epsilon \notin V_\epsilon$ for some positive constant $\rho(\epsilon)$, called the degree of identifiability, and constant c and δ, possibly dependent on ϵ, where

$$d(v, V_\epsilon) := \inf_{v' \in V_\epsilon} \|v - v'\|.$$

We propose the following estimator of V_ϵ:

$$\widehat{V}_\epsilon = \{v \in \mathcal{V} : \widehat{\theta}(v) \leq \inf_{v \in \mathcal{V}} \widehat{\theta}(v) + \ell_n c_n + \epsilon\}. \tag{3.14}$$

Theorem 2. *Suppose that conditions V.1 and V.2 hold. Then with probability converging to one, the set V_ϵ is a subset of the estimator $\widehat{V}_\epsilon$. Moreover, the Hausdorff distance between these two sets approaches zero at the following rate:*

$$d_H(\widehat{V}_\epsilon, V_\epsilon) \lesssim_p r_n = \gamma_n^{1/\rho(\epsilon)}.$$

Moreover, the Lebesgue measure of the difference between the two sets approaches zero at the following rate:

$$mes(\widehat{V}_\epsilon \setminus V_\epsilon) \lesssim_p r_n = \gamma_n^{d/\rho(\epsilon)},$$

where d is the dimension of the Euclidean space containing $\mathcal{V}$.

Thus the rate of convergence depends on the uniform rate of convergence γ_n of $v \mapsto \widehat{\theta}(v)$ to $v \mapsto \theta(v)$ and on the degree of identifiability $\rho(\epsilon)$ of the ϵ-argmin set V_ϵ.

The following lemma presents a case where condition V.2 holds under reasonable conditions and the degree of identifiability $\rho(\epsilon)$ is one.

Lemma 8. *Let $\epsilon \geq 0$ be fixed, and suppose that $\mathcal{V}$ is a convex body in $\mathbb{R}^d$ and V_ϵ is in the interior of $\mathcal{V}$. Suppose that there exists a function $\eta(\cdot)$ such that*

$$\theta(v) = \max(\eta(v), \theta_0),$$

where $\eta : \mathcal{V} \mapsto \mathbb{R}$ is continuously differentiable on $\mathcal{V}$ with $\|\nabla\eta(v)\|$ bounded away from zero on

$$\partial V_\epsilon := \{v \in \mathcal{V} : \theta(v) - \theta_0 = \eta(v) - \theta_0 = \epsilon\}.$$

Then condition V.2 holds with

$$\rho(\epsilon) = 1, c = \inf_{v \in \partial V_\epsilon} \|\nabla \eta(v)\|/2 > 0, \text{ and } \delta = \inf_{d(v, V_\epsilon) \geq d_0} (\eta(v) - \theta_0 - \epsilon) > 0$$

for some $d_0 > 0$.

3.7. **Inference on the identified set Θ_I and on the true parameter value θ^*.** We can use our one-sided confidence bands for lower and upper bounds to construct confidence intervals for the identified set, as well as for the true parameter θ^*. As in the introduction, we suppose that the identified set is of the form $\Theta_I = \left[\theta_0^l, \theta_0^u\right]$, where $\theta_0^l = \sup_{v \in \mathcal{V}^l} \theta^l(v)$ is the maximum of a collection of lower bounds, and $\theta_0^u = \sup_{v \in \mathcal{V}^u} \theta^u(v)$ the minimum of a collection of upper bounds on parameter θ^*. So far, we have described how to consistently estimate such bounds, as well as how to construct one-sided confidence bands. We now describe how these one-sided confidence bands can be used to construct two-sided bands for either Θ_I or θ^*.

We can construct two-sided bands for the identified set Θ_I as follows. Let $\widehat{\theta}_p^l$ and $\widehat{\theta}_p^u$ denote the end-points of one-sided bands so that

$$P\left(\theta_0^u \leq \widehat{\theta}_p^u\right) \geq p + o(1) \text{ and } P\left(\theta_0^l \geq \widehat{\theta}_p^l\right) \geq p + o(1).$$

Then, by Bonferroni's inequality, the region $[\widehat{\theta}_p^l, \widehat{\theta}_p^u]$ with $p = 1 - \alpha/2$ is an asymptotically valid $1 - \alpha$ confidence interval for Θ_I,

$$P\left([\theta_0^l, \theta_0^u] \subseteq [\widehat{\theta}_p^l, \widehat{\theta}_p^u]\right) \geq 1 - P\left(\theta_0^l < \widehat{\theta}_p^l\right) - P\left(\theta_0^u > \widehat{\theta}_p^u\right) \geq 1 - \alpha + o(1). \tag{3.15}$$

We can construct two-sided bands for the true parameter value θ^* as follows: Let $\widehat{\Delta}_n^+ \equiv \widehat{\Delta}_n 1[\widehat{\Delta}_n > 0]$, where $\widehat{\Delta}_n = \widehat{\theta}_{1/2}^u - \widehat{\theta}_{1/2}^l$, and $\widehat{p}_n \equiv 1 - \Phi(\tau_n \widehat{\Delta}_n^+)\alpha$, where $\Phi(\cdot)$ is the standard normal CDF, τ_n is a sequence of constants satisfying $\tau_n \to \infty$ and $\tau_n |\widehat{\Delta}_n^+ - \Delta_n| \to_p 0$, where $\Delta_n = \theta_0^u - \theta_0^l$. Notice that since $1/2 \leq \Phi(c) \leq 1$ for $c \geq 0$, we have that $\widehat{p}_n \in [1 - \alpha, 1 - \alpha/2]$. Then under conditions similar to stated below we have:

$$\inf_{\theta^* \in [\theta_0^l, \theta_0^u]} P\left(\theta^* \in [\theta_{\widehat{p}_n}^l, \theta_{\widehat{p}_n}^u]\right) \geq 1 - \alpha + o(1). \tag{3.16}$$

We note that the confidence intervals are valid uniformly with respect to the location of the true parameter value θ^* within the bounds. Moreover, this statement allows the model and thus also the width of the identification regions Δ_n to change with the sample size. Thus these confidence intervals are also valid uniformly with respect to Δ_n.

Before stating the formal result, some further notation is required. In what follows we shall use the additional superscripts $j = u$ (for upper bound) or $j = l$ (for lower bounds)

relative to the main text. Thus, all statistics, estimators, and sets receive such indices; moreover, we define the studentized empirical processes as follows

$$Z_n^u(v) = \frac{\theta^u(v) - \hat{\theta}^u(v)}{s^u(v)},\ Z_n^l(v) = \frac{\hat{\theta}^l(v) - \theta^l(v)}{s^l(v)},$$

where the second expression has the sign reversed.

The following theorem provides a formal statement of the validity of our proposed confidence intervals for θ^*.

Theorem 3. *Consider a sequence of models indexed by n such that the following conditions hold. Assume C*.1(b) holds for each $j \in \{u, l\}$, so that $a_n^j \cdot (\sup_{v \in V^j} Z_n^j(v) - b_n^j) \leq_d \mathcal{E}_n^j(V^j) + o_p(1)$, where each $\mathcal{E}_n^j(V^j) = O_p(1)$ has a known distribution and satisfies the stated anti-concentration property. Assume that C.2 holds so that for each $j \in \{u, l\}$, $\widehat{a}_n^j \cdot (\sup_{v \in \widehat{V}^j} Z_n^j(v) - \widehat{b}_n^j) - a_n^j \cdot (\sup_{v \in V^j} Z_n^j(v) - b_n^j) \to_p 0$. Further suppose that $\widehat{c}^j(p)$ is a consistent upper bound on $c_n^j(p) :=$ the $p-th$ quantile of $\mathcal{E}_n^j(V^j)$, where $n = \infty$ under C.1(b), namely for each j, $\widehat{c}^j(p) \geq c_n^j(p) + o_p(1)$. Let τ_n be a sequence of positive constants such that (A.7) holds. Then if $\Delta_n \geq 0$, (3.16) holds.*

Regarding the choice of τ_n, we note that since $|\widehat{\Delta}_n^+ - \Delta_n| \to_p 0$ typically at a polynomial rate in n, there are many admissible choices of τ_n, for example $\tau_n = \log n$. In practice it may be desirable to use a different choice, for example, $\tau_n = \sigma_n^{-1}/\log n$, where σ_n is a standardizing sequence for $\widehat{\Delta}_n - \Delta_n$ in the sense that $\sigma_n^{-1}(\widehat{\Delta}_n - \Delta_n) = O_p(1)$. More specifically, σ_n could be the standard deviation of $\widehat{\Delta}_n - \Delta_n$. Another choice, which is more readily available in our context is $\sigma_n = \max[\widehat{\theta}_{3/4}^u - \widehat{\theta}_{1/4}^u, \widehat{\theta}_{3/4}^l - \widehat{\theta}_{1/4}^l]$.

The construction above employs reasoning analogous to that of Imbens and Manski (2004) and Stoye (2009), though the specifics differ since the former approaches do not apply here. The reasoning behind our construction is as follows. If the width Δ_n of the identification region is bounded away from zero, then θ^* can be close to either the lower bound θ_0^l or the upper bound θ_0^u but not both, so in this case the end-points $\widehat{\theta}_{1-\alpha}^u$ and $\widehat{\theta}_{1-\alpha}^l$ from one-sided intervals suffice for a two-sided interval. If Δ_n is zero or approaches zero, then θ^* can be close to both the lower bound θ_0^l and the upper bound θ_0^u simultaneously, so in this case the more conservative end-points $\widehat{\theta}_{1-\alpha/2}^u$ and $\widehat{\theta}_{1-\alpha/2}^l$ are needed for a valid two-sided confidence interval. To smoothly and robustly interpolate between the two situations, we use the end-points $\widehat{\theta}_{\widehat{p}_n}^u$ and $\widehat{\theta}_{\widehat{p}_n}^l$ from one-sided intervals with the level $\widehat{p}_n \in [1-\alpha, 1-\alpha/2]$ varying smoothly as a function of Δ_n.

4. Monte Carlo Experiments

In this section we present the results of some Monte Carlo experiments that illustrate the finite-sample performance of our method. We consider a Monte Carlo design that is similar to that of Manski and Pepper (2009). In particular, we consider the lower bound on $\theta^* = E[Y_i(t)|V_i = v]$ under the monotone instrumental variable (MIV) assumption, where t is a treatment, $Y_i(t)$ is the corresponding potential outcome, and V_i is a monotone instrumental variable. The lower bound on $E[Y_i(t)|V_i = v]$ can be written as

$$\max_{u \leq v} E\left[Y_i \cdot 1\{Z_i = t\} + y_0 \cdot 1\{Z_i \neq t\} | V_i = u\right], \tag{4.1}$$

where Y_i is the observed outcome, Z_i is a realized treatment, and y_0 is the left endpoint of the support of Y_i, see Manski and Pepper (2009). Throughout the Monte Carlo experiments, the parameter of interest is $\theta^* = E[Y_i(1)|V_i = 1.5]$.

4.1. **Data-Generating Processes.** We consider two cases of data-generating processes (DGPs). In the first case, which we call DGP1, $V_0 = \mathcal{V}$ and the MIV assumption has no identifying power. In other words, the boundary-generating function is flat on $\mathcal{V}$, in which case the bias of the analog estimator is most acute, see Manski and Pepper (2009). In the second case, which we call DGP2, the MIV assumption has identifying power, and V_0 is a strict subset of $\mathcal{V}$.

Specifically, for both DGPs we generated 1000 independent samples from the following model:

$$V_i \sim \text{Unif}[-2, 2], Z_i = 1\{\varphi_0(V_i) + \varepsilon_i > 0\}, \text{ and } Y_i = \mu_0(V_i) + \sigma_0(V_i)U_i,$$

where $\varepsilon_i \sim N(0,1)$, $\eta_i \sim N(0,1)$, $U_i = \min\{\max\{-1.96, \eta_i\}, 1.96\}$, and $(V_i, \eta_i, \varepsilon_i)$ are statistically independent, where $i = 1, \ldots, n$. For DGP1, $\varphi_0(v) \equiv 0$, $\mu_0(v) \equiv 0$, and $\sigma_0(v) = |v|$. In this case, the bound-generating function

$$\theta_l(v) := E\left[Y_i \cdot 1\{Z_i = 1\} + y_0 \cdot 1\{Z_i \neq 1\} | V_i = v\right]$$

is completely flat ($\theta_l(v) = -0.98$ for each $v \in \mathcal{V} = [-2, 1.5]$). For DGP2, an alternative specification is considered:

$$\varphi_0(v) = v1(v \leq 1) + 1(v > 1), \mu_0(v) = 2[v1(v \leq 1) + 1(v > 1)], \text{ and } \sigma_0(v) = |v|.$$

In this case, $\theta_l(v) = \mu_0(v)\Phi[\varphi_0(v)] - 1.96\Phi[-\varphi_0(v)]$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Thus, $v \mapsto \theta_l(v)$ is strictly increasing on $[-2, 1]$ and is flat on $[1, 2]$, and $V_0 = [1, 1.5]$ is a strict subset of $\mathcal{V} = [-2, 1.5]$.

We considered sample sizes $n = 500$ and $n = 1000$, and we implemented both series and kernel-type estimators to estimate the bound-generating function $\theta_l(v)$ in (4.1). For both estimators, we computed critical values via simulation as described in Appendix C.2, and we implemented our method both with and without estimating V_ϵ. For the latter, the precision-corrected curve is maximized on the interval between the 5th percentile of V_i and the point 1.5. We do this in order to avoid undue influence of outliers at the boundary of the support of V_i. For the former, V_ϵ is estimated by $\widehat{V}_\epsilon$ in (3.14) with $\epsilon = 10^{-6}$, $c_n = \sqrt{\log n}$, and $\ell_n = 2\sqrt{\log n} \cdot \sup_{v \in \mathcal{V}} s(v)$.

4.2. **Series Estimation.** For basis functions we used cubic B-splines with knots equally spaced over the sample quantiles of V_i. The number K of approximating functions was obtained by the following simple rule-of-thumb:

$$K = \underline{\widehat{K}}, \quad \widehat{K} := \widehat{K}_{cv} \times n^{-1/5} \times n^{2/7}, \tag{4.2}$$

where $\underline{a}$ is defined as the largest integer that is smaller than or equal to a, and $\widehat{K}_{cv}$ is the minimizer of the leave-one-out least squares cross validation score from the set $\{5, 6, 7, 8, 9\}$. If $\theta_l(v)$ is twice continuously differentiable, then a cross-validated K has the form $K \propto n^{1/5}$ asymptotically. Hence, the multiplicative factor $n^{-1/5} \times n^{2/7}$ in (4.2) ensures that the bias is asymptotically negligible from under-smoothing.[6]

We obtained the precision-corrected curve for the lower bound by subtracting the product of a critical value and an asymptotic pointwise standard error from the estimated function. At each data point of V_i, we computed the pointwise standard error of our estimate using an asymptotic heteroscedasticity-robust formula.

4.3. **Kernel-Type Estimation.** We used local linear smoothing since it is known to behave better at the boundaries of the support than the standard kernel method. We used the kernel function $K(s) = \frac{15}{16}(1 - s^2)^2 1(|s| \leq 1)$ and the following rule of thumb bandwidth:

$$h = \widehat{h}_{ROT} \times \widehat{s}_v \times n^{1/5} \times n^{-2/7}, \tag{4.3}$$

[6]To check the sensitivity of simulation results, we considered alternative bandwidths such as $K \pm 1$ or $K \pm 2$ and found that the simulation results were not very sensitive within the local range around our rule-of-thumb choice.

where $\widehat{h}_{ROT}$ is the rule-of-the-thumb bandwidth for estimation of $\theta_l(v)$ with studentized V, as prescribed in Section 4.2 of Fan and Gijbels (1996). The exact form of $\widehat{h}_{ROT}$ is

$$\widehat{h}_{ROT} = 2.036 \left[\frac{\tilde{\sigma}^2 \int w_0(v)dv}{n^{-1} \sum_{i=1}^n \left\{ \tilde{\theta}_l^{(2)}(\tilde{V}_i) \right\}^2 w_0(\tilde{V}_i)} \right]^{1/5} n^{-1/5},$$

where $\tilde{V}_i$'s are studentized V_i's, $\tilde{\theta}_l^{(2)}(\cdot)$ is the second-order derivative of the global quartic parametric fit of $\theta_l(v)$ with studentized V_i, $\tilde{\sigma}^2$ is the simple average of squared residuals from the parametric fit, $w_0(\cdot)$ is a uniform weight function that has value 1 for any $\tilde{V}_i$ that is between the 10th and 90th sample quantiles of $\tilde{V}_i$. Again, the factor $n^{1/5} \times n^{-2/7}$ is multiplied in (4.3) to ensure that the bias is asymptotically negligible due to under-smoothing.[7]

At each data point of V_i, we computed an estimate of the pointwise standard error with the asymptotic standard error formula $[nhf_V(v)]^{-1} \int K^2(u)du\, \sigma^2(v)$, where f_V is the density of V and $\sigma^2(v)$ is the conditional variance function. We estimated f_V and $\sigma^2(v)$ using the standard kernel density and regression estimators with the same bandwidth h.

4.4. **Simulation Results.** Table 1 summarizes the results of Monte Carlo experiments. To evaluate the relative performance of our new estimator, we also consider a simple analog estimator of the left-hand side of (4.1).

First, we consider Monte Carlo results for the series estimator for DGP1 with $n = 500$. In this case, not surprisingly, the simple analog estimator suffers from substantial biases since the true bound-generating function is flat on $\mathcal{V}$. However, our new estimator, which is asymptotically median unbiased, has negligible mean bias and even smaller median bias. One potential concern with the new estimator is that it may have a larger variance due to the fact that we need to estimate the pointwise standard error for each point. However, it turns out that with DGP1, the new estimator has smaller standard deviation (SD) and also smaller mean absolute deviation. As a result, the new estimator enjoys substantial gains relative to the analog estimator in terms of the root mean square error (RMSE). It is interesting to comment on estimation of V_ϵ in this case. Since the true argmax set V_0 is equal to $\mathcal{V}$, an estimated V_ϵ should be the entire set $\mathcal{V}$. Note that the simulation results are similar since for many simulation draws, $\widehat{V}_\epsilon = \mathcal{V}$. Similar conclusions hold for the sample size $n = 1000$. Note that the biases of the sample analog

[7] As in series estimation, we considered alternative bandwidths such as $0.8h$ or $1.2h$ and found that the qualitative findings of Monte Carlo experiments were the same.

estimator are still quite large, even though it is a consistent estimator. The discrepancies between nominal and actual coverage probabilities are not large.

We now move to DGP2. In this case, the true argmax set V_0 is $[1, 1.5]$. In this case, our estimator is upward median unbiased and the coverage probability is conservative. The Monte Carlo results are consistent with asymptotic theory. As in DGP1, the sample analog estimator suffers from upward biases. However, unlike in DGP1, our new proposed estimator has a slightly larger RMSE than the analog estimator with $n = 500$. In DGP2, the true argmax set V_0 is a strict subset of $\mathcal{V}$. Hence, we expect that it is important to estimate V_ϵ. On average, the estimated sets were $[-0.847, 1.5]$ when $n = 500$ and $[-0.147, 1.5]$ when $n = 1,000$. As can be seen from the table, our method performed better when V_ϵ is estimated in terms of making the bound estimates and confidence intervals less conservative. However, there was no gain for the sample analog method even with the estimated V_ϵ. When $n = 1,000$ and V_ϵ is estimated, the RMSE of the new proposed estimator is more than 10% lower than that of the sample analog estimator.

We now comment on local linear estimation. Overall, simulation results are quite similar for both the series estimator and the local linear estimator. With DGP1, the differences between the two estimators are negligible, but with DGP2, it seems that the series estimator performs slightly better than the local linear estimator. We conclude from the Monte Carlo experiments that our inference method performs well in coverage probabilities and that our proposed estimator outperforms the sample analog estimator, especially when the MIV assumption has no identifying power.

5. An Empirical Application

In this section, we illustrate our inference procedure by applying it to a MIV-MTR (monotone instrument variable - monotone treatment response) bound of Manski and Pepper (2000, Proposition 2). The parameter of interest is $E[Y_i(t)|V_i = v]$, where t is a treatment, $Y_i(t)$ is a potential outcome variable corresponding to a treatment t, and V_i is a scalar explanatory variable. Let Z_i denote the realized treatment that is possibly self-selected by individuals. The source of the identification problem here is that for each individual i, we only observe $Y_i \equiv Y_i(Z_i)$ along with (Z_i, V_i), but not $Y_i(t)$ with $t \neq Z_i$. The MIV-MTR bounds take the form

$$\sup_{u \leq v} E\left[Y_i^l | V_i = u\right] \leq E[Y_i(t)|V_i = v] \leq \inf_{u \geq v} E\left[Y_i^u | V_i = u\right],$$

where $Y_i^l = Y_i \cdot 1\{t \geq Z_i\} + y_0 \cdot 1\{t < Z_i\}$, $Y_i^u = Y_i \cdot 1\{t \leq Z_i\} + y_1 \cdot 1\{t > Z_i\}$, and $[y_0, y_1]$ is the support of Y_i. Thus the bound-generating functions are $\theta^l(v) = E\left[Y_i^l | V_i = v\right]$ and $\theta^u(v) = E\left[Y_i^u | V_i = v\right]$ with intersection sets $\mathcal{V}^l = (-\infty, v]$ for the lower bound and $\mathcal{V}^u = [v, \infty)$ for the upper bound. Note that the MIV-MTR bounds are uninformative if the support of Y is unbounded. In the empirical illustration below, we use the sample minimum and maximum as the boundary points of the support.

We use data from the National Longitudinal Survey of Youth of 1979 (NLSY79); in particular, we use the same data extract as Carneiro and Lee (2009) giving us $n = 2044$ observations. The outcome variable Y_i is the logarithm of hourly wages in 1994. In order to alleviate problems induced by possible measurement error and the occurrence of missing wages, Y_i is constructed as a 5 year average of all non-missing wages reported in the five year interval centered in the year 1994. The treatment variable t is years of schooling. The monotone instrumental variable V_i is the Armed Forces Qualifying Test score (AFQT, a measure of cognitive ability), normalized to have mean zero in the NLSY population. The MIV assumption here stipulates that the conditional expectation of potential log wages at any level of schooling is nondecreasing in AFQT score. The use of AFQT as a MIV can tighten the bound, but its empirical implementation carries some challenges since the bounds are the suprema and infima of nonparametric estimates.[8] Table 2 presents descriptive statistics for our sample.

Our targets are the MIV-MTR bounds for $E[Y_i(t)|v]$ at $v = 0$ (the mean value of AFQT) for high school graduates ($t = 12$) and college graduates ($t = 16$). We estimate the bound-generating functions $E\left[Y_i^l | V_i = u\right]$ and $E\left[Y_i^u | V_i = u\right]$ by local linear smoothing. For each nonparametric function, we use the same kernel function and rule-of-thumb as in Section 4.3. In addition, we used the critical value in (3.11) and estimated V_ϵ as in Section 4.3.

Table 3 summarizes our empirical results. The first row shows naïve sample analog estimates, which are based on the maxima and minima of the bound-generating functions. The second row presents our median downward-unbiased (upward-unbiased) estimates for the upper (lower) bounds. We see that our estimate and the analog estimate for the upper bound of average log wages for college graduates differ quite substantially. The economic implication of this difference is large: the upper bound for the return to college (defined as $E[Y_i(16)|V_i = 0] - E[Y_i(12)|V_i = 0]$) is $2.87 - 2.12 = 0.75$ based on the naïve

[8]The NBER working paper version of Manski and Pepper (1998) also considered AFQT as a MIV. See the comments in the NBER working paper version of Manski and Pepper (1998, Section 6.2) for discussion of the difficulty of carrying out inference.

sample analog estimates, whereas it is $3.18 - 2.03 = 1.15$ based on our proposed new estimates. The resulting difference between the two estimates of the upper bound for the return to college is 40%, a 10% difference in terms of one year of college education.

We now consider 95% one-sided confidence intervals, which are given in the third row of the table. If we combine upper and lower bounds together, then we obtain a 90 % confidence interval for average log wages of each education group. Note that the 90% confidence interval for the potential average college log wages is wider that the 90% confidence interval of the high school wages.[9] This is because the estimate of the upper bound-generating function for college wages is rather imprecise.

In order to illustrate the sources of the difference between naïve sample analog estimates and our estimates in Figures 4 and 5, we plot estimated bound-generating functions $v \mapsto E\left[Y_i^j | V_i = v\right], j = l, u$ as well as precision-corrected bound-generating functions for college graduates. In Figure 4 we see that the sudden drop of the estimated bound-generating function in the right tail for college graduates tightens the empirical MIV-MTR bound, but the tightness of this bound could be due to reduced precision of the local linear estimator at the boundary. On the other hand, our new method automatically corrects for varying degree of precision.

6. Conclusion

In this paper we provided a novel method for inference on intersection bounds. Bounds of this form are common in the recent literature, but two issues have posed difficulties for valid asymptotic inference and bias-corrected estimation. First, the application of the supremum and infimum operators to boundary estimates results in finite-sample bias. Second, unequal sampling error of estimated boundary functions complicates inference. We overcame these difficulties by applying a precision-correction to the estimated boundary functions before taking their intersection. We employed strong approximation to justify the magnitude of the correction in order to achieve the correct asymptotic size. As a by-product, we proposed a bias-corrected estimator for intersection bounds based on an asymptotic median adjustment. We provided formal conditions that justified our approach in both parametric and nonparametric settings, the latter using either kernel or series estimators. As such, our method is the first to provide valid inference for nonparametric specifications of a continuum of conditional moment inequalities.

[9]To check sensitivity to the choice of critical values, we obtained the corresponding confidence intervals using critical values in (3.12). It turns out that the resulting 90% confidence intervals are almost identical: [1.96,2.88] for high school wages and [2.30, 3.46] for college wages, respectively.

At least two of our results may be of independent interest beyond the scope of inference on intersection bounds. First, our result on the strong approximation of series estimator is new. This essentially provides a functional central limit theorem for any series estimator that admits a linear asymptotic expansion, and is applicable quite generally. Second, our method for inference applies to any value that can be defined as a linear programming problem with either finite or infinite dimensional constraint set. Estimators of this form can arise in a variety of contexts, including, but not limited to intersection bounds. We therefore anticipate that although our motivation lay in inference on intersection bounds, our results may have further application.

Appendix A. Proofs

Proof of Theorem 1. We prove the results assuming condition C*.1 only, since condition C.1 is a special case with $\mathcal{E}_n(V) = \mathcal{E}_\infty(V)$.

Part 1. Observe that

$$\begin{aligned}
P[\theta_0 \leq \widehat{\theta}_p] &= P[\inf_{v\in\widehat{V}}[\widehat{\theta}(v) - \theta_0 + [\widehat{b}_n + \widehat{c}(p)/\widehat{a}_n]s(v)] \geq 0] \\
&\geq P[\inf_{v\in\widehat{V}}[\widehat{\theta}(v) - \theta(v) + [\widehat{b}_n + \widehat{c}(p)/\widehat{a}_n]s(v)] \geq 0] \\
&= P[\widehat{a}_n[\widehat{\theta}(v) - \theta(v)]/s(v) + \widehat{a}_n\widehat{b}_n \geq -\widehat{c}(p), \forall v \in \widehat{V}] \\
&= P[\widehat{a}_n[Z_n(v) - \widehat{b}_n] \leq \widehat{c}(p), \forall v \in \widehat{V}] \\
&= P[\widehat{a}_n[\sup_{v\in\widehat{V}} Z_n(v) - \widehat{b}_n] \leq \widehat{c}(p)] \\
&= P[a_n[\sup_{v\in V} Z_n(v) - b_n] \leq \widehat{c}(p) + o_p(1)],
\end{aligned}$$

where we used that $\theta(v) \geq \theta_0$ for all $v \in \mathcal{V}$ as well as condition C.2. Then we observe that using condition C.1*(b) and the anti-concentration property

$$\begin{aligned}
P[a_n[\sup_{v\in V} Z_n(v) - b_n] \leq \widehat{c}(p) + o_p(1)] &= P[\mathcal{E}_n(V) \leq \widehat{c}(p) + o_p(1)] \\
&\geq P[\mathcal{E}_n(V) \leq c_n(p) + o_p(1)] \\
&\geq P[\mathcal{E}_n(V) \leq c_n(p)] - P[\mathcal{E}_n(V) \in [c_n(p) \pm o_p(1)]] \\
&\geq p - o(1),
\end{aligned}$$

which proves part 1.

Part 2. Using Condition C*.1(a) and C.2, we obtain

$$\begin{aligned}
P[\theta_0 \leq \widehat{\theta}_p] &= P[\inf_{v\in\widehat{V}}[\widehat{\theta}(v) - \theta_0 + [\widehat{b}_n + \widehat{c}(p)/\widehat{a}_n]s(v)] \geq 0] \\
&= P[\widehat{a}_n[\sup_{v\in\widehat{V}} \frac{\theta_0 - \widehat{\theta}(v)}{s(v)} - \widehat{b}_n] \leq \widehat{c}(p)] \\
&= P[a_n[\sup_{v\in V_0} \frac{\theta_0 - \widehat{\theta}(v)}{s(v)} - b_n] \leq \widehat{c}(p) + o_p(1)] \\
&= P[a_n[\sup_{v\in V_0} Z_n(v) - b_n] \leq \widehat{c}(p) + o_p(1)] \\
&= P[\mathcal{E}_n(V_0) \leq c_n(p) + o_p(1)] \\
&= P[\mathcal{E}_n(V_0) \leq c_n(p)] + w_n \text{ where } |w_n| \leq P[\mathcal{E}_n(V_0) \in [c_n(p) \pm o_p(1)]] \\
&= p + o(1),
\end{aligned}$$

where we also used that $\theta(v) = \theta_0$ for all $v \in V_0$, and the continuity of the limit distribution of $\mathcal{E}_\infty(V_0)$. □.

Proof of Lemma 1. From the Donsker condition and by the Continuous Mapping Theorem, we have that

$$\sup_{v\in V_0} Z_n(v) =_d \mathcal{E}_\infty(V_0) + o_p(1) = \sup_{v\in V_0} Z_\infty(v) + o_p(1).$$

Moreover, the distribution of the limit variable is continuous by the non-degeneracy of the covariance kernel. This verifies condition C.1(a).

By the stochastic equicontinuity, we have that

$$|\sup_{v\in\widehat{V}} Z_n(v) - \sup_{v\in V} Z_n(v)| \leq \sup_{|v-v'|\leq d_H(\widehat{V},V)} |Z_n(v) - Z_n(v')| = o_p(1),$$

for any sequence of sets $\widehat{V}$ such that $d_H(\widehat{V}, V) = O_p(r_n) = o_p(1)$. This implies condition C.2. □.

Proof of Lemma 2. This is immediate from the statement of the conditions. □.

Proof of Lemma 3. We have by Taylor expansion that

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}(v) - \theta(v)) &=_d \frac{\partial\theta(v, \gamma_0 + o_p(1))'}{\partial\gamma}(\sqrt{n}(\widehat{\gamma} - \gamma_0)) \\
&=_d \frac{\partial\theta(v, \gamma_0 + o_p(1))'}{\partial\gamma}(\Omega^{1/2}\mathcal{N} + o_p(1)) \\
&=_d g(v)'\mathcal{N} + o_p(1).
\end{aligned}$$

Then in $\ell^\infty(\mathcal{V})$,

$$\begin{aligned} Z_n(v) = \frac{(\theta(v) - \widehat{\theta}(v))}{s(v)} &=_d \frac{g(v)'\mathcal{N} + o_p(1)}{\|g(v)\| + o_p(1)} \\ &=_d \frac{g(v)'\mathcal{N}}{\|g(v)\|} + o_p(1). \end{aligned}$$

□.

Proof of Lemma 4 By assumption we have that $a_n \cdot \sup_{v\in V} |Z_n(v) - Z_n'(v)| = o_p(1)$ for any a_n, including a_n stated in the Lemma. Now we set $\mathcal{E}_n(V) = a_n[\sup_{v\in V} Z_n'(v) - b_n]$. This random variable need not have a limit distribution, but its exact distribution can be obtained by simulation. Thus, in our words, its distribution is known.

Case 1 (Dimension of regressor v is one). In this case, we can also use Hotelling's tubing method to conservatively estimate the quantiles of $\mathcal{E}_n(V)$. Expressions for dimensions greater than one are less tractable, but they can also be stated at the cost of complicated notation.

Indeed, from the Hotelling-Naik tubing method we obtain that

$$P[\sup_{v\in V} Z_n'(v) \geq k] \leq (1 - \Phi(k)) + \frac{\kappa_n(V)}{2\pi} e^{-k^2/2}$$

where $\kappa_n(V) = \int_V \|\nabla \alpha_n(v)\| dv$. As $k \to \infty$, we have

$$P[\sup_{v\in V} Z_n'(v) \geq k] \leq \frac{\kappa_n(V)}{2\pi} e^{-k^2/2}[1 + o(1)].$$

For any p, we choose $k = k_n(p) = a_n + p/a_n$. Note that

$$a_n = \sqrt{2\log(\kappa_n(V)/2\pi)} \;\Leftrightarrow\; \frac{\kappa_n(V)}{2\pi} e^{-a_n^2/2} = 1.$$

Then

$$P[\sup_{v\in V} Z_n'(v) \geq k_n(p)] \leq \exp\left(-p - \frac{p^2}{2a_n^2}\right)[1 + o(1)],$$

equivalently

$$P[a_n[\sup_{v\in V} Z_n'(v) - a_n] \geq p] \leq \exp\left(-p - \frac{p^2}{2a_n^2}\right)[1 + o(1)]. \tag{A.1}$$

Using the above relations, we conclude that the quantiles of $\mathcal{E}_n(V)$ can be estimated conservatively by the quantiles of an exponential distribution or by the quantiles of an exponential-distribution-like function $F_n(p) := 1 - \exp\left(-p - \frac{p^2}{2a_n^2}\right)$. Thus, we have established N.1(b) when the dimension of the regressors equals one.

CASE 2. (DIMENSION OF REGRESSOR v IS ANY). With regressors of higher dimension, we can also use the following argument. Since the metric entropy of $\{Z_n(v)', v \in V\}$ under the L_2 pseudometric ρ satisfies

$$N(\epsilon, V, \rho) \leq \left(\frac{c \cdot L}{\epsilon}\right)^d, \quad L = \sup_{v \in \mathcal{V}} \|\nabla \alpha_n(v)\| \cdot \text{diam}(V) \lesssim K^p \text{ for some constant } p < \infty,$$

we have by Samorodnitsky-Talagrand's inequality (van der Vaart and Wellner (1996), p. 442) that for ℓ large enough and some constant C

$$P(\sup_{v \in V} Z_n'(v) > \ell) \leq 2\,(C \cdot L)^d\,(1 - \Phi(\ell)).$$

Since

$$T_n := \sqrt{\log 2(C \cdot L)^d} \lesssim \sqrt{\log K},$$

the bound implies by Feller's inequality

$$P(\sup_{v \in V} Z_n'(v) > \ell) \leq \frac{\ell^{-1}}{\sqrt{2\pi}} e^{-\ell^2/2 + T_n^2/2}, \quad T_n \lesssim \sqrt{\log K}$$

We thus conclude that

$$\sup_{v \in V} |Z_n'(v)| = O_p(\sqrt{\log K}).$$

For any p, we choose $k = k_n(p) = a_n + p/a_n$, where a_n is the largest solution to

$$\frac{2}{\sqrt{2\pi}}\,(C \cdot L)^d\, a_n^{-1} e^{-a_n^2/2} = 1$$

which implies that as $n \to \infty$, using the assumption that $\sup_v \|\nabla_v g_n(v)/\|g_n(v)\|\| = O(K^p)$,

$$a_n \sim \sqrt{2d \log(2LC/\sqrt{2\pi})} \sim \sqrt{2d \log(2L/\sqrt{2\pi})} \lesssim \sqrt{2d \log K}.$$

Then

$$P[\sup_{v \in V} Z_n'(v) \geq k_n(p)] \leq \frac{a_n}{k_n(p)} \exp\left(-p - \frac{p^2}{2a_n^2}\right)[1 + o(1)] \leq \exp\left(-p - \frac{p^2}{2a_n^2}\right)[1 + o(1)],$$

equivalently

$$P[a_n[\sup_{v \in V} Z_n'(v) - a_n] \geq p] \leq \exp\left(-p - \frac{p^2}{2a_n^2}\right)[1 + o(1)]. \tag{A.2}$$

Thus, we have established N.1(b). □.

Proof of Lemma 5. In order to establish N.2, we can use a crude approach based on Cauchy-Schwarz inequalities

$$\begin{aligned} a_n \Big| \sup_{v\in\widehat{V}} Z_n'(v) - \sup_{v\in V} Z_n'(v)| \quad &\leq a_n \sup_{|v-v'|\leq r_n} |(\alpha_n(v)-\alpha_n(v'))'\mathcal{N}_n| \\ &\leq a_n \sup_{|v-v'|\leq r_n, |\bar{v}-v'|\leq r_n} \|\nabla\alpha_n(\bar{v})\|\|v-v'\|\|\mathcal{N}_n\| \\ &\leq a_n \sup_{v\in\mathcal{V}} \|\nabla\alpha_n(v)\| r_n O_p(\sqrt{K}). \end{aligned}$$

Provided $a_n \sup_{v\in\mathcal{V}} \|\nabla\alpha_n(v)\| r_n\sqrt{K} \to 0$, we have the result. However, a substantially better condition follows from a careful use of Samorodnitsky-Talagrand's inequalities for Gaussian processes as shown below. The strategy shown below has been used by Belloni and Chernozhukov (2007) to bound oscillations of Gaussian processes of increasing dimension over vanishing neigborhoods. Here, we adopt their strategy to our case, which is quite a bit different due to particular structure of the function $\alpha_n(v)$.

We will use the following Samorodnitsky-Talagrand maximal inequality for Gaussian processes (Proposition A.2.7 in Van der Vaart and Wellner (1998)). Let X be a separable zero-mean Gaussian process indexed by a set T. Suppose that for some $\kappa > \sigma(X) = \sup_{t\in T}\sigma(X_t)$, $0 < \epsilon_0 \leq \sigma(X)$, we have

$$N(\varepsilon, T, \rho) \leq \left(\frac{\kappa}{\varepsilon}\right)^v, \text{ for } 0 < \varepsilon < \epsilon_0,$$

where $N(\varepsilon, T, \rho)$ is the covering number of T by ε-balls w.r.t. the standard deviation metric $\rho(t,t') = \sigma(X_t - X_{t'})$. Then there exist an universal constant D such that for every $\lambda \geq \sigma^2(X)(1+\sqrt{v})/\epsilon_0$ we have

$$P\left(\sup_{t\in T} X_t > \lambda\right) \leq \left(\frac{D\kappa\lambda}{\sqrt{v}\sigma^2(X)}\right)^v (1 - \Phi(\lambda/\sigma(X))).$$

We apply this result to the zero-mean Gaussian process $X_n : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ defined as

$$X_{n,t} = (\alpha_n(v) - \alpha_n(v'))'\mathcal{N}_n, \quad t = (v, v') : \ |v - v'| \leq r_n.$$

It follows that $\sup_{t\in T} X_{n,t} = \sup_{|v-v'|\leq r_n}(\alpha_n(v) - \alpha_n(v'))'\mathcal{N}_n$. For the process X_n we have:

$$\sigma(X_n) \leq \sup_{\|v-v'\|\leq r_n} \|\alpha_n(v) - \alpha_n(v')\| \leq \sup_{v\in\mathcal{V}} \|\nabla\alpha_n(v)\| r_n.$$

Furthermore we have that

$$N(\epsilon, T, \rho) \leq \left(\frac{C\cdot L}{\epsilon}\right)^d, \quad L := \sup_{v\in\mathcal{V}} \cdot \|\nabla\alpha_n(v)\| \cdot r_n \cdot \mathrm{diam}(\mathcal{V}^2) \lesssim \sup_{v\in\mathcal{V}} \|\nabla\alpha_n(v)\| \cdot r_n,$$

so that the bound on covering numbers holds with $\kappa \lesssim L$, and $v = d$. Applying the Samorodnitsky-Talagrand inequality we conclude that for every $\ell \to \infty$, $\epsilon_0 = \sigma(X_n)$,

$$Pr\{\sup_{t\in T} X_{n,t} > \ell\sigma(X_n)\} \lesssim (1 - \Phi(\ell)) \to 0.$$

Therefore, we conclude that $\sup_{t\in T} X_t = O_p(\sigma(X_n))$. Thus,

$$\begin{aligned} a_n\Big|\sup_{v\in\widehat{V}} Z_n'(v) - \sup_{v\in V} Z_n'(v)\Big| &\leq a_n \sup_{|v-v'|\leq r_n} |(\alpha_n(v) - \alpha_n(v'))'\mathcal{N}_n| \\ &= O_p(a_n \sup_{v\in\mathcal{V}} \|\nabla\alpha_n(v)\| r_n), \end{aligned}$$

which is $o_p(1)$ by our assumption. Hence, we have shown that

$$a_n \cdot \sup_{v\in\widehat{V}} Z_n'(v) - a_n \cdot \sup_{v\in V} Z_n'(v) = o_p(1).$$

Since $a_n = b_n$, it only remains to show that $(\widehat{a}_n - a_n)\sup_{v\in\widehat{V}} Z_n'(v) \to_p 0$. Note that $\sup_{v\in\widehat{V}} Z_n'(v) = O_p(a_n)$ and $(\widehat{a}_n - a_n) = (\widehat{a}_n^2 - a_n^2)/(\widehat{a}_n + a_n)$. Therefore,

$$(\widehat{a}_n - a_n)\sup_{v\in\widehat{V}} Z_n'(v) = O_p(\widehat{a}_n^2 - a_n^2),$$

which is $o_p(1)$ by assumption. □.

Proof of Lemma 6. Arguments similar to those used in the proof of Lemma 3.4 of Ghosal, Sen, and van der Vaart (2000) yield

$$\frac{w_n(v)'\mathbf{U_n}}{\|w_n(v)\|} =_d Z_n''\left(h_n^{-1}v\right) + r_n'(v), \tag{A.3}$$

where

$$\sup_{v\in V} |r_n'(v)| = O_p\left(h_n\sqrt{\log h_n^{-1}}\right).$$

Now note that Conditions K.1 and K.2, along with (A.3), imply that

$$\sup_{v\in V}\left|Z_n(v) - Z_n''(h_n^{-1}v)\right| = o_p(a_n(V)^{-1}).$$

Since the distribution of $Z_n''(s)$ does not depend on n, for the purpose of statistical inference, it suffices to consider the asymptotic behavior of a Gaussian process, say $Z'(s)$, that has the same covariance function as $Z_n''(s)$. We first derive the asymptotic behavior of the tail probability of the maximum of $Z'(s)$ over s on a set $\mathcal{S}$ with a fixed measure, mes($\mathcal{S}$). Define

$$\Psi(a) = \frac{1}{\sqrt{2\pi}}\int_a^\infty \exp\left(-\frac{1}{2}x^2\right) dx.$$

Recall that

$$\lambda = \frac{-\int K(u)K''(u)du}{\int K^2(u)du}.$$

We can prove that

$$\Pr\left(\max_{s\in\mathcal{S}} Z'(s) > a\right) = \text{mes}(\mathcal{S})\left(\frac{\lambda}{2\pi}\right)^{d/2} a^d \Psi(a)[1+o(1)] \tag{A.4}$$

as $a \to \infty$. To show this, we use the double sum method developed in Piterbarg (1996), applying in particular Piterbarg's Lemma 7.1. Note that for each j,

$$\rho(s_j) = 1 - \frac{\lambda}{2}s_j^2 + o(s_j^2), \quad s_j \to 0,$$

and that $\rho(s_j) = 0$ for $s_j > 2$ (K has support in $[-1,1]$). Hence,

$$\rho_d\left[(2/\lambda)^{1/2} s\right] = 1 - \sum_{j=1}^{d} s_j^2 + o\left(\sum_{j=1}^{d} s_j^2\right)$$

as $s \to 0$. Thus, the Gaussian process $Z'(s)$ has a stationary structure $\left(E^{(d)}, \alpha^{(d)}\right)$ with $C = \text{diag}(\sqrt{2/\lambda}, \ldots, \sqrt{2/\lambda})$, $E^{(d)} = (1, \ldots, 1)$ and $\alpha^{(d)} = (2, \ldots, 2)$ (using the notation in Piterbarg (1996)). Then an application of Corollary 7.1 of Piterbarg (1996) gives

$$\Pr\left(\max_{s\in\mathcal{S}} Z'(s) > a\right) = H_{E^{(d)},\alpha^{(d)}} \text{mes}(\mathcal{S}) a^d \Psi(a)(1+o(1)) \tag{A.5}$$

as $a \to \infty$, where $H_{E^{(d)},\alpha^{(d)}}$ is the Pickands' constant (see Section 4 of Piterbarg (1996) for its definition). In our case, $H_{E^{(d)},\alpha^{(d)}} = (\pi)^{-d/2}$ by (F.4) and Lemma 6.4 of Piterbarg (1996). Then, (A.4) follows immediately from (A.5).

Then arguments almost identical to those used in the proof of Theorem A.3 of Lee, Linton, and Whang (2009), which is based on the proof of Theorem G.1 of Piterbarg (1996), yield the following: for any x,

$$\begin{aligned}
&\Pr\left(a_n(V)\left[\sup_{v\in V} Z_n''\left(h_n^{-1}v\right) - a_n(V)\right] < x\right) \\
&= \exp\left\{-\exp\left(-x - \frac{x^2}{2a_n(V)^2}\right)\left[1 + \frac{x}{a_n(V)^2}\right]^{d-1}\right\} + o(1),
\end{aligned} \tag{A.6}$$

where $a_n(V)$ is defined in (3.7). Since $a_n(V) \to \infty$, (3.8) is proved. □.

Proof of Lemma 7. This lemma can be proved using arguments almost identical to those used to prove Lemma 5. In particular, as in the proof of Lemma 5, we apply Samorodnitsky-Talagrand's maximal inequality to the following zero-mean Gaussian

process $X_n : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, which is defined as

$$X_{n,t} = (\alpha_n(v) - \alpha_n(v'))'\mathbf{U_n}, \quad t = (v, v') : \ |v - v'| \leq r_n$$

with $\alpha_n(v) = \frac{w_n(v)}{\|w_n(v)\|}$. Then we have that

$$\begin{aligned} a_n \Big| \sup_{v \in \widehat{V}} Z_n'(v) - \sup_{v \in V} Z_n'(v) \Big| \quad &\leq a_n \sup_{|v-v'| \leq r_n} |(\alpha_n(v) - \alpha_n(v'))'\mathbf{U_n}| \\ &= O_p(a_n \sup_{v \in \mathcal{V}} \|\nabla \alpha_n(v)\| r_n). \end{aligned}$$

This proves the first conclusion of the lemma. To show the second conclusion of the lemma, note that

$$\|\nabla \alpha_n(v)\| \leq \frac{\|\nabla w_n(v)\|}{\|w_n(v)\|} + \frac{\|\nabla \|w_n(v)\|\|}{\|w_n(v)\|}.$$

Furthermore, since

$$\|w_n(v)\| = \left\{ \frac{1}{n h_n^d f_V^2(v)} \sum_{i=1}^n \sigma^2(V_i) \mathbf{K}^2 \left(\frac{v - V_i}{h_n} \right) \right\}^{1/2},$$

we have that

$$\nabla \|w_n(v)\| = \|w_n(v)\|^{-1} \left\{ \frac{1}{n h_n^{d+1} f_V^2(v)} \sum_{i=1}^n \sigma^2(V_i) \mathbf{K} \left(\frac{v - V_i}{h_n} \right) (\nabla \mathbf{K}) \left(\frac{v - V_i}{h_n} \right) \right\},$$

$$\begin{aligned} \|\nabla w_n(v)\| = \Big\{ \frac{1}{n h_n^d f_V^4(v)} \sum_{i=1}^n \sum_{j=1}^d \sigma^2(V_i) \Big[h_n^{-1} f_V(v) \, (\nabla_j \mathbf{K}) \left(\frac{v - V_i}{h_n} \right) \\ - \mathbf{K} \left(\frac{v - V_i}{h_n} \right) \nabla_j f_V(v) \Big]^2 \Big\}^{1/2}, \end{aligned}$$

where $\nabla_j \mathbf{K}$ and $\nabla_j f_V$ are the j-th elements of $\nabla \mathbf{K}$ and ∇f_V. Then under Condition K, $\|\nabla \alpha_n(v)\|$ is at most $O_p(h_n^{-1})$ uniformly over v. Therefore, we have proved the second conclusion of the lemma. □.

Proof of Theorem 2. Let

$$\zeta_n = c_n \sup_{v \in \mathcal{V}} s(v), \quad \gamma_n = \ell_n c_n, \quad \widehat{\theta}_0 = \min_{v \in \mathcal{V}} \widehat{\theta}(v) + \ell_n c_n.$$

Note that wp $\to 1$, $\sup_{v \in V_\epsilon} [\widehat{\theta}(v)] \leq \widehat{\theta}_0 + \epsilon$. This follows from two observations. First, by construction $\zeta_n = o_p(\ell_n c_n)$, so wp $\to 1$

$$\sup_{v \in V_\epsilon} [\widehat{\theta}(v)] \leq \sup_{v \in V_\epsilon} [\theta(v) + O_p(\zeta_n)] \leq \sup_{v \in V_\epsilon} [\theta(v) + (\ell_n/2) c_n] \leq \theta_0 + \epsilon + (\ell_n/2) c_n.$$

Second, wp $\to 1$

$$\begin{aligned}
\widehat{\theta}_0 + \epsilon &\geq \inf_{v \in \mathcal{V}} \{\theta(v) - O_p(\zeta_n) + \ell_n c_n\} + \epsilon, \\
&\geq \inf_{v \in \mathcal{V}} \{\theta(v) - (\ell_n/2)c_n + \ell_n c_n\} + \epsilon, \\
&\geq \inf_{v \in \mathcal{V}} \{\theta(v) + (\ell_n/2)c_n\} + \epsilon, \\
&\geq \theta_0 + (\ell_n/2)c_n + \epsilon.
\end{aligned}$$

Hence wp $\to 1$

$$V_\epsilon \subseteq \widehat{V}_\epsilon \text{ and } \sup_{v \in V_\epsilon} d(v, \widehat{V}_\epsilon) = 0.$$

Next,

$$\begin{aligned}
\sup_{v \in \widehat{V}_\epsilon} d(v, V_\epsilon) &= \sup\{d(v, V_\epsilon) : \widehat{\theta}(v) \leq \widehat{\theta}_0 + \epsilon\} \\
&\leq \sup\{d(v, V_\epsilon) : \theta(v) - \theta_0 - \epsilon \leq O_p(\zeta_n) + \gamma_n\} \\
&\leq \sup\{d(v, V_\epsilon) : \theta(v) - \theta_0 - \epsilon \leq \gamma_n \cdot (1 + o_p(1))\} \\
&\leq \sup\{d(v, V_\epsilon) : (cd(v, V_\epsilon))^{\rho(\epsilon)} \wedge \delta \leq \gamma_n(1 + o_p(1))\} \\
&\leq \sup\{x : (cx)^{\rho(\epsilon)} \wedge \delta \leq \gamma_n(1 + o_p(1))\} \\
&= \frac{[(\gamma_n + o_p(1))]^{1/\rho(\epsilon)}}{c} \quad \text{wp } \to 1.
\end{aligned}$$

The first claim of the theorem follows. The second claim follows from the inclusion $V_\epsilon \subseteq \widehat{V}_\epsilon$, so that

$$\text{mes}(\widehat{V}_\epsilon \setminus V_\epsilon) \lesssim [\sup_{v \in \widehat{V}_\epsilon} d(v, V_\epsilon)]^d \lesssim_p (\gamma_n)^{d/\rho(\epsilon)}. \quad \square$$

Proof of Lemma 8. Take any $v \notin V_\epsilon$. A projection of v on the set V_ϵ is defined as

$$v_\epsilon \in \arg \min_{v' \in \mathcal{V}: \theta(v') - \theta_0 \leq \epsilon} \|v - v'\|^2.$$

The Lagrangian characterization of the solution to this problem is of the form:

$$v - v_\epsilon = \lambda \nabla \eta(v_\epsilon)$$

for some scalar $\lambda > 0$. This is true because the solution is necessarily an interior one by V_ϵ belonging to the interior of $\mathcal{V}$ and the latter being a convex body in $\mathbb{R}^d$. Hence

$$v - v_\epsilon = \|v - v_\epsilon\| \frac{\nabla \eta(v_\epsilon)}{\|\nabla \eta(v_\epsilon)\|} = d(v, V_\epsilon) \frac{\nabla \eta(v_\epsilon)}{\|\nabla \eta(v_\epsilon)\|}.$$

By Taylor expansion we have that for some v_ϵ^* on the line joining v and v_ϵ

$$\theta(v) - \theta_0 - \epsilon = \eta(v) - \theta_0 - \epsilon = \nabla\eta(v_\epsilon^*)'(v - v_\epsilon) = \nabla\eta(v_\epsilon^*)'\frac{\nabla\eta(v_\epsilon)}{\|\nabla\eta(v_\epsilon)\|}d(v, V_\epsilon).$$

If $d(v, V_\epsilon) > 0$ is small enough, say $d(v, V_\epsilon) \leq d_0$, then by continuity of $\nabla\eta(v)$ we have that

$$\nabla\eta(v_\epsilon^*)'\frac{\nabla\eta(v_\epsilon)}{\|\nabla\eta(v_\epsilon)\|} \geq \frac{1}{2}\nabla\eta(v_\epsilon)'\frac{\nabla\eta(v_\epsilon)}{\|\nabla\eta(v_\epsilon)\|} = \|\nabla\eta(v_\epsilon)\|/2 \geq c,$$

where $c = \inf_{v\in\partial V_\epsilon} \|\nabla\eta(v)\|/2$. Thus, for $\delta = \inf_{d(v,V_\epsilon)\geq d_0}(\theta(v)-\theta_0-\epsilon) = \inf_{d(v,V_\epsilon)\geq d_0}(\eta(v)-\theta_0-\epsilon) > 0$,

$$\theta(v) - \theta_0 - \epsilon \geq cd(v, V_\epsilon)1\{d(v, V_\epsilon) \leq d_0\} + \delta 1\{d(v, V_\epsilon) > d_0\} \geq (cd(v, V_\epsilon)) \wedge \delta.$$

Finally, note that $\delta > 0$ by continuity of $\theta(v)$, by the definition of V_ϵ as ϵ-argmin of $\theta(v)$, and by $d_0 > 0$. □

Proof of Theorem 3. Recall that we construct the two-sided bands for the true parameter value θ^* as follows: Let

$$\widehat{\Delta}_n^+ \equiv \widehat{\Delta}_n 1[\widehat{\Delta}_n > 0], \text{ where } \widehat{\Delta}_n = \widehat{\theta}_{1/2}^u - \widehat{\theta}_{1/2}^l, \text{ and } \widehat{p}_n \equiv 1 - \Phi(\tau_n\widehat{\Delta}_n^+)\alpha,$$

where $\Phi(\cdot)$ is the standard normal CDF and $\tau_n \to \infty$ is a sequence of constants satisfying

$$\tau_n\{(a_n^j)^{-1} + b_n^j\}\bar{s}^j \to 0. \tag{A.7}$$

This condition implies that $\tau_n|\widehat{\Delta}_n^+ - \Delta_n| \to_p 0$, where $\Delta_n = \theta_0^u - \theta_0^l$. We also define $\bar{s}^j = \sup_{v\in\mathcal{V}^j} s^j(v)$, $j \in \{u, l\}$.

Step 1. We use the notation

$$p_n := 1 - \Phi(\tau_n\Delta_n)\alpha, \quad \Delta_n^u = \theta_0^u - \theta^*, \quad \Delta_n^l = \theta^* - \theta_0^l.$$

In what follows we allow θ^* to be an arbitrary sequence of constants within the identified set, so that its value can change depending on n; likewise, we allow $\Delta_n \geq 0$ to change with n.

The probability that θ^* lies outside the confidence interval is

$$P\left\{\theta^* \notin \left[\hat{\theta}_{\widehat{p}_n}^l, \hat{\theta}_{\widehat{p}_n}^u\right]\right\} \leq P\left\{\theta^* < \hat{\theta}_{\widehat{p}_n}^l\right\} + P\left\{\theta^* > \hat{\theta}_{\widehat{p}_n}^u\right\}. \tag{A.8}$$

Focusing on the second term, we have

$$P\left\{\theta^* > \hat{\theta}_{\widehat{p}_n}^u\right\} = P\left\{\theta_0^u > \theta_0^u - \theta^* + \inf_{v\in\hat{V}^u}\left[\hat{\theta}^u(v) + \left(\widehat{b}_n^u + \hat{c}^u(\widehat{p}_n)/\widehat{a}_n^u\right)s^u(v)\right]\right\},$$

from the definition of $\hat{\theta}^u_{\widehat{p}_n}$. We can show that for some $\varepsilon > 0$ and some $\varepsilon_n \searrow 0$

$$\begin{aligned} P\left\{\theta^* > \hat{\theta}^u_{\widehat{p}_n}\right\} &\leq P\left\{\mathcal{E}^u_n(V^u) > c^u_n\left(p_n - \varepsilon\right) + a^u_n \frac{\Delta^u_n}{\bar{s}^u} + \varepsilon_n\right\} + o(1) \\ &\leq \underbrace{P\left\{\mathcal{E}^u_n(V^u) > c^u_n\left(p_n - \varepsilon\right) + a^u_n \frac{\Delta^u_n}{\bar{s}^u}\right\}}_{\mathcal{A}} + o(1). \end{aligned}$$

The first inequality follows similarly to the proof of Theorem 1, also using that $s^u \leq \bar{s}^u$, that $\widehat{p}_n = p_n + o_p(1)$ so that for any $\varepsilon > 0$, $\widehat{p}_n \geq p_n - \varepsilon$ with probability approaching one, and the assumption on τ_n. The second inequality follows from the anti-concentration property. We can conclude analogously that

$$P\left\{\theta^* < \hat{\theta}^l_{p_n}\right\} \leq \underbrace{P\left\{\mathcal{E}^l_n(V^l) > c^l_n\left(p_n - \varepsilon\right) + a^l_n \frac{\Delta^l_n}{\bar{s}^l}\right\}}_{\mathcal{B}} + o(1).$$

Thus we have that for each $\varepsilon > 0$

$$P\left\{\theta^* \notin \left[\hat{\theta}^l_{\widehat{p}_n}, \hat{\theta}^u_{\widehat{p}_n}\right]\right\} \leq \mathcal{A} + \mathcal{B} + o(1).$$

In Step 2 below we show that for each $\varepsilon > 0$, $\mathcal{A} + \mathcal{B} \leq \alpha + \varepsilon + o(1)$, so that

$$P\left\{\theta^* \notin \left[\hat{\theta}^l_{\widehat{p}_n}, \hat{\theta}^u_{\widehat{p}_n}\right]\right\} \leq \alpha + o(1).$$

This gives us the required conclusion since θ^* is an arbitrary sequence of constants within the identified set, dependent upon n.

Step 2. Let $[0, \infty]$ be the standard one-point compactification of $[0, \infty)$, endowed with the metric $d(x, y) = |\lambda(x) - \lambda(y)|$, where $\lambda(x) = 1 - \exp(-x)$. This space is compact, so that every sequence in this space has a convergent subsequence.

Here we first consider sequences along which $\tau_n \Delta_n \to c \in [0, \infty]$, and show that

$$\mathcal{A} + \mathcal{B} \leq \alpha + \varepsilon + o(1) \quad \text{if } \tau_n \Delta_n \to c \in [0, \infty] \tag{A.9}$$

Given this, we show that

$$\mathcal{A} + \mathcal{B} \leq \alpha + \varepsilon + o(1) \tag{A.10}$$

holds for every sequence by way of contradiction. Indeed, suppose that $\mathcal{A} + \mathcal{B} > \alpha + \varepsilon + \delta$ for some $\delta > 0$ along a subsequence. Then we can find a convergent subsequence in $[0, \infty]$ with respect to d. Thus, we can find a subsequence such that $\tau_k \Delta_k \to c \in [0, \infty]$ and $\mathcal{A} + \mathcal{B} > \alpha + \varepsilon + \delta$ for k large enough, which gives us a contradiction to (A.9).

We now have to show (A.9). Suppose first that $c = 0$ in (A.9), then in this case $p_n = 1 - \alpha/2 + o(1)$ and

$$\mathcal{A} \leq 1 - (p_n - \varepsilon) = \alpha/2 + \varepsilon + o(1), \quad \mathcal{B} \leq 1 - (p_n - \varepsilon) + o(1) = \alpha/2 + \varepsilon + o(1).$$

Suppose that $0 < c \leq \infty$ in (A.9), then by $\tau_n(a_n^j)^{-1}\bar{s}_j \to 0$,

$$\frac{a_n^j}{\bar{s}^j}\Delta_n = [\tau_n(a_n^j)^{-1}\bar{s}_j]^{-1}\tau_n\Delta_n \to \infty.$$

Since $\Delta_n = \Delta_n^l + \Delta_n^u$, this implies that for every subsequence there exists a further subsequence indexed by k such that (a) $a_k^u\Delta_k^u \to \infty$ or (b) $a_k^l\Delta_k^l \to \infty$. In case (a) we get $p_k = 1 - \Phi(c)\alpha + o(1)$ and

$$\mathcal{B} \leq 1 - (p_n - \varepsilon) = \Phi(c)\alpha + \varepsilon + o(1), \quad \mathcal{A} \leq P\left\{\mathcal{E}_k^u(V^u) > a_k^u\frac{\Delta_k^u}{\bar{s}^u}\right\} + o(1) = o(1);$$

in case (b) we get we get $p_k = 1 - \Phi(c)\alpha + o(1)$ and

$$\mathcal{A} \leq 1 - (p_k - \varepsilon) \leq \Phi(c)\alpha + \varepsilon + o(1), \quad \mathcal{B} \leq P\left\{\mathcal{E}_k^l(V^l) > a_k^l\frac{\Delta_k^l}{\bar{s}^l}\right\} + o(1) = o(1).$$

So we get for all such subsequences that $\mathcal{A} + \mathcal{B} \leq \alpha + \varepsilon + o(1)$. Given this, we can claim that this relation holds for every sequence by the way of contradiction. Indeed, suppose that $\mathcal{A} + \mathcal{B} > \alpha + \varepsilon + \delta$ for $\delta > 0$ along a subsequence. But since we can find at least one further subsequence along which $\mathcal{A} + \mathcal{B} > \alpha + \varepsilon + \delta$ for $\delta > 0$ holds and that also satisfies either case (a) or (b) above, we obtain a contradiction. □

Appendix B. Strong Approximations For Nonparametric Estimators

B.1. **Strong Approximations for Series Estimators.** Here we establish strong approximations for series estimators of the form considered in section 3.4.

Theorem 4 (Strong Approximation for a Generic Series Estimator)**.** *Let a_n be a sequence of constants $a_n \to \infty$. In this paper it suffices to consider $a_n = \sqrt{\log n}$. We assume the following conditions on a generic series estimation problem. (a) The series estimator $\widehat{\theta}(v)$ for the function $\theta(v)$ has the form $\widehat{\theta}(v) = p(v)'\widehat{\beta}$, where $p_n(v) := (p_1(v), \ldots, p_K(v))$ is a collection of K-dimensional approximating functions such that*

$K \to \infty$, and $\widehat{\beta}$ is a K-vector of estimates. (b) The estimator $\widehat{\beta}$ satisfies an asymptotically linear representation around some K-dimensional vector β :

$$\Omega_n^{-1/2}\sqrt{n}(\widehat{\beta}-\beta) = n^{-1/2}\Omega_n^{-1/2}Q_n^{-1}\sum_{i=1}^{n} p_n(V_i)\epsilon_i + r_n, \quad \|r_n\| = o_p(a_n^{-1}),$$

$$(V_i,\epsilon_i) \text{ are i.i.d. with } E[\epsilon_i p(V_i)] = 0, E[\epsilon_i^2 p_n(V_i)p_n(V_i)'] =: S_n,\ Q_n^{-1}S_n(Q_n^{-1})' =: \Omega_n,$$

where Q_n^{-1} is some non-random invertible matrix, which is not necessarily symmetric, and eigenvalues of S_n^{-1} are bounded above by s_n. (c) The function $\theta(v)$ admits the approximation $\theta(v) = p_n(v)'\beta + A_n(v)$, where the approximation error $A_n(v)$ satisfies $\sup_{v\in\mathcal{V}}\sqrt{n}|A_n(v)|/\|g_n(v)\| = o(a_n^{-1})$, $g_n(v) := p_n(v)'\Omega_n^{1/2}$. (d) Finally, $E[|\epsilon_i|^3]$ and $\sup_{v\in\mathcal{V}}\max_j|p_j(v)|$ are uniformly bounded in n, and $a_n^6 s_n^3 K^5/n \to 0$. Then we can find a random normal vector $\mathcal{N}_n = N(0, I_K)$ such that

$$\|\Omega_n^{-1/2}\sqrt{n}(\widehat{\beta}-\beta) - \mathcal{N}_n\| = o_p(a_n^{-1}).$$

As a consequence we obtain the following approximation for the series estimator

$$\sup_{v\in\mathcal{V}}\left|\frac{\sqrt{n}(\widehat{\theta}(v)-\theta(v))}{\|g_n(v)\|} - \frac{g_n(v)'}{\|g_n(v)\|}\mathcal{N}_n\right| = o_p(a_n^{-1}).$$

Remarks. Sufficient conditions for linear approximation (b) are well known in the literature on series estimation, e.g. Andrews (1991) and Newey (1995). Conditions imposed in (a)-(c) are rather weak. The condition on the boundedness of components p_j of the vector p is weak, and is satisfied by B-splines, trigonometric series, and a variety of other bases. As shown in the proof, the Condition (b), namely that $\sup_{v\in\mathcal{V}}\max_j|p_j(v)| < \infty$ and $s_n^3 a_n^6 K^5/n \to 0$ can be replaced by an alternative condition, which is $s_n^{3/2}a_n^3K^{5/2}\max_{v\in\mathcal{V}}\sum_{j=1}^{K}|p_j(v)|^3/n^{1/2} \to 0$, which will cover more general cases.

Proof of Theorem 4. The proof has two steps: in the first, we couple the estimator $\sqrt{n}(\widehat{\beta}-\beta)$ with the normal vector; in the second, we establish the strong approximation for the series estimate of the function.

Step 1. He we shall apply Yurinskii coupling, see Yurinskii (1977) and Pollard (2002) (page 244).

Let $\xi_1, ..., \xi_n$ be independent K-vectors with $E\xi_i = 0$ for each i, and $\Delta := \sum_i E\|\xi_i\|^3$ finite. Let $S = \xi_1 + ... + \xi_n$. For each $\delta > 0$ there exists a random vector T with a

$N(0, \text{var } (S))$ distribution such that

$$P\{\|S-T\| > 3\delta\} \leq C_0 B \left(1 + \frac{|\log(1/B)|}{K}\right) \text{ where } B := \Delta K \delta^{-3},$$

for some universal constant C_0.

In order to apply the coupling, consider

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i, \quad \xi_i = \Omega_n^{-1/2} Q_n^{-1} p_n(V_i)\epsilon_i \sim (0, I_K),$$

Then we have that

$$\begin{aligned}
E\|\xi_i\|^3 &\leq \text{ maxeig } (S_n^{-1})^{3/2} \cdot E\|p_n(V_i)\epsilon_i\|^3 \\
&= s_n^{3/2} \cdot K^{3/2} E\left(\epsilon_i^2 \frac{1}{K} \sum_{j=1}^{K} p_{nj}(V_i)^2\right)^{3/2} \\
&\leq s_n^{3/2} \cdot K^{3/2} \max_{v\in\mathcal{V}} \frac{1}{K} \sum_{j=1}^{K} |p_{nj}(v)|^3 E|\epsilon_i|^3 \\
&\leq s_n^{3/2} \cdot K^{3/2} \max_{j} \sup_{v\in\mathcal{V}} |p_{nj}(v)|^3 E|\epsilon_i|^3 \\
&\lesssim s_n^{3/2} K^{3/2},
\end{aligned}$$

using the assumption that $E|\epsilon_i|^3$ and $\max_j \sup_{v\in\mathcal{V}} |p_{nj}(v)|$ are uniformly bounded in n. Therefore, by Yurinskii's coupling, for each $\delta > 0$

$$\begin{aligned}
P\left\{\left|\frac{\sum_{i=1}^n \xi_i}{\sqrt{n}} - \mathcal{N}_n\right| \geq 3\delta a_n^{-1}\right\} &\lesssim \frac{nK s_n^{3/2} K^{3/2}}{(\delta a_n^{-1}\sqrt{n})^3} \\
&= \frac{a_n^3 s_n^{3/2} K^{5/2}}{(\delta n^{1/2})} \to 0, \\
&\quad \text{by } (a_n)^6 s_n^3 K^5/n \to 0.
\end{aligned}$$

This proves the first part of the lemma. Also, to justify the remark given after the lemma, we have that

$$E\|\xi_i\|^3 \lesssim s_n^{3/2} K^{3/2} \max_{v\in\mathcal{V}} \frac{1}{K} \sum_{j=1}^{K} |p_j(v)|^3 E|\epsilon_i|^3.$$

Therefore, by Yurinskii's coupling, for each $\delta > 0$

$$P\left\{\left|\frac{\sum_{i=1}^n \xi_i}{\sqrt{n}} - \mathcal{N}_n\right| \geq 3\delta a_n^{-1}\right\} \lesssim \frac{s_n^{3/2}(a_n)^3 n K^{5/2} \max_{v\in\mathcal{V}} \frac{1}{K}\sum_{j=1}^K |p_j(v)|^3 E|\epsilon_i|^3}{(\delta\sqrt{n})^3} \to 0,$$

$$\text{if} \quad s_n^{3/2}(a_n)^3 K^{5/2} \max_{v\in\mathcal{V}} \sum_{j=1}^K |p_j(v)|^3/n^{1/2} \to 0.$$

Finally by combining the preceding step with the assumption on the linearization error r_n, we obtain

$$\begin{aligned}\|\Omega_n^{-1/2}\sqrt{n}(\widehat{\beta}-\beta) - \mathcal{N}_n\| &\leq \|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i - \mathcal{N}_n\| + \|\Omega_n^{-1/2}\sqrt{n}(\widehat{\beta}-\beta) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\| \\ &= o_p(a_n^{-1}) + r_n = o_p(a_n^{-1}).\end{aligned}$$

Step 2. Using the result of Step 1 and that

$$\frac{\sqrt{n}p(v)'(\widehat{\beta}-\beta)}{\|g_n(v)\|} = \frac{\sqrt{n}g_n(v)'\Omega_n^{-1/2}(\widehat{\beta}-\beta)}{\|g_n(v)\|}$$

we conclude that

$$\begin{aligned}|S_n(v)| &:= \Big|\frac{\sqrt{n}g_n(v)'\Omega_n^{-1/2}(\widehat{\beta}-\beta)}{\|g_n(v)\|} - \frac{g_n(v)'\mathcal{N}_n}{\|g_n(v)\|}\Big| \\ &\leq \Big\|\sqrt{n}\Omega_n^{-1/2}(\widehat{\beta}-\beta) - \mathcal{N}_n\Big\| = o_p(a_n^{-1}),\end{aligned} \tag{B.1}$$

uniformly in $v \in \mathcal{V}$. Finally,

$$\begin{aligned}&\sup_{v\in\mathcal{V}}\Big|\frac{\sqrt{n}(\widehat{\theta}(v)-\theta(v))}{\|g_n(v)\|} - \frac{g_n(v)'\mathcal{N}_n}{\|g_n(v)\|}\Big| \\ &\quad\leq \sup_{v\in\mathcal{V}}\Big|\frac{\sqrt{n}(\widehat{\theta}(v)-\theta(v))}{\|g_n(v)\|} - \frac{\sqrt{n}g_n(v)'\Omega_n^{-1/2}(\widehat{\beta}-\beta)}{\|g_n(v)\|}\Big| \\ &\quad+ \sup_{v\in\mathcal{V}}\Big|\frac{\sqrt{n}g_n(v)'\Omega_n^{-1/2}(\widehat{\beta}-\beta)}{\|g_n(v)\|} - \frac{g_n(v)'\mathcal{N}_n}{\|g_n(v)\|}\Big| \\ &\quad= \sup_{v\in\mathcal{V}}|\sqrt{n}A_n(v)/\|g_n(v)\|| + \sup_{v\in\mathcal{V}}|S_n(v)| = o_p(a_n^{-1}) + o_p(a_n^{-1}),\end{aligned}$$

using the assumption on the approximation error $A_n(v) = \theta(v) - p_n(v)'\beta$ and the bound (B.1). □

B.2. **Strong Approximations for Kernel-Type Estimators.** This section provides low-level sufficient conditions for K.1 and K.2. In particular, we focus on a case when

a bound-generating function $\theta(\cdot)$ is estimated by a kernel-type estimator of conditional expectation functions. Let $F_{U|V}(\cdot|v)$ denote the cumulative distribution function of U given $V = v$.

Theorem 5. *Assume that (1) the joint distribution of (U, V) is absolutely continuous with respect to Lebesgue measure on $[0,1]^{d+1}$; (2) $f_V(v)$ and $\sigma^2(v)$ are Lipschitz continuous and bounded away from zero on their support $[0,1]^d$; (3) $\sigma(v)$ is continuously differentiable and its derivative is bounded; (4) $F^{-1}_{U|V}(\varepsilon|v)$ is bounded uniformly in (ε, v) and its partial derivatives with respect to ε and v are also uniformly bounded; (5) as $n \to \infty$, the kernel estimator of $\theta(v)$ has an asymptotic linear expansion:*

$$(nh_n^d)^{1/2}(\widehat{\theta}(v) - \theta(v)) = \frac{1}{(nh_n^d)^{1/2} f_V(v)} \sum_{i=1}^{n} \sigma(V_i) U_i \mathbf{K}\left(\frac{v - V_i}{h_n}\right) + R_n(v),$$

where $\mathbf{K}$ is a d-dimensional kernel function with compact support $[-1,1]^d$, $\int \mathbf{K}(u)du = 1$, and is twice continuously differentiable, h_n is a sequence of bandwidths that converges to zero, and the remainder term satisfies

$$\sup_{v \in \mathcal{V}} |R_n(v)| = o_p(a_n^{-1});$$

(6) Further, assume that

$$\frac{nh_n^d}{a_n (\log n)^2} \to \infty \quad \textit{and} \quad \frac{a_n \log n}{n^{1/(d+1)} h_n} \to 0.$$

Then there exists a sequence of Gaussian processes $G_n(\cdot)$, indexed by $\mathcal{V}$, with continuous sample paths and with

$$E[G_n(v)] = 0 \quad \textit{for } t \in \mathcal{V},$$
$$E[G_n(v_1) G_n(v_2)] = E[\phi_{h_n, v_1}(U, V) \phi_{h_n, v_2}(U, V)]$$

for v_1 and $v_2 \in \mathcal{V}$, such that

$$\sup_{v \in \mathcal{V}} \left| \frac{1}{(nh_n^d)^{1/2} f_V(v)} \sum_{i=1}^{n} \sigma(V_i) U_i \mathbf{K}\left(\frac{v - V_i}{h_n}\right) - \frac{G_n(v)}{h_n^{d/2} f_V(v)} \right|$$
$$= O\left[n^{-1/(2d+2)} \left(h_n^{-1} \log n\right)^{1/2} + (nh_n^d)^{-1/2} \log n \right] \quad a.s.$$

Condition (1) assumes that (U, V) are continuous random variables with support on the unit cube. There is no loss of generality by restricting the support to be the unit cube, provided that the support is known and is a Cartesian product of compact connected intervals. The bounded support assumption on U is standard in settings with

partial identification. Otherwise, the bound may not exist. Conditions (2)-(4) are mild smoothness conditions. Condition (5) provides standard regularity conditions for kernel estimation. This holds for kernel mean regression estimators and also local polynomial estimators under fairly general conditions. One important restriction that is implicit in the asymptotic expansion is that the asymptotic bias is negligible. This could be achieved by undersmoothing, which would prevent us from using optimal bandwidths. Alternatively, one could use a bias corrected one-sided confidence intervals. Condition (6) ensures that

$$\sup_{v\in\mathcal{V}}\left|\frac{1}{(nh_n^d)^{1/2}f_V(v)}\sum_{i=1}^n\sigma(V_i)U_i\mathbf{K}\left(\frac{v-V_i}{h_n}\right)-\frac{G_n(v)}{h_n^{d/2}f_V(v)}\right|=o\left(a_n^{-1}\right)\quad a.s.$$

Proof. To prove this theorem, we use Theorem 1.1 of Rio (1994). Define $\varepsilon = F_{U|V}(U|V)$. For any positive h, define $\phi_{h,s}(\varepsilon, v) = \sigma(v)F^{-1}_{U|V}(\varepsilon|v)K_d\left[h^{-1}(s-v)\right]$. For any real numbers a and b satisfying $0 < a < b \leq 1$, let $\mathcal{K}_{a,b}$ be a class of functions

$$\mathcal{K}_{a,b} = \{\phi_{h,s}(\cdot,\cdot) : s \in \mathbb{R}^d, h \in [a,b]\}.$$

First, it is standard to show that $\mathcal{K}_{h/4,h}$ is a VC class of functions for each h. Second, the UBV (uniformly of bounded variation) and LUBV (locally UBV) conditions of Rio (1994) are satisfied. To see this, first note that for some universal constant $C < \infty$,

$$\int\int\left|\frac{\partial\phi_{h,s}(\varepsilon,v)}{\partial\varepsilon}\right|+\sum_{j=1}^d\left|\frac{\partial\phi_{h,s}(\varepsilon,v)}{\partial v^{(j)}}\right|d\varepsilon dv \leq Ch^{d-1},$$

where $v^{(j)}$ is the j-th element of v. Furthermore, as in equation (4.1) of Rio (1994), note that for some universal constant $\tilde{C} < \infty$,

$$\int\int_{(\varepsilon,v)\in\mathbb{C}(\eta)}\left|\frac{\partial\phi_{h,s}(\varepsilon,v)}{\partial\varepsilon}\right|+\sum_{j=1}^d\left|\frac{\partial\phi_{h,s}(\varepsilon,v)}{\partial v^{(j)}}\right|d\varepsilon dv \leq \tilde{C}h^{-1}\min(\eta h^d, \eta^{d+1}),$$

where $\mathbb{C}(\eta)$ is a tube in $\mathbb{R}^{d+1}$ with edges of length η. Then Theorem 1.1 of Rio (1994) gives the following:

$$\sup_{v\in\mathcal{V}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\phi_{h_n,v}(U_i,V_i)-G_n(v)\right|=O\left[n^{-1/(2d+2)}\left(h_n^{d-1}\log n\right)^{1/2}+n^{-1/2}\log n\right]\quad a.s. \tag{B.2}$$

Since the density of $f_V(v)$ is bounded away from zero, Theorem 5 follows immediately from (B.2). □

Appendix C. Implementation

In this Appendix we describe implementation of our procedure. We begin by detailing the steps required for parametrically estimated bound-generating functions, and then describe implementation for nonparametric cases. Finally, we describe how one-sided bands for upper and lower bounds on θ^* can be combined to perform inference on either Θ_I or θ^*.

Note that below we focus on the upper bound, but if instead θ_0 were the lower bound for θ^*, given by the supremum of a bound-generating function, the algorithm would be entirely symmetric.[10]

C.1. **Parametric Boundary Estimation.** We start by considering implementation when the bound-generating function is estimated parametrically, i.e. where conditions P.1 and P.2 hold. We provide a simple approach that relies on simulation from the multivariate normal distribution:

(1) Compute a consistent set estimate $\widehat{V}$ for the minimizing set V_0:
$$\widehat{V} = \{v \in \mathcal{V} : \widehat{\theta}(v) \leq \inf_{v \in \mathcal{V}} \widehat{\theta}(v) + \ell_n c_n\}$$
with $\ell_n = 2\sqrt{\log n} \cdot \sup_{v \in \mathcal{V}} s(v)$ and $c_n = 1$.
(2) For each $v \in \widehat{V}$, compute $\widehat{g}(v) = \partial\theta(v, \widehat{\gamma}) / \partial\gamma \cdot \widehat{\Omega}^{1/2}$, where $\widehat{\Omega}$ is a consistent estimator for the asymptotic variance of $\sqrt{n}(\widehat{\gamma} - \gamma_0)$.
(3) Simulate a large number R of draws from $\mathcal{N}(0, I_K)$, denoted $Z_1, ..., Z_R$, where $K = \dim(\gamma)$ and I_K is the identity matrix, and compute $\widehat{k}(p) = p$-quantile of $\{\max_{v \in \widehat{V}} \left(\widehat{g}(v)' Z_r / \|\widehat{g}(v)\|\right), r = 1, ..., R\}$.
(4) Compute $\widehat{\theta}_p = \min_{v \in \widehat{V}} [\widehat{\theta}(v) + \widehat{k}(p) s(v)]$. Selecting $p = 1/2$ provides a median-unbiased estimator for θ_0, while selecting $p = 1 - \alpha$ provides a one-sided confidence interval such that $P(\theta_0 \leq \widehat{\theta}_p) = 1 - \alpha$.

An important special case is when the support of v is finite, so that $\mathcal{V} = \{v_1, ...v_J\}$. In this case, the algorithm above applies where $\theta(v, \gamma) = \sum_{j=1}^{J} \gamma_j 1[v = v_j]$, i.e. where for each j, $\theta(v_j, \gamma) = \gamma_j$ and $\widehat{g}(v) = (1[v = v_1], ..., 1[v = v_J]) \cdot \widehat{\Omega}^{1/2}$.

[10] Specifically, the steps below would apply with the following two modifications. First, the set estimate $\widehat{V}_\epsilon$ in step 1 would be given by $\widehat{V}_\epsilon = \{v \in \mathcal{V} : \widehat{\theta}(v) \geq \sup_{v \in \mathcal{V}} \widehat{\theta}(v) - \ell_n c_n - \epsilon\}$. Second, one would *subtract*, rather than add, a precision adjustment from the analog estimates for the lower bound in step (4), and then compute the *maximum* after applying this precision-adjustment, i.e. $\widehat{\theta}_p = \max_{v \in \widehat{V}} [\widehat{\theta}(v) - \widehat{k}(p) s(v)]$. Note that now $\widehat{k}(p)$ approximates the p-quantile of $\max_{v \in \widehat{V}} [\widehat{\theta}(v) - \theta(v)]/s(v)$. However, no changes need to made to the computation of $\widehat{k}(p)$ due to the symmetry of the normal distribution.

C.2. **Nonparametric Boundary Estimation.** Here we generalize the previous procedure to nonparametric series and kernel boundary estimators. The basic steps are the same, though some adjustments are necessary. In particular, the set estimator in the first step, $\widehat{V}_\epsilon$ will converge to V_ϵ, which is generally not equal to V_0, but contains V_0 with probability approaching 1. Setting $\epsilon = 0$ may also be feasible, but this implicitly puts more stringent growth restrictions on the number of series terms and bandwidth, which may be difficult to verify in practice.

C.2.1. *Series Estimators.* In practice, implementation with a series estimator does not substantially differ from the parametric case:

(1) Compute a consistent estimate $\widehat{V}_\epsilon$ for V_ϵ:

$$\widehat{V}_\epsilon = \{v \in \mathcal{V} : \widehat{\theta}(v) \leq \inf_{v \in \mathcal{V}} \widehat{\theta}(v) + \ell_n c_n + \epsilon\}$$

with $\ell_n = 2\sqrt{\log n} \cdot \sup_{v \in \mathcal{V}} s(v)$ and $c_n = \sqrt{\log n}$.

(2) For each $v \in \widehat{V}_\epsilon$, compute $\widehat{g}(v) = p_n(v)' \widehat{\Omega}^{1/2}$, where $\widehat{\Omega}$ is a consistent estimate of asymptotic variance of $\hat{\beta}$.

(3) Simulate a large number R of draws from $N(0, I_K)$, denoted $Z_1, ..., Z_R$. Compute $\widehat{k}(p) = p$-quantile of $\{\max_{v \in \widehat{V}} \left(\widehat{g}(v)' Z_r / \|\widehat{g}(v)\|\right), r = 1, ..., R\}$.

(4) Compute $\widehat{\theta}_p = \min_{v \in \widehat{V}_\epsilon} [\widehat{\theta}(v) + \widehat{k}(p) s(v)]$. Selecting $p = 1/2$ provides a median-unbiased estimator for θ_0, while selecting $p = 1 - \alpha$ provides a one-sided confidence interval such that $P(\theta_0 \leq \widehat{\theta}_p) = 1 - \alpha$.

We can also bypass simulation of the stochastic process by employing expansion (A.2) in the proof of Lemma 4 in Appendix A. This choice of $\widehat{k}(p)$ is convenient because it does not involve simulation; however, it could be too conservative in some applications. Thus, we recommend using simulation in applications, unless the computational cost is too high.

C.2.2. *Kernel Estimators.* The steps are as follows:

(1) Compute a consistent estimate $\widehat{V}_\epsilon$ for V_ϵ, as given in (3.14), e.g.

$$\widehat{V}_\epsilon = \{v \in \mathcal{V} : \widehat{\theta}(v) \leq \inf_{v \in \mathcal{V}} \widehat{\theta}(v) + \ell_n c_n + \epsilon\}$$

with $\ell_n = 2\sqrt{\log n} \cdot \sup_{v \in \mathcal{V}} s(v)$ and $c_n = \sqrt{\log n}$.

(2) For each $v \in \widehat{V}_\epsilon$, compute $\omega_n(v)$ as given in condition K.1, using consistent sample analog estimators.

(3) Simulate a large number R of draws from $N(0, I_n)$, denoted $Z_1, ..., Z_n$. Compute $\widehat{k}(p) = p$-quantile of $\{\max_{v \in \widehat{V}} \left(\omega_n(v)' Z_r / \|\omega_n(v)\|\right), r = 1, ..., R\}$.

(4) Compute $\widehat{\theta}_p = \min_{v \in \widehat{V}_\epsilon} [\widehat{\theta}(v) + \widehat{k}(p) s(v)]$. Selecting $p = 1/2$ provides a median-unbiased estimator for θ_0, while selecting $p = 1 - \alpha$ provides a one-sided confidence interval such that $P(\theta_0 \leq \widehat{\theta}_p) = 1 - \alpha$.

The researcher also has the option of employing an analytical approximation in place of simulation if desired. Such critical values are provided by (3.10), (3.11), and (3.12), all of which are asymptotically equivalent.

References

Andrews, D. W. K. (1991): "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59(2), 307–345.

Andrews, D. W. K., and S. Han (2009): "Invalidity of the Bootstrap and m Out of n Bootstrap for Interval Endpoints Defined by Moment Inequalities," *Econometrics Journal*, 12(s1), S172–S199.

Andrews, D. W. K., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," working paper, Cowles Foundation.

Andrews, D. W. K., and X. Shi (2009): "Inference for Parameters Defined by Conditional Moment Inequalities," working paper, Cowles Foundation.

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

Berry, S. T., and E. Tamer (2007): "Identification in Models of Oligopoly Entry," in *Advances in Econometrics, Ninth World Congress*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 2, pp. 46–85. Cambridge University Press.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Unemployment Composition Using Bounds," *Econometrica*, 75(2), 323–363.

Bugni, F. (2009): "Bootstrap Inference in Partially Identified Models," working paper, Duke University.

Canay, I. (2009): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," Northwestern University.

Carneiro, P., and S. Lee (2009): "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality," *Journal of Econometrics*, 149, 191–208.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Chernozhukov, V., R. Rigobon, and T. Stoker (2007): "Set Identification with Tobin Regressors," working paper, MIT.

Chesher, A. D. (2007): "Endogeneity with Discrete Outcomes," CEMMAP working paper CWP 05/07.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London, UK.

Fan, Y. (2009): "Confidence Sets for Distributions of Treatment Effects with Covariates," working paper, Vanderbilt University.

Galichon, A., and M. Henry (2009): "A Test of Non-identifying Restrictions and Confidence Regions for Partially Identified Parameters," *Journal of Econometrics*, forthcoming.

Ghosal, S., A. Sen, and A. W. van der Vaart (2000): "Testing Monotonicity of Regression," *Annals of Statistics*, 28, 1054–1082.

Gonzalez, L. (2005): "Nonparametric Bounds on the Returns to Language Skills," *Journal of Applied Econometrics*, 20, 771–795.

Haile, P. A., and E. Tamer (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 111(1), 1–51.

Härdle, W., and O. Linton (1994): "Applied Nonparametric Methods," in *The Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, pp. 2295–2339. North-Holland.

Imbens, G., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Kim, K. i. (2009): "Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities," working paper, University of Minnesota.

Kreider, B., and J. Pepper (2007): "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102(478), 432–441.

Lee, S., O. Linton, and Y.-J. Whang (2009): "Testing for Stochastic Monotonicity," *Econometrica*, 77(2), 585–602.

Lee, S., and R. Wilke (2009): "Reform of Unemployment Compensation in Germany: A Nonparametric Bounds Analysis Using Register Data," *Journal of Business and Economic Statistics*, 27(2), 193–205.

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24(3), 343–360.

——— (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80(2), 319–323.

——— (1994): "The Selection Problem," in *Advances in Econometrics, Sixth World Congress*, ed. by C. Sims, vol. 1, pp. 143–170. Cambridge University Press.

——— (1997): "Monotone Treatment Response," *Econometrica*, 65(6), 1311–1334.

——— (2003): *Partial Identification of Probability Distributions*. Springer-Verlag, New York.

Manski, C. F., and J. Pepper (1998): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *NBER working paper*.

Manski, C. F., and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997–1010.

——— (2009): "More on Monotone Instrumental Variables," *Econometrics Journal*, 12(s1), S200–S216.

Menzel, K. (2009): "Estimation and Inference with Many Weak Moment Inequalities," working paper, MIT.

Nevo, A., and A. M. Rosen (2008): "Identification with Imperfect Instruments," CEMMAP working paper CWP16/08.

Newey, W. K. (1995): "Convergence Rates for Series Estimators," in *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao*, ed. by G. Maddalla, P. Phillips, and T. Srinavasan, pp. 254–275. Blackwell, Cambridge, U.S.A.

Nicoletti, C., F. Foliano, and F. Peracchi (2007): "Estimating Income Poverty in The Presence Of Missing Data and Measurement Error Problems," ISER Working Papers, WP 2007-15.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2005): "The Method of Moments with Inequality Constraints," working paper, Harvard University.

Piterbarg, V. I. (1996): *Asymptotic Methods in the Theory of Gaussian Processes and Fields*. American Mathematical Society, Providence, RI.

Pollard, D. (2002): *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, Cambridge.

Rio, E. (1994): "Local Invariance Principles and Their Application to Density Estimation," *Probability Theory and Related Fields*, 98, 21–45.

Romano, J. P., and A. M. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.

——— (2009): "Inference for the Identified Set in Partially Identified Econometric Models," working paper, Stanford University and University of Chicago.

Rosen, A. M. (2008): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107–117.

Rudelson, M., and R. Vershynin (2007): "Anti-Concentration Inequalities," Conference Proceeding for Phenomena in High Dimensions, Samos, Greece, 2007.

Rudelson, M., and R. Vershynin (2008): "The Littlewood-Offord problem and invertibility of random matrices," *Adv. Math.*, 218(2), 600–633.

Stoye, J. (2009): "More on Confidence Regions for Partially Identified Parameters," *Econometrica*, forthcoming.

Tao, T., and V. Vu (2009): "From the Littlewood-Offord problem to the Circular Law: Universality of the spectral distribution of random matrices," *Bull. Amer. Math. Soc.*, 46, 377–396.

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

Yurinskii, V. (1977): "On the Error of the Gaussian Approximation for Convolutions," *Theory of Probability and Its Applications*, 22, 236–247.

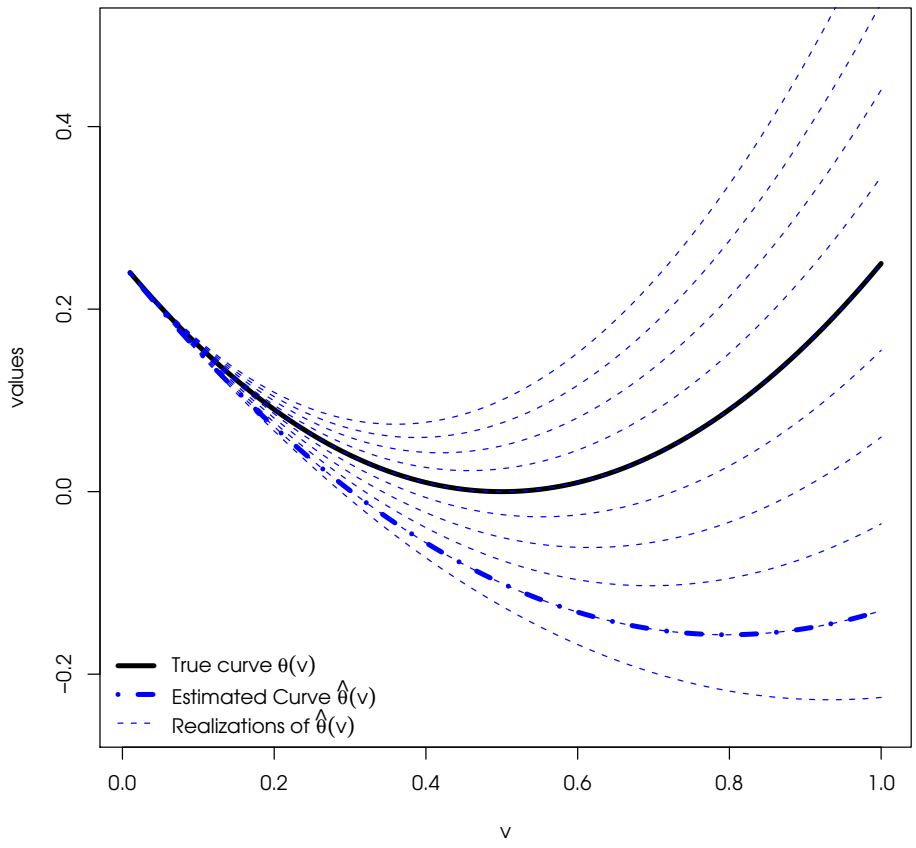

FIGURE 1. This figure illustrates how variation in the precision of the analog estimator at different points may impede inference. The solid curve is the true bound-generating function $\theta(v)$, while the dash-dot curve is a single realization of its estimator, $\widehat{\theta}(v)$. The lighter dashed curves depict eight additional representative realizations of the estimator, illustrating its precision at different values of v. The minimum of the estimator $\widehat{\theta}(v)$ is indeed quite far from the minimum of $\theta(v)$, making the empirical upper bound unduly tight.

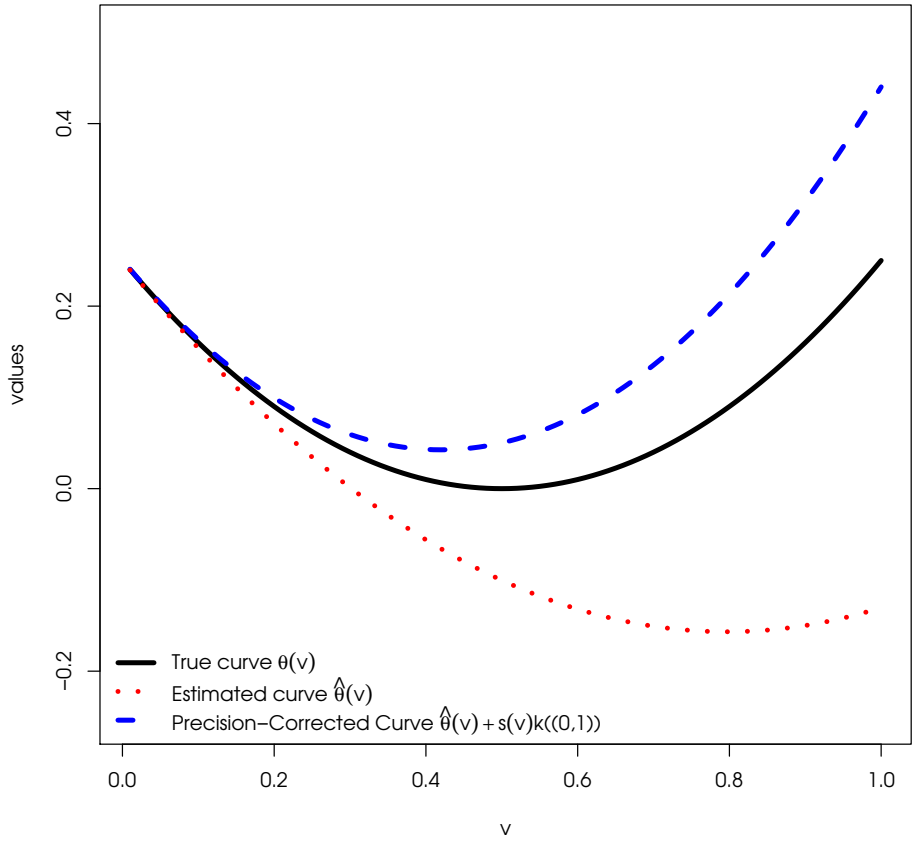

FIGURE 2. This figure depicts a precision-corrected curve (dashed curve) that adjusts the boundary estimate $\widehat{\theta}(v)$ (dotted curve) by an amount proportional to its point-wise standard error. The minimum of the precision-corrected curve is closer to the minimum of the true curve (solid) than the minimum of $\widehat{\theta}(v)$, removing the downward bias.

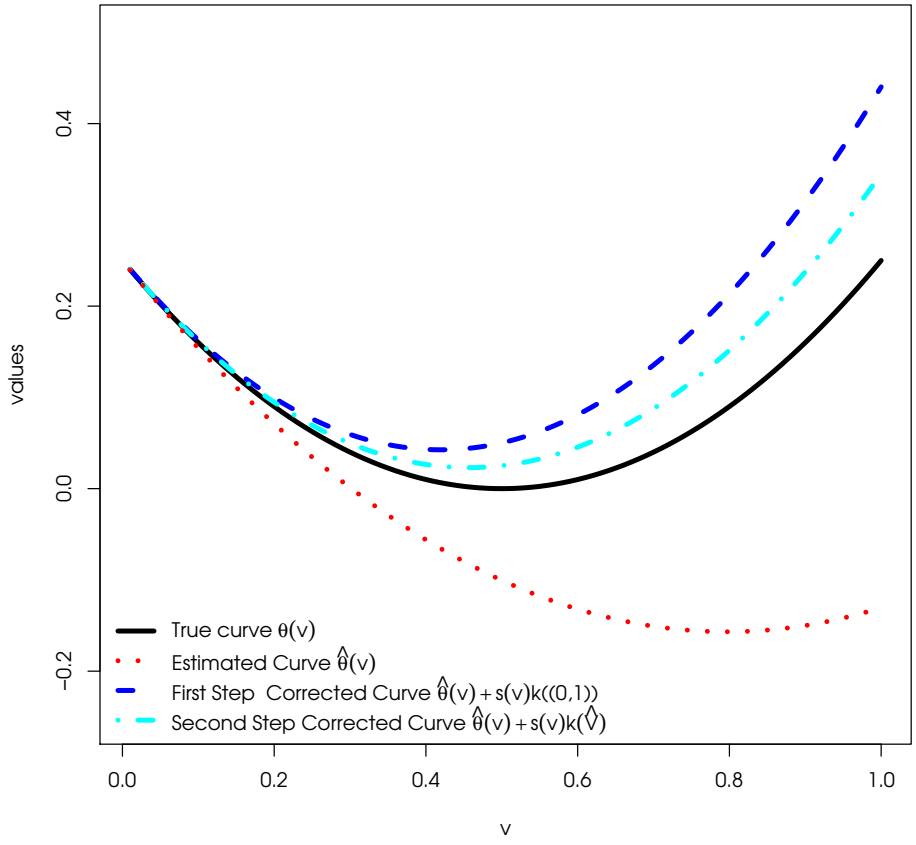

FIGURE 3. This figure depicts a precision-corrected curve (dashed curve) that adjusts the boundary estimate $\widehat{\theta}(v)$ (dotted curve) by an amount proportional to its point-wise standard error. The dash-dot curve represents an improvement on the precision-corrected curve obtained by employing an estimator for the set of minimizing values. The minimum of this dash-dotted curve is closer to the minimum of $\theta(v)$ than the initial precision-corrected curve.

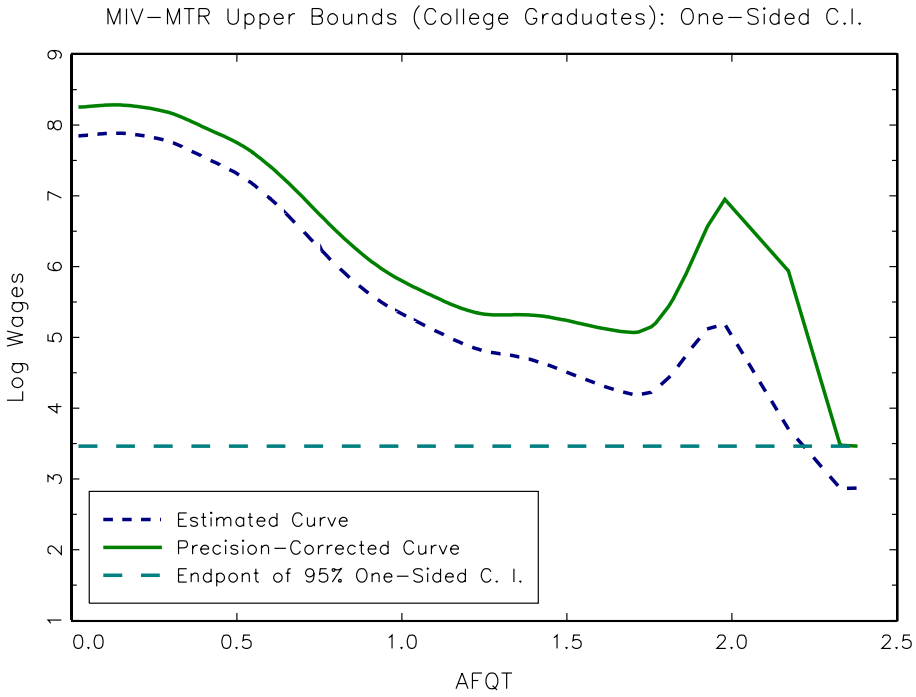

FIGURE 4. This figure provides the estimated upper bound on the log wages for college graduates. The minimum of the estimated boundary function (dashed curve) occurs in the right-tail of the distribution, where the curve is less precisely estimated. The estimate may therefore not provide an accurate representation of the true boundary function in this region. Our method employs the precision-corrected curve (solid curve) to account for varying levels of precision of the estimate.

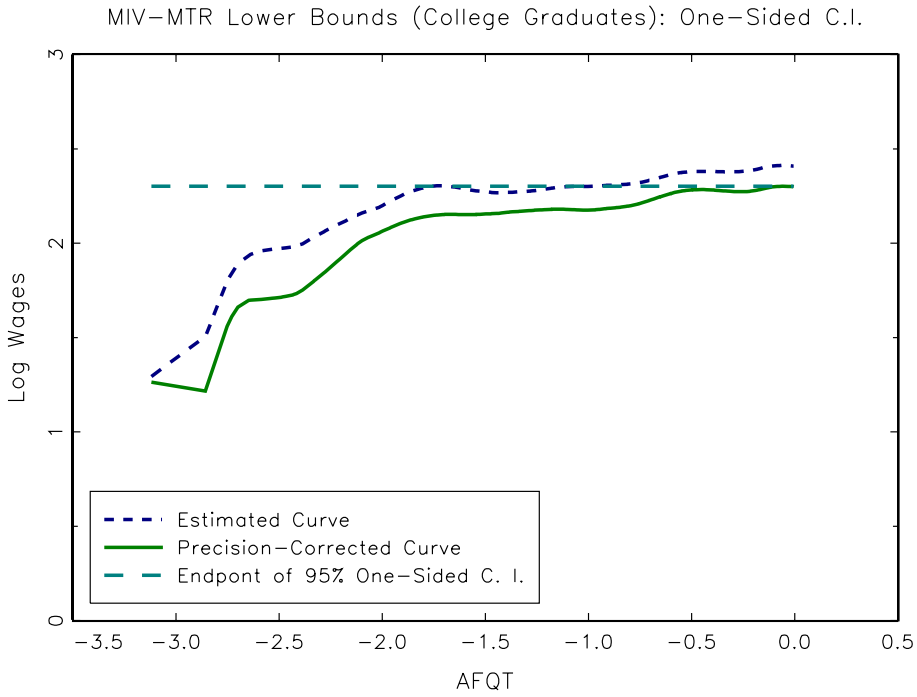

FIGURE 5. This figure provides the estimated lower bound on the log wages for college graduates. The maximum of the estimated boundary function (dashed curve) is in a region where it is relatively precisely estimated. The maximum of the precision-corrected curve (solid curve) is therefore quite near the maximum of the estimated curve, though the latter is slightly higher.

TABLE 1. Results for Monte Carlo Experiments [1,000 replications per experiment]

DGP	Sample Size	Average Smoothing Parameter	Estimating V?	Method	Mean Bias	Median Bias	SD	MAD	RMSE	Cov. Prob. 0.50	Cov. Prob. 0.95
Series Estimation											
1	500	8.949	No	Analog	0.255	0.238	0.137	0.256	0.290		
				New	0.007	0.000	0.096	0.074	0.096	0.496	0.958
1	500	8.916	Yes	Analog	0.248	0.227	0.132	0.248	0.280		
				New	0.002	-0.006	0.096	0.075	0.096	0.531	0.965
1	1000	9.716	No	Analog	0.187	0.179	0.091	0.187	0.208		
				New	0.003	-0.002	0.069	0.054	0.069	0.510	0.955
1	1000	9.696	Yes	Analog	0.189	0.177	0.099	0.189	0.213		
				New	0.003	0.000	0.071	0.055	0.071	0.501	0.965
2	500	9.313	No	Analog	0.172	0.176	0.211	0.221	0.272		
				New	-0.171	-0.159	0.214	0.220	0.274	0.782	0.978
2	500	9.372	Yes	Analog	0.164	0.159	0.214	0.216	0.270		
				New	-0.136	-0.127	0.250	0.227	0.284	0.696	0.953
2	1000	10.430	No	Analog	0.140	0.142	0.159	0.172	0.212		
				New	-0.134	-0.129	0.166	0.173	0.213	0.796	0.974
2	1000	10.440	Yes	Analog	0.144	0.147	0.162	0.177	0.217		
				New	-0.064	-0.053	0.178	0.150	0.189	0.626	0.942
Local Linear Estimation											
1	500	0.584	No	Analog	0.208	0.192	0.119	0.209	0.240		
				New	0.012	0.001	0.088	0.067	0.088	0.491	0.943
1	500	0.584	Yes	Analog	0.208	0.192	0.119	0.209	0.240		
				New	0.012	0.001	0.088	0.067	0.088	0.491	0.943
1	1000	0.548	No	Analog	0.153	0.141	0.081	0.153	0.173		
				New	0.007	0.004	0.061	0.048	0.061	0.478	0.951
1	1000	0.548	Yes	Analog	0.153	0.141	0.081	0.153	0.173		
				New	0.007	0.004	0.061	0.048	0.061	0.478	0.951
2	500	0.324	No	Analog	0.165	0.163	0.220	0.222	0.276		
				New	-0.242	-0.248	0.220	0.275	0.327	0.864	0.979
2	500	0.324	Yes	Analog	0.166	0.163	0.221	0.222	0.276		
				New	-0.198	-0.203	0.237	0.254	0.309	0.804	0.970
2	1000	0.266	No	Analog	0.151	0.142	0.164	0.179	0.223		
				New	-0.187	-0.195	0.163	0.211	0.248	0.862	0.984
2	1000	0.266	Yes	Analog	0.151	0.143	0.165	0.179	0.224		
				New	-0.126	-0.134	0.168	0.175	0.210	0.775	0.970

Notes: The "Analog" and "New" methods refer to the sample analog method and our new proposed method. For each method, we report the mean and median biases, standard deviation (SD), mean absolute deviation (MAD), root mean squared error (RMSE), and empirical coverage probabilities at 50% and 95% levels.

TABLE 2. Descriptive Statistics ($n = 2044$)

Variable	Mean	Median	Std. dev.	Minimum	Maximum
Log hourly wages (Y)	2.54	2.50	0.58	0.28	9.06
Years of schooling (Z)	13.43	12.00	2.56	5.00	20.00
AFQT score (V)	0.00	0.12	0.99	-3.12	2.38

TABLE 3. Estimation Results

	High School Graduates		College Graduates	
Estimation method	Lower bound	Upper bound	Lower bound	Upper bound
Naïve sample analog estimator	2.12	2.75	2.41	2.87
New estimator	2.03	2.84	2.35	3.18
95% confidence interval	1.97	2.88	2.31	3.44

DEPARTMENT OF ECONOMICS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY; DEPARTMENT OF ECONOMICS, UNIVERSITY COLLEGE LONDON; DEPARTMENT OF ECONOMICS, UNIVERSITY COLLEGE LONDON