

Chernozhukov, Victor; Lee, Sokbae; Rosen, Adam

Working Paper

Intersection bounds: Estimation and inference

cemmap working paper, No. CWP33/12

Provided in Cooperation with:

Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Lee, Sokbae; Rosen, Adam (2012) : Intersection bounds: Estimation and inference, cemmap working paper, No. CWP33/12, Centre for Microdata Methods and Practice (cemmap), London,
<https://doi.org/10.1920/wp.cem.2012.3312>

This Version is available at:

<https://hdl.handle.net/10419/79516>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Intersection bounds: estimation and inference

Victor Chernozhukov
Sokbae Lee
Adam Rosen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP33/12

INTERSECTION BOUNDS: ESTIMATION AND INFERENCE

VICTOR CHERNOZHUKOV, SOKBAE LEE, AND 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. We show that many bounds characterizations in econometrics, for instance bounds on parameters under conditional moment inequalities, can be formulated as intersection bounds. Our approach is especially convenient for 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 size of the identified set, we also offer a median-bias-corrected estimator of such bounds as a by-product of our inferential procedures. We develop theory for large sample inference based on the strong approximation of a sequence of series or kernel-based empirical processes by a sequence of “penultimate” Gaussian processes. These penultimate processes are generally not weakly convergent, and thus non-Donsker. Our theoretical results establish that we can nonetheless perform asymptotically valid inference based on these processes. Our construction also provides new adaptive inequality/moment selection methods. We provide conditions for the use of nonparametric kernel and series estimators, including a novel result that establishes strong approximation for any general series estimator admitting linearization, which may be of independent interest.

KEY WORDS. Bound analysis, conditional moments, partial identification, strong approximation, infinite dimensional constraints, linear programming, concentration inequalities, anti-concentration inequalities, non-Donsker empirical process methods, moderate deviations, adaptive moment selection.

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

Date: September 1, 2012. First version: December 2007.

We are especially grateful to D. Chetverikov, K. Kato, Y. Luo, A. Santos, five anonymous referees, and the editor for making several extremely useful suggestions that have led to substantial improvements. We thank T. Armstrong, R. Blundell, A. Chesher, F. Molinari, W. Newey, C. Redmond, N. Roys, S. Stouli, and J. Stoye for detailed discussion and suggestions, and participants at numerous seminars and conferences for their comments. This paper is a revised version of “Inference on Intersection Bounds” initially presented and circulated at the University of Virginia and the Harvard/MIT econometrics seminars in December 2007, and presented at the March 2008 CEMMAP/Northwestern conference on “Inference in Partially Identified Models with Applications.” We gratefully acknowledge financial support from the National Science Foundation, Economic and Social Research Council (RES-589-28-0001, RES-000-22-2761) and European Research Council (ERC-2009-StG-240910-ROMETA).

Victor Chernozhukov: Department of Economics, Massachusetts Institute of Technology, vchern@mit.edu.

Sokbae Lee: Department of Economics, Seoul National University and CeMMAP, sokbae@gmail.com.

Adam Rosen: Department of Economics, University College London and CeMMAP, adam.rosen@ucl.ac.uk.

1. INTRODUCTION

This paper develops a practical and novel method for estimation and inference on intersection bounds. Such bounds arise in settings where the parameter of interest, denoted θ^* , is known to lie within the bounds $[\theta^l(v), \theta^u(v)]$ for each v in some set $\mathcal{V} \subseteq \mathbb{R}^d$, which may be uncountably infinite. The identification region for θ^* is then

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

Intersection bounds stem naturally from exclusion restrictions (Manski (2003)) and appear in numerous applied and theoretical examples.¹ A leading case is that where the bounding functions are conditional expectations with continuous conditioning variables, yielding conditional moment inequalities. 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 covers both parametric and non-parametric estimators of bounding functions $\theta^l(\cdot)$ and $\theta^u(\cdot)$. We provide formal justification for parametric, series, and kernel-type estimators via asymptotic theory based on the strong approximation of a sequence of empirical processes by a sequence of Gaussian 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 addition, we generalize existing results on the strong approximation of kernel-type estimators to regression models with multivariate outcomes, and we provide a novel multiplier method to approximate the distribution of such estimators. For each of these estimation methods, the paper provides

- (i) confidence regions that achieve a desired asymptotic level,

¹Examples include average treatment effect bounds from instrumental variable restrictions (Manski (1990)), bounds on the distribution of treatment effects in a randomized experiment (Heckman, Smith, and Clements (1997)), treatment effect bounds from nonparametric selection equations with exclusion restrictions (Heckman and Vytlacil (1999)), 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)), changes in the distribution of wages (Blundell, Gosling, Ichimura, and Meghir (2007)), the study of disability and employment (Kreider and Pepper (2007)), unemployment compensation reform (Lee and Wilke (2009)), set identification with Tobin regressors (Chernozhukov, Rigobon, and Stoker (2010)), endogeneity with discrete outcomes (Chesher (2010)), estimation of income poverty measures (Nicoletti, Foliano, and Peracchi (2011)), bounds on average structural functions and treatment effects in triangular systems (Shaikh and Vytlacil (2011)), and set identification with imperfect instruments (Nevo and Rosen (2012)).

- (ii) novel adaptive inequality selection (AIS) needed to construct sharp critical values, which in some cases result in confidence regions with exact asymptotic size,²
- (iii) convergence rates for the boundary points of these regions,
- (iv) a characterization of local alternatives against which the associated tests have non-trivial power,
- (v) half-median-unbiased estimators of the intersection bounds.

Moreover, our paper also extends inferential theory based on empirical processes in Donsker settings to non-Donsker cases. The empirical processes arising in our problems do not converge weakly to a Gaussian process, but can be strongly approximated by a sequence of “penultimate” Gaussian processes, which we use directly for inference without resorting to further approximations, such as extreme value approximations as in Bickel and Rosenblatt (1973). These new methods may be of independent interest for a variety of other problems.

Our results also apply to settings where a parameter of interest, say μ , is characterized by intersection bounds of the form (1.1) on an auxiliary function $\theta(\mu)$. Then the bounding functions have the representation

$$\theta^l(v) := \theta^l(v; \mu) \text{ and } \theta^u(v) := \theta^u(v; \mu), \quad (1.2)$$

and thus inference statements for $\theta^* := \theta(\mu)$ bounded by $\theta^l(\cdot)$ and $\theta^u(\cdot)$ can be translated to inference statements for the parameter μ . This includes cases where the bounding functions are a collection of conditional moment functions indexed by μ . When the auxiliary function is additively separable in μ , the relation between the two is simply a location shift. When the auxiliary function is nonseparable in μ inference statements on θ^* still translate to inference statements on μ , though the functional relation between the two is more complex.

This paper overcomes significant complications for estimation of and inference on intersection bounds. First, because the bound estimates are suprema and infima of parametric or nonparametric estimators, closed-form characterization of their asymptotic distributions are typically unavailable or difficult to establish. As a consequence, researchers have often used the canonical bootstrap for inference, yet the recent literature indicates that the canonical bootstrap is not generally consistent in such settings, see e.g. Andrews and Han (2009), Bugni (2010), and Canay (2010).³ Second, since sample analogs of the bounds of Θ_I are the

²The previous literature, e.g. Chernozhukov, Hong, and Tamer (2007) and contemporaneous papers, such as Andrews and Shi (forthcoming), use “non-adaptive” cutoffs such as $C\sqrt{\log n}$. Ideally C should depend on the problem at hand and so careful calibration might be required in practice. Our new AIS procedure provides data-driven, adaptive cutoffs, which do not require calibration.

³The recent papers Andrews and Shi (forthcoming) and Kim (2009) provide justification for subsampling procedures for the statistics they employ for inference with conditional moment inequalities. We discuss these papers further in our literature review below.

suprema and infima of estimated bounding functions, they have substantial finite sample bias, and estimated bounds tend to be much tighter than the population bounds. This has been noted by Manski and Pepper (2000, 2009), and some heuristic bias adjustments have been proposed by Haile and Tamer (2003) and Kreider and Pepper (2007).

We solve the problem of estimation and inference for intersection bounds by proposing bias-corrected estimators of the upper and lower bounds, as well as confidence intervals. Specifically, our approach employs a precision-correction to the estimated bounding functions $v \mapsto \hat{\theta}^l(v)$ and $v \mapsto \hat{\theta}^u(v)$ before applying the supremum and infimum operators. We adjust the estimated bounding 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) confidence sets for either the identified set Θ_I or the true parameter value θ^* , or (ii) bias-corrected estimators for the lower and upper bounds. Our bias-corrected estimators are half-median-unbiased in the sense that the upper bound estimator $\hat{\theta}^u$ exceeds θ^u and the lower bound estimator $\hat{\theta}^l$ falls below θ^l each with probability at least one half asymptotically. Due to the presence of the inf and sup operators in the definitions of θ^u and θ^l , achieving unbiasedness is impossible in general, as shown by Hirano and Porter (2012), and which motivates the half-unbiasedness property. Bound estimators with this property are also proposed by Andrews and Shi (forthcoming), henceforth AS. An attractive feature of our approach is that the only difference in the construction of our estimators and confidence intervals is the choice of a critical value. Thus, practitioners need not implement two entirely different methods to construct estimators and confidence bands with desirable properties.

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 (forthcoming), Andrews and Guggenberger (2009), Andrews and Soares (2010), Beresteanu and Molinari (2008), Bugni (2010), Canay (2010), Chernozhukov, Hong, and Tamer (2007), Galichon and Henry (2009), Romano and Shaikh (2008), Romano and Shaikh (2010), and Rosen (2008), among others. We contribute to this literature by considering inference with a continuum of inequalities. Contemporaneous and independently written research on conditional moment inequalities includes AS, Kim (2009), and Menzel (2009). Our approach differs from all of these. Whereas we treat the problem with fundamentally nonparametric methods, AS provide inferential statistics that transform the

model's conditional restrictions to unconditional ones through the use of instrument functions.⁴ In this sense our approach is similar in spirit to that of Härdle and Mammen (1993) (although they use the L^2 norm and we use a sup test), while the approach of AS parallels that of Bierens (1982) for testing a parametric specification against a nonparametric alternative. As such, these approaches are complementary, each with their relative merits, as we describe in further detail below. Kim (2009) proposes an inferential method related to that of AS, but where data dependent indicator functions play the role of instrument functions. Menzel (2009) considers problems where the number of moment inequalities defining the identified set is large relative to the sample size. He provides results on the use of a subset of such restrictions in any finite sample, where the number of restrictions employed grows with the sample size, and examines the sensitivity of estimation and inference methods to the rate with which the number of moments used grows with the sample size.

The classes of models to which our approach and others in the recent literature apply have considerable overlap, most notably in models comprised of conditional moment inequalities, equivalently models whose bounding functions are conditional moment functions. Relative to other approaches, our approach is especially convenient for inference in parametric and nonparametric models with a continuum of inequalities that are separable in parameters, i.e. those admitting representations of the form

$$\sup_{v \in \mathcal{V}} \theta^l(v) \leq \theta^* \leq \inf_{v \in \mathcal{V}} \theta^u(v),$$

as in (1.1). Our explicit use of nonparametric estimation of bounding functions renders our method applicable in settings where the bounding functions depend on exogenous covariates in addition to the variable V , i.e. where the function $\theta(x)$ at a point x is the object of interest, with

$$\sup_{v \in \mathcal{V}} \theta^l(x, v) \leq \theta(x) \leq \inf_{v \in \mathcal{V}} \theta^u(x, v).$$

When the functions $\theta^l(x, v)$ and $\theta^u(x, v)$ are nonparametrically specified, these can be estimated by either the series or kernel-type estimators we study in Section 4. At present most other approaches do not appear to immediately apply when we are interested in $\theta(x)$

⁴Thus, the two approaches also require different assumptions. We rely on the strong approximation of a studentized version of parametric or nonparametric bounding function estimators (e.g. conditional moment functions in the context of conditional moment inequalities), while AS require that a functional central limit theorem hold for the *transformed* unconditional moment functions, which involve instrument functions not present in this paper.

at a point x , when covariates X are continuously distributed, with the exception of the recent work by Fan and Park (2011) in the context of IV and MIV bounds.⁵

To better understand the comparison between our point and interval estimators and those of AS when both are applicable, consider as a simple example the case where $\theta^* \leq E[Y|V]$ almost surely, with $E[Y|V = v]$ continuous in v . Then the upper bound on θ^* is $\theta_0 = \inf_{v \in \mathcal{V}} E[Y|V = v]$ over some region \mathcal{V} . θ_0 is a nonparametric functional and can in general only be estimated at a nonparametric rate. That is, one can not construct point or interval estimators that converge to θ_0 at superefficient rates, i.e. rates that exceed the optimal nonparametric rate for estimating $\theta(v) := \theta^u(v) = E[Y|V = v]$.⁶ Our procedure delivers point and interval estimators that can converge to θ_0 at this rate, up to an undersmoothing factor. However, there exist point and interval estimators that can achieve faster (super-efficient) convergence rates at *some* values of $\theta(\cdot)$. In particular, if the bounding function $\theta(\cdot)$ happens to be flat on the contact set $V_0 = \{v \in \mathcal{V} : \theta(v) = \theta_0\}$, meaning that V_0 is a set of positive Lebesgue measure, then the point and interval estimator of AS can achieve the convergence rate of $n^{-1/2}$. As a consequence, their procedure for testing $\theta_{na} \leq \theta_0$ against $\theta_{na} > \theta_0$, where $\theta_{na} = \theta_0 + C/\sqrt{n}$ for $C > 0$, has non-trivial asymptotic power, while our procedure does not. If, however, $\theta(\cdot)$ is not flat on V_0 , then the testing procedure of AS no longer has power against the aforementioned $n^{-1/2}$ alternatives, and results in point and interval estimators that converge to θ_0 at a sub-optimal rate.⁷ In contrast, our procedure delivers point and interval estimators that can converge at nearly the optimal rate, and hence can provide better power in these cases. In applications both flat and non-flat cases are important, and we therefore believe that both testing procedures are useful.⁸ For further comparisons, we refer the reader to our Monte-Carlo Section and to Supplemental Appendices K and L, which confirm these points both analytically and numerically.⁹

⁵The complication is that inclusion of additional covariates in a nonparametric framework requires a method for localization of the bounding function around the point x . With some non-trivial work and under appropriate conditions, the other approaches can likely be adapted to this context.

⁶Suppose for example that $V_0 = \arg \inf_{v \in \mathcal{V}} \theta(v)$ is singleton, with $\theta_0 = \theta(v)$ for some $v \in \mathcal{V}$. Then θ_0 is a nonparametric function evaluated at a single point, which cannot be estimated faster than the optimal nonparametric rate. Lower bounds on the rates of convergence in nonparametric models are characterized e.g. by Stone (1982) and Tsybakov (2009). Having a uniformly super-efficient procedure would contradict these lower bounds.

⁷With regard to confidence intervals/interval estimators, we mean here that the upper bound of the confidence interval does not converge at this rate.

⁸Note also that non-flat cases can be justified as generic if e.g. one takes $\theta(\cdot)$ as a random draw from the Sobolev ball equipped with the Gaussian (Wiener) measure.

⁹See Supplemental Appendix K for specific examples, and see Armstrong (2011b) for a comprehensive analysis of the power properties the procedure of Andrews and Shi (forthcoming). We also note that this qualitative comparison of local asymptotic power properties conforms with related results regarding tests of parametric models versus nonparametric (PvNP) alternatives, which involve moment *equalities*. Recall that our test relies on nonparametric estimation of bound-generating functions, which often take the form of

There have also been some more recent additions to the literature on conditional moment inequalities. Lee, Song, and Whang (forthcoming) develop a test for conditional moment inequalities using a one-sided version of L^p -type functionals of kernel estimators. Their approach is based on a least favorable configuration that permits valid but possibly conservative inference using standard normal critical values. Armstrong (2011b) and Chetverikov (2011) both propose interesting and important approaches to estimation and inference based on conditional moment inequalities, which can be seen as introducing full studentization in the procedure of AS, fundamentally changing its behavior. The resulting procedures use a collection of fully-studentized nonparametric estimators for inference, bringing them closer to the approach of this paper. Their implicit nonparametric estimators are locally constant with an adaptively chosen bandwidth. Our approach is specifically geared to smooth cases, where $\theta^u(\cdot)$ and $\theta^l(\cdot)$ are continuously differentiable of order $s \geq 2$ resulting in more precise estimates and hence higher power in these cases. On the other hand, in less smooth cases, the procedures of Armstrong (2011b) and Chetverikov (2011) automatically adapt to deliver optimal estimation and testing procedures, and so can perform somewhat better than our approach in these cases. Armstrong (2011a) derives the convergence rate and asymptotic distribution for a test statistic related to that in Armstrong (2011b) when evaluated at parameter values on the boundary of the identified set, drawing a connection to the literature on nonstandard M-estimation. Ponomareva (2010) studies bootstrap procedures for inference using kernel-based estimators, including one that can achieve asymptotically exact inference when the bounding function is uniquely maximized at a single point and locally quadratic. Our simulation-based approach does not rely on these conditions for its validity, but will automatically achieve asymptotic exactness with appropriately chosen smoothing parameters in a sufficiently regular subset of such cases.

Plan of the Paper. 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 under high level conditions. In Section 4 we provide conditions and theorems for validity for parametric and nonparametric series and kernel-type estimators. We provide several examples that demonstrate the use of primitive conditions to verify the conditions of Section 3. This includes sufficient conditions for the

conditional moment inequalities, and is similar in spirit to the approach of e.g. Härdle and Mammen (1993) in the PvNP testing literature. On the other hand, the statistics employed by AS rely on a transformation of conditional restrictions to unconditional ones in similar spirit to Bierens (1982). Tests of the latter type have been found to have power against some $n^{-1/2}$ alternatives, while the former do not. However, tests of the first type typically have non-trivial power against a larger class of alternatives, and so achieve higher power against some classes of alternatives. For further details see for example Horowitz and Spokoiny (2001) and the references therein.

application of these estimators to models comprised of conditional moment inequalities. In Section 5 we provide a theorem that establishes strong approximation for series estimators admitting an asymptotic linear representation, and which covers the examples of Section 4. Likewise, we provide a theorem that establishes strong approximation for kernel-type estimators in Section G.2 of an on-line supplement. In Section 6 we provide step-by-step implementation guides for parametric and non-parametric series and kernel-type estimators. In Section 7 we illustrate the performance of our method using both series and kernel-type estimators in Monte Carlo experiments, which we compare to that of AS in terms of coverage frequency and power. Our method performs well in these experiments, and we find that our approach and that of AS perform favorably in different models, depending on the shape of the bounding function. Section 8 concludes. In Appendices A - D we recall the definition of strong approximation and provide proofs, including the proof of the strong approximation result for series estimators. The on-line supplement contains further appendices. The first of these, Appendix E provides proofs omitted from the main text.¹⁰ Appendices F-H concern kernel-type estimators, providing primitive conditions for their application to conditional moment inequalities, strong approximation results and the multiplier method enabling inference via simulation, and proofs. Appendix I provides additional details on the use of primitive conditions to verify an asymptotic linear expansion needed for strong approximation of series estimators and Appendix J gives some detailed arguments for local polynomial estimation of conditional moment inequalities. Appendix K provides local asymptotic power analysis that supports the findings of our Monte Carlo experiments. Appendix L provides further Monte Carlo evidence.

Notation. For any two reals a and b , $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. $Q_p(X)$ denotes the p -th quantile of random variable X . We use $\text{wp} \rightarrow 1$ as shorthand for “with probability approaching one as $n \rightarrow \infty$.” We write $\mathcal{N}_k =_d N(0, I_k)$ to denote that the k -variate random vector \mathcal{N}_k is distributed multivariate normal with mean zero and variance the $k \times k$ identity matrix. To denote probability statements conditional on observed data, we write statements conditional on \mathcal{D}_n . \mathbb{E}_n and \mathbb{P}_n denote the sample mean and empirical measure, respectively. That is, given i.i.d. random vectors X_1, \dots, X_n , we have $\mathbb{E}_n f = \int f d\mathbb{P}_n = n^{-1} \sum_{i=1}^n f(X_i)$. In addition, let $\mathbb{G}_n f = \sqrt{n}(\mathbb{E}_n - E)f = n^{-1/2} \sum_{i=1}^n [f(X_i) - Ef(X)]$. The notation $a_n \lesssim b_n$ means that $a_n \leq Cb_n$ for all n ; $X_n \lesssim_{P_n} c_n$ abbreviates $X_n = O_{P_n}(c_n)$. $X_n \rightarrow_{P_n} \infty$ means that for any constant $C > 0$, $P_n(X_n < C) \rightarrow 0$. We use \mathbf{V} to denote a generic compact subset of \mathcal{V} , and we write $\text{diam}(\mathbf{V})$ to denote the diameter of \mathbf{V} in the Euclidian metric. $\|\cdot\|$ denotes the Euclidean norm, and for any two

¹⁰Specifically, Appendix E contains the proofs of Lemmas 2 and 4.

sets A, B in Euclidean space, $d_H(A, B)$ denotes the Hausdorff pseudo-distance between A and B with respect to the Euclidean norm. C stands for a generic positive constant, which may be different in different places, unless stated otherwise. For a set V and an element v in Euclidean space, let $d(v, V) := \inf_{v' \in V} \|v - v'\|$. For a function $p(v)$, let $\text{lip}(p)$ denote the Lipschitz coefficient, that is $\text{lip}(p) := L$ such that $\|p(v_1) - p(v_2)\| \leq L\|v_1 - v_2\|$ for all v_1 and v_2 in the domain of $p(v)$.

2. MOTIVATING EXAMPLES AND INFORMAL OVERVIEW OF RESULTS

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

Example A: 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 different treatment is not known. Although 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 at other prices.

Suppose only that the support of the outcome space is known, $Y \in [0, 1]$, but no other assumptions are made regarding the distribution of counterfactual outcomes. Manski (1989, 1990) provides worst-case bounds on mean treatment outcomes for any treatment t conditional on observables $(X, V) = (x, v)$,

$$\theta^l(x, v) \leq E[Y(t) | X = x, V = v] \leq \theta^u(x, v),$$

where the bounds are

$$\begin{aligned} \theta^l(x, v) &:= E[Y \cdot 1\{Z = t\} | X = x, V = v], \\ \theta^u(x, v) &:= E[Y \cdot 1\{Z = t\} + 1\{Z \neq t\} | X = x, V = v], \end{aligned}$$

where Z is the observed treatment. If V is an instrument satisfying $E[Y(t) | X, V] = E[Y(t) | X]$, then for any fixed x , sharp bounds on $\theta^* := \theta^*(x) := E[Y(t) | X = x]$ are given by

$$\sup_{v \in \mathcal{V}} \theta^l(x, v) \leq \theta^*(x) \leq \inf_{v \in \mathcal{V}} \theta^u(x, v),$$

for any $\mathcal{V} \subseteq \text{support}(V|X = x)$, where the subset \mathcal{V} will be taken as known for estimation purposes. 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.

Example B: Bounding Distributions to Account for Selection. Similar analysis applies to inference on distributions whose observations are censored due to selection. This approach is used by Blundell, Gosling, Ichimura, and Meghir (2007) to study changes in male and female wages. The starting point of their analysis is that the cumulative distribution $F(w|x, v)$ of wages W at any point w , conditional on observables $(X, V) = (x, v)$ must satisfy the worst case bounds

$$\theta^l(x, v) \leq F(w|x, v) \leq \theta^u(x, v), \quad (2.1)$$

where D is an indicator of employment, and hence observability of W , so that

$$\begin{aligned} \theta^l(x, v) &:= E[D \cdot 1\{W \leq w\} | X = x, V = v], \\ \theta^u(x, v) &:= E[D \cdot 1\{W \leq w\} + (1 - D) | X = x, V = v]. \end{aligned}$$

This relation is used to bound quantiles of conditional wage distributions. Additional restrictions motivated by economic theory are then used to tighten the bounds.

One such restriction is an exclusion restriction of the continuous variable out-of-work income, V . They consider the use of V as either an excluded or monotone instrument. The former restriction implies bounds on the parameter $\theta^* := F(w|x)$,

$$\sup_{v \in \mathcal{V}} \theta^l(x, v) \leq F(w|x) \leq \inf_{v \in \mathcal{V}} \theta^u(x, v), \quad (2.2)$$

for any $\mathcal{V} \subseteq \text{support}(V|X = x)$, while the weaker monotonicity restriction implies the following bounds on $\theta^* := F(w|x, v_0)$ for any v_0 in $\text{support}(V|X = x)$,

$$\sup_{v \in \mathcal{V}_l} \theta^l(x, v) \leq F(w|x, v_0) \leq \inf_{v \in \mathcal{V}_u} \theta^u(x, v), \quad (2.3)$$

where $\mathcal{V}_l = \{v \in \mathcal{V} : v \leq v_0\}$ and $\mathcal{V}_u = \{v \in \mathcal{V} : v \geq v_0\}$.

Example C: (Conditional) Conditional Moment Inequalities. Our inferential method can also be used for pointwise inference on parameters restricted by (possibly conditional) conditional moment inequalities. Such restrictions arise naturally in empirical work in industrial organization, see for example Pakes, Porter, Ho, and Ishii (2005) and Berry and Tamer (2007).

To illustrate, consider the restriction

$$E[m_j(X, \mu_0) | Z = z] \geq 0 \text{ for all } j = 1, \dots, J \text{ and } z \in \mathcal{Z}_j. \quad (2.4)$$

where each $m_j(\cdot, \cdot)$, $j = 1, \dots, J$, is a real-valued function, (X, Z) are observables, and μ_0 is the parameter of interest. Note that this parameter can depend on a particular covariate value. Suppose for instance that $Z = (Z_1, Z_2)$ and interest lies in the subgroup of the population with $Z_1 = z_1$, so that the researcher wishes to condition on $Z_1 = z_1$. In this case $\mu_0 = \mu_0(z_1)$ depends on z_1 . Conditioning on this value, we have from (2.4) that

$$E[m_j(X, \mu_0) | Z_1 = z_1, Z_2 = z_2] \geq 0 \text{ for all } j = 1, \dots, J \text{ and } z_2 \in \text{Supp}(Z_2 | Z_1 = z_1),$$

which is equivalent to (2.4) with $\mathcal{Z}_j = \text{Supp}(Z | Z_1 = z_1)$. Note also that regions \mathcal{Z}_j can depend on the inequality j as in (2.3) of the previous example, and that the previous two examples can in fact be cast as special cases of this one.

Suppose that we would like to test (2.4) at level α for the conjectured parameter value $\mu_0 = \mu$ against an unrestricted alternative. To see how this can be done, define

$$v = (z, j), \quad \mathcal{V} := \{(z, j) : z \in \mathcal{Z}_j, j \in \{1, \dots, J\}\} \text{ and } \theta(\mu, v) := E[m_j(X, \mu) | Z = z]$$

and $\hat{\theta}(\mu, v)$ a consistent estimator. Under some continuity conditions this is equivalent to a test of $\theta_0(\mu) := \inf_{v \in \mathcal{V}} \theta(\mu, v) \geq 0$ against $\inf_{v \in \mathcal{V}} \theta(\mu, v) < 0$. Our method for inference delivers a statistic

$$\hat{\theta}_\alpha(\mu) = \inf_{v \in \mathcal{V}} [\hat{\theta}(\mu, v) + \hat{k} \cdot s(\mu, v)]$$

such that $\lim_{n \rightarrow \infty} P(\theta_0(\mu) \geq \hat{\theta}_\alpha(\mu)) \leq \alpha$. Here, $s(\mu, v)$ is the standard error of $\hat{\theta}(\mu, v)$ and \hat{k} is an estimated critical value, as we describe below. If $\hat{\theta}_\alpha(\mu) < 0$, we reject the null hypothesis, while if $\hat{\theta}_\alpha(\mu) \geq 0$, we do not.

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

$$\theta^* \leq \theta_0 := \inf_{v \in \mathcal{V}} \theta(v), \quad (2.5)$$

where $v \mapsto \theta(v)$ is a bounding function, and \mathcal{V} is the set over which the infimum is taken. We focus on describing our method for the upper bound (2.5), as the lower bound is entirely symmetric. In fact, any combination of upper and lower bounds can be combined into upper bounds on an auxiliary function of θ^* of the form (2.5), and this can be used for inference on θ^* , as we describe in Section 6.¹¹

¹¹Alternatively, one can combine one-sided intervals for lower and upper bounds for inference on the identified set Θ_I using Bonferroni's inequality, or for inference on θ^* using the method described in Chernozhukov, Lee,

What are good *estimators* and *confidence regions* for the bound θ_0 ? A natural idea is to base estimation and inference on the sample analog: $\inf_{v \in \mathcal{V}} \hat{\theta}(v)$. However, this estimator does not perform well in practice. First, the analog estimator tends to be downward biased in finite samples. As discussed in the introduction, this will typically result in bound estimates that are much narrower than those in the population, see e.g. Manski and Pepper (2000) and Manski and Pepper (2009) for more on this point. Second, inference must appropriately take account of sampling error of the estimator $\hat{\theta}(v)$ across all values of v . Indeed, different levels of precision of $\hat{\theta}(v)$ at different points can severely distort the perception of the minimum of the bounding function $\theta(v)$. Figure 1 illustrates these problems geometrically. The solid curve is the true bounding function $v \mapsto \theta(v)$, and the dash-dotted thick curve is its estimate $v \mapsto \hat{\theta}(v)$. The remaining dashed curves represent eight additional potential realizations of the estimator, illustrating its precision. 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 :

$$\hat{\theta}(p) := \inf_{v \in \mathcal{V}} [\hat{\theta}(v) + k(p) \cdot s(v)], \quad (2.6)$$

where $s(v)$ is the standard error of $\hat{\theta}(v)$, and $k(p)$ is a critical value, the selection of which is described below. That is, our estimator $\hat{\theta}(p)$ minimizes the *precision-corrected curve* given by $\hat{\theta}(v)$ plus critical value $k(p)$ 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 $\inf_{v \in \mathcal{V}} \hat{\theta}(v)$.

These issues are important both in theory and in practice, as can be seen in the application of the working paper Chernozhukov, Lee, and Rosen (2009). There we used the data from the National Longitudinal Survey of Youth of 1979 (NLSY79), as in Carneiro and Lee (2009), to estimate bounds on expected log wages Y_i as a function of years of schooling t . We used Armed Forces Qualifying Test score (AFQT) normalized to have mean zero as a monotone instrumental variable, and estimated the MIV-MTR (monotone instrument variable - monotone treatment response) bounds of Manski and Pepper (2000). Figures 3

and Rosen (2009) Section 3.7, which is a slight generalization of methods previously developed by Imbens and Manski (2004) and Stoye (2009).

and 4 highlight the same issues as the schematic figures 1 and 2 with the NLSY data and the MIV-MTR upper bound for the parameter $\theta^* = P[Y_i(t) > y|V_i = v]$, at $y = \log(24)$ ($\sim 90^{th}$ percentile of hourly wages) and $v = 0$ for college graduates ($t = 16$).¹²

Figure 3 shows the nonparametric series estimate of the bounding function using B-splines as described in Section 7.2 (solid curve) and 20 bootstrap estimates (dashed curves). The precision of the estimate is worst when the AFQT is near 2, as demonstrated by the bootstrap estimates. At the same time, the bounding function has a decreasing shape with the minimum at AFQT = 2. Figure 4 shows a precision-corrected curve (solid curve) that adjusts the bound estimate $\hat{\theta}(v)$ (dashed curve) by an amount proportional to its point-wise standard error, and the horizontal dashed line shows the end point of a 95% one-sided confidence interval. As in Figure 2, the minimizer of the precision-corrected curve is quite far from that of the uncorrected estimate of the bounding function.

The degree of precision correction, both in these figures and in general, is driven by the critical value $k(p)$. The main input in the selection of $k(p)$ for the estimator $\hat{\theta}(p)$ in (2.6) is the standardized process

$$Z_n(v) = \frac{\theta(v) - \hat{\theta}(v)}{\sigma(v)},$$

where $\sigma(v)/s(v) \rightarrow 1$ uniformly in v . Generally, the finite-sample distribution of the process Z_n is unknown, but we can approximate it uniformly by a sequence of Gaussian processes Z_n^* such that for an appropriate sequence of constants \bar{a}_n

$$\bar{a}_n \sup_{v \in \mathcal{V}} |Z_n(v) - Z_n^*(v)| = o_p(1). \quad (2.7)$$

For any compact set \mathbf{V} , used throughout to denote a generic compact subset of \mathcal{V} , we then approximate the quantiles of $\sup_{v \in \mathbf{V}} Z_n^*(v)$ either by analytical methods based on asymptotic approximations, or by simulation. We then use the p -quantile of this statistic, $k_{n,\mathbf{V}}(p)$, in place of $k(p)$ in (2.6). We show that in general simulated critical values provide sharper inference, and therefore advocate their use.

For the estimator in (2.6) to exceed θ_0 with probability no less than p asymptotically, we require that $\text{wp} \rightarrow 1$ the set \mathbf{V} contains the argmin set

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

A simple way to achieve this is to use $\mathbf{V} = \mathcal{V}$, which leads to asymptotically valid but conservative inference. For construction of the critical value $k_{n,\mathbf{V}}(p)$ above we thus propose

¹²The parameter θ^* used for this illustration differs from the conditional expectations bounded in Chernozhukov, Lee, and Rosen (2009). For further details regarding the application and the data we refer to that version, available at <http://cemmap.ifs.org.uk/wps/cwp1909.pdf>.

the use of a preliminary set estimator \widehat{V}_n for V_0 using a novel adaptive inequality selection procedure. Because the critical value $k_{n,V}(p)$ is non-decreasing in V for n large enough, this yields an asymptotic critical value no larger than those based on $V = \mathcal{V}$. The set estimator \widehat{V}_n is shown to be sandwiched between two non-stochastic sequences of sets, a lower envelope V_n and an upper envelope \overline{V}_n with probability going to one. We show in Lemma 1 that our inferential procedure using \widehat{V}_n concentrates on the lower envelope V_n , which is a neighborhood of the contact set V_0 . This validates our use of the set estimator \widehat{V}_n . The upper envelope \overline{V}_n , a larger - but nonetheless shrinking - neighborhood of the contact set V_0 , plays an important role in the derivation of estimation rates and local power properties of our procedure. Specifically, because this set contains V_0 wp $\rightarrow 1$, the tail behavior of $\sup_{v \in \overline{V}_n} Z_n^*(v)$ can be used to bound the estimation error of $\widehat{\theta}_{n0}(p)$ relative to θ_{n0} .

Moreover, we show that in some cases inference based on simulated critical values using \widehat{V}_n in fact concentrates on V_0 rather than just V_n . This requires that the scaled penultimate process $\bar{a}_n Z_n^*$ behaves sufficiently well (i.e. is equicontinuous) within r_n neighborhoods of V_0 , where r_n denotes the rate of convergence of the set estimator \widehat{V}_n to the argmin set V_0 . When this holds the tail behavior of $\sup_{v \in V_0} Z_n^*(v)$ rather than $\sup_{v \in \overline{V}_n} Z_n^*(v)$ bounds the estimation error of our estimator. This typically leads to small improvements in the convergence rate of our estimator and the local power properties of our approach. The conditions for this to occur include the important special case where V_0 is singleton, and where the bounding function is locally quadratic, though can hold more generally. The formal conditions are given in Section 3.5, where we provide conditions for consistency and rates of convergence of \widehat{V}_n for V_0 , and in Section 3.6 where we provide the aforementioned equicontinuity condition and a formal statement of the result regarding when inference concentrates on V_0 .

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. Our theory for nonparametric estimation relies on undersmoothing, although for locally constant or sign-preserving estimation of bounding functions, this does not appear essential since the approximation bias is conservatively signed. Our inference algorithm still applies to nonparametric estimates of bounding functions without undersmoothing, although our theory would require some minor modifications to handle this case. We do not formally pursue this here, but we provide some simulation results for kernel estimation without undersmoothing as part of the additional Monte Carlo experiments reported in supplementary appendix L.

For all estimators, parametric and nonparametric, we employ strong approximation analysis to approximate the quantiles of $\sup_{v \in \mathcal{V}} Z_n(v)$, and we verify our conditions separately for each case. The formal definition of strong approximation is provided in Appendix A.

3. ESTIMATION AND INFERENCE THEORY UNDER GENERAL CONDITIONS

3.1. Basic Framework. In this and subsequent sections we allow the model and the probability measure to depend on n . Formally, we work with a probability space (A, \mathcal{A}, P_n) throughout. This approach is conventionally used in asymptotic statistics to ensure robustness of statistical conclusions with respect to perturbations in P_n . It guarantees the validity of our inference procedure under any sequence of probability laws P_n that obey our conditions, including the case with fixed P . We thus generalize our notation in this section to allow model parameters to depend on n .

The basic setting is as follows:

Condition C. 1 (Setting). *There is a non-empty compact set $\mathcal{V} \subset \mathcal{K} \subset \mathbb{R}^d$, where \mathcal{V} can depend on n , and \mathcal{K} is a bounded fixed set, independent of n . There is a continuous real valued function $v \mapsto \theta_n(v)$. There is an estimator $v \mapsto \hat{\theta}_n(v)$ of this function, which is an a.s. continuous stochastic process. There is a continuous function $v \mapsto \sigma_n(v)$ representing non-stochastic normalizing factors bounded by $\bar{\sigma}_n := \sup_{v \in \mathcal{V}} \sigma_n(v)$, and there is an estimator $v \mapsto s_n(v)$ of these factors, which is an a.s. continuous stochastic process, bounded above by $\bar{s}_n := \sup_{v \in \mathcal{V}} s_n(v)$.*

We are interested in constructing point estimators and one-sided interval estimators for

$$\theta_{n0} = \inf_{v \in \mathcal{V}} \theta_n(v).$$

The main input in this construction is the standardized process

$$Z_n(v) = \frac{\theta_n(v) - \hat{\theta}_n(v)}{\sigma_n(v)}.$$

In the following we require that this process can be approximated by a standardized Gaussian process in the metric space $\ell^\infty(\mathcal{V})$ of bounded functions mapping \mathcal{V} to \mathbb{R} , which can be simulated for inference.

Condition C. 2 (Strong Approximation). *(a) Z_n is strongly approximated by a sequence of penultimate Gaussian processes Z_n^* having zero mean and a.s. continuous sample paths:*

$$\sup_{v \in \mathcal{V}} |Z_n(v) - Z_n^*(v)| = o_{P_n}(\delta_n),$$

where $E_{P_n}[(Z_n^*(v))^2] = 1$ for each $v \in \mathcal{V}$, and $\delta_n = o(\bar{a}_n^{-1})$ for the sequence of constants \bar{a}_n defined in Condition C.3 below. (b) Moreover, for simulation purposes, there is a process Z_n^* , whose distribution is zero-mean Gaussian conditional on the data \mathcal{D}_n and such that $E_{P_n}[Z_n^*(v) \mid \mathcal{D}_n]^2 = 1$ for each $v \in \mathcal{V}$, that can approximate an identical copy \bar{Z}_n^* of Z_n^* , where \bar{Z}_n^* is independent of \mathcal{D}_n , namely there is $o(\delta_n)$ term such that

$$P_n \left[\sup_{v \in \mathcal{V}} |\bar{Z}_n^*(v) - Z_n^*(v)| > o(\delta_n) \mid \mathcal{D}_n \right] = o_{P_n}(1/\ell_n)$$

for some $\ell_n \rightarrow \infty$ chosen below.

For convenience we refer to Appendix A, where the definition of strong approximation is recalled. The penultimate process Z_n^* is often called a coupling, and we construct such couplings for parametric and nonparametric estimators under both high-level and primitive conditions. It is convenient to work with Z_n^* , since we can rely on the fine properties of Gaussian processes. Note that Z_n^* depends on n and generally does not converge weakly to a fixed Gaussian process, and therefore is not asymptotically Donsker. Nonetheless we can perform either analytical or simulation-based inference based on these processes.

Our next condition captures the so-called *concentration* properties of Gaussian processes:

Condition C. 3 (Concentration). *For all n sufficiently large and for any compact, non-empty $V \subseteq \mathcal{V}$, there is a normalizing factor $a_n(V)$ satisfying*

$$1 \leq a_n(V) \leq a_n(\mathcal{V}) =: \bar{a}_n, \quad a_n(V) \text{ is increasing in } V,$$

such that

$$\mathcal{E}_n(V) := a_n(V) \left(\sup_{v \in V} Z_n^*(v) - a_n(V) \right)$$

obeys

$$P_n[\mathcal{E}_n(V) \geq x] \leq P[\mathcal{E} \geq x], \tag{3.1}$$

where \mathcal{E} is a random variable with continuous distribution function such that for some $\eta > 0$,

$$P(\mathcal{E} > x) \leq \exp(-x/\eta).$$

The *concentration* condition will be verified in our applications by appealing to the Talagrand-Samorodnitsky inequality for the concentration of the suprema of Gaussian processes, which is sharper than the classical concentration inequalities.¹³ These concentration properties play a key role in our analysis, as they determine the uniform speed of convergence $\bar{a}_n \bar{\sigma}_n$ of the estimator $\hat{\theta}_n(p)$ to θ_{n0} . In particular this property implies that for

¹³For details see Lemma 12 in Appendix C.1.

any compact $V_n \subseteq \mathcal{V}$, $E_{P_n}[\sup_{v \in V_n} Z_n^*(v)] \lesssim \bar{a}_n$. As there is concentration, there is an opposite force, called *anti-concentration*, which implies that under C.2(a) and C.3 for any $\delta_n = o(1/\bar{a}_n)$ we have

$$\sup_{x \in \mathbb{R}} P_n \left(\left| \sup_{v \in V_n} Z_n^*(v) - x \right| \leq \delta_n \right) \rightarrow 0. \quad (3.2)$$

This follows from a generic anti-concentration inequality derived in Chernozhukov and Kato (2011), quoted in Appendix B for convenience. Anti-concentration simplifies the construction of our confidence intervals. Finally, the exponential tail property of \mathcal{E} plays an important role in the construction of our adaptive inequality selector, introduced below, since it allows us to bound moderate deviations of the one-sided estimation noise of $\sup_{v \in V} Z_n^*(v)$.

Our next assumption requires uniform consistency as well as suitable estimates of σ_n :

Condition C.4 (Uniform Consistency). *We have that*

$$(a) \bar{a}_n \bar{\sigma}_n = o(1) \quad \text{and} \quad (b) \sup_{v \in \mathcal{V}} \left| \frac{s_n(v)}{\sigma_n(v)} - 1 \right| = o_{P_n} \left(\frac{\delta_n}{\bar{a}_n + \ell \ell_n} \right),$$

where $\ell \ell_n \nearrow \infty$ is a sequence of constants defined below.

In what follows we let

$$\ell_n := \log n, \text{ and } \ell \ell_n := \log \ell_n,$$

but it should be noted that ℓ_n can be replaced by other slowly increasing sequences.

3.2. The Inference and Estimation Strategy. For any compact subset $V \subseteq \mathcal{V}$ and $\gamma \in (0, 1)$, define:

$$\kappa_{n,V}(\gamma) := Q_\gamma \left(\sup_{v \in V} Z_n^*(v) \right).$$

The following result is useful for establishing validity of our inference procedure.

Lemma 1 (Inference Concentrates on a Neighborhood V_n of V_0). *Under C.1-C.4*

$$P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \leq x \right) \geq P_n \left(\sup_{v \in V_n} Z_n^*(v) \leq x \right) - o(1),$$

uniformly in $x \in [0, \infty)$, where

$$V_n := \{v \in \mathcal{V} : \theta_n(v) \leq \theta_{n0} + \kappa_n \sigma_n(v)\}, \text{ for } \kappa_n := \kappa_{n,\mathcal{V}}(\gamma'_n), \quad (3.3)$$

where γ'_n is any sequence such that $\gamma'_n \nearrow 1$ with $\kappa_n/(\bar{a}_n + \ell \ell_n) \lesssim 1$.

Thus, with probability converging to one, the inferential process concentrates on a neighborhood of V_0 given by V_n . The “size” of the neighborhood is determined by κ_n , a high

quantile of $\sup_{v \in \mathcal{V}} Z_n^*(v)$, which summarizes the maximal one-sided estimation error over \mathcal{V} . We use this to construct half-median-unbiased estimators for θ_{n0} as well as one-sided interval estimators for θ_{n0} with correct asymptotic level, based on analytical and simulation methods for obtaining critical values proposed below.

Remark 1 (Sharp Concentration of Inference). In general, it is not possible for the inferential processes to concentrate on smaller subsets than V_n . However, as shown, in Section 3.6, in some special cases, e.g. when V_0 is a well-identified singleton, the inference process will in fact concentrate on V_0 . In this case our simulation-based construction will automatically adapt to deliver median-unbiased estimators for θ_{n0} as well as one-sided interval estimators for θ_{n0} with correct asymptotic size. Indeed, in the special but extremely important case of V_0 being singleton we can achieve

$$\mathbb{P}_n \left(\sup_{v \in V_n} Z_n^*(v) > x \right) = \Pr \left(N(0, 1) > x \right) - o(1),$$

under some regularity conditions. In this case, our simulation-based procedure will automatically produce a critical value that approaches the p -th quantile of the standard normal, delivering asymptotically exact inference. \blacksquare

Our construction relies first on an *auxiliary critical value* $k_{n,\mathcal{V}}(\gamma_n)$, chosen so that $\text{wp} \rightarrow 1$,

$$k_{n,\mathcal{V}}(\gamma_n) \geq \kappa_{n,\mathcal{V}}(\gamma'_n), \quad (3.4)$$

where we set $\gamma_n := 1 - .1/\ell_n \nearrow 1$ and $\gamma'_n = \gamma_n - o(1) \nearrow 1$. This critical value is used to obtain a preliminary set estimator

$$\widehat{V}_n = \left\{ v \in \mathcal{V} : \widehat{\theta}_n(v) \leq \min_{\tilde{v} \in \mathcal{V}} \left(\widehat{\theta}_n(\tilde{v}) + k_{n,\mathcal{V}}(\gamma_n) s_n(\tilde{v}) \right) + 2k_{n,\mathcal{V}}(\gamma_n) s_n(v) \right\}, \quad (3.5)$$

The set estimator \widehat{V}_n is then used in the construction of the *principal critical value* $k_{n,\widehat{V}_n}(p)$, $p \geq 1/2$, where we require that $\text{wp} \rightarrow 1$,

$$k_{n,\widehat{V}_n}(p) \geq \kappa_{n,\widehat{V}_n}(p - o(1)). \quad (3.6)$$

The principal critical value is fundamental to our construction of confidence regions and estimators, which we now define.

Definition 1 (Generic Interval and Point Estimators). *Let $p \geq 1/2$, then our interval estimator takes the form:*

$$\widehat{\theta}_{n0}(p) = \inf_{v \in \mathcal{V}} \left[\widehat{\theta}_n(v) + k_{n,\widehat{V}_n}(p) s_n(v) \right], \quad (3.7)$$

where the half-median unbiased estimator corresponds to $p = 1/2$.

The principal and auxiliary critical values are constructed below so as to satisfy (3.4) and (3.6) using either analytical or simulation methods. As a consequence, we show in Theorems 1 and 2 that

$$\mathbb{P}_n \left\{ \theta_{n0} \leq \widehat{\theta}_{n0}(p) \right\} \geq p - o(1), \quad (3.8)$$

for any fixed $1/2 \leq p < 1$. The construction relies on the new set estimator \widehat{V}_n , which we call an *adaptive inequality selector* (AIS), since it uses the problem-dependent cutoff $k_{n,\mathcal{V}}(\gamma_n)$, which is a bound on a high quantile of $\sup_{v \in \mathcal{V}} Z_n^*(v)$. The analysis therefore must take into account the moderate deviations (tail behavior) of the latter.

Before proceeding to the details of its construction, we note that the argument for establishing the coverage results and analyzing power properties of the procedure depends crucially on the following result proven in Lemma 2 below:

$$\mathbb{P}_n \left\{ V_n \subseteq \widehat{V}_n \subseteq \overline{V}_n \right\} \rightarrow 1,$$

where

$$\overline{V}_n := \{v \in \mathcal{V} : \theta_n(v) \leq \theta_{n0} + \bar{\kappa}_n \sigma_n(v)\}, \text{ for } \bar{\kappa}_n := 4(\bar{a}_n + \eta \ell \ell_n / \bar{a}_n). \quad (3.9)$$

Thus, the preliminary set estimator \widehat{V}_n is sandwiched between two deterministic sequences of sets, facilitating the analysis of its impact on the convergence of $\widehat{\theta}_{n0}(p)$ to θ_{n0} .

3.3. Analytical Method and Its Theory. Our first construction is quite simple and demonstrates the main – though not the finest – points. This construction uses the majorizing variable \mathcal{E} appearing in C.3.

Definition 2 (Analytical Method for Critical Values). *For any compact set V and any $p \in (0, 1)$, we set*

$$k_{n,V}(p) = a_n(V) + c(p)/a_n(V), \quad (3.10)$$

where $c(p) = Q_p(\mathcal{E})$ is the p -th quantile of the majorizing variable \mathcal{E} defined in C.3. Note that under the requirement of C.3 that $a_n(V)$ is increasing in V we have that $V \mapsto k_{n,V}(p)$ is weakly monotone increasing in V for sufficiently large n .

The first main result is as follows.

Theorem 1 (Analytical Inference, Estimation, Power under C.1-C.4). *Suppose C.1-C.4 hold. Consider the interval estimator given in Definition 1 with critical value function given in Definition 2. Then, for a given $p \in [1/2, 1)$,*

1. The interval estimator has asymptotic level p :

$$\mathbb{P}_n \left\{ \theta_{n0} \leq \hat{\theta}_{n0}(p) \right\} \geq p - o(1).$$

2. The estimation risk is bounded by, $wp \rightarrow 1$ under \mathbb{P}_n ,

$$\left| \hat{\theta}_{n0}(p) - \theta_{n0} \right| \leq 4\bar{\sigma}_n \left(a_n(\bar{V}_n) + \frac{O_{\mathbb{P}_n}(1)}{a_n(\bar{V}_n)} \right) \lesssim_{\mathbb{P}_n} \bar{\sigma}_n \bar{a}_n.$$

3. Hence, any, possibly data-dependent, alternative $\theta_{na} > \theta_{n0}$ such that

$$\theta_{na} \geq \theta_{n0} + 4\bar{\sigma}_n \left(a_n(\bar{V}_n) + \frac{\mu_n}{a_n(\bar{V}_n)} \right), \quad \mu_n \rightarrow_{\mathbb{P}_n} \infty,$$

is rejected with probability converging to 1 under \mathbb{P}_n .

Thus, $(-\infty, \hat{\theta}_{n0}(p)]$ is a valid one-sided interval estimator for θ_{n0} . Moreover, $\hat{\theta}_{n0}(1/2)$ is a half-median-unbiased estimator for θ_{n0} in the sense that

$$\lim_{n \rightarrow \infty} \mathbb{P}_n \left[\theta_{n0} \leq \hat{\theta}_{n0}(1/2) \right] \geq 1/2.$$

The rate of convergence of $\hat{\theta}_{n0}(p)$ to θ_{n0} is bounded above by the uniform rate $\bar{\sigma}_n \bar{a}_n$ for estimation of the bounding function $v \mapsto \theta_n(v)$. This implies that the test of $H_0 : \theta_{n0} = \theta_{na}$ that rejects if $\theta_{na} > \hat{\theta}_{n0}(p)$ asymptotically rejects all local alternatives that are more distant than $\bar{\sigma}_n \bar{a}_n$, including fixed alternatives as a special case. In Section 4 below we show that in parametric cases this results in power against $n^{-1/2}$ local alternatives. For series estimators $\bar{a}_n \bar{\sigma}_n$ is proportional to $(\log n)^c \sqrt{K/n}$ where c is some positive constant, and $K \rightarrow \infty$ is the number of series terms. For kernel-type estimators of bounding functions the rate $\bar{a}_n \bar{\sigma}_n$ is proportional to $(\log n)^c / \sqrt{nh^d}$ where c is some positive constant and h is the bandwidth, assuming some undersmoothing is done. For example, if the bounding function is s -times differentiable, $\bar{\sigma}_n$ can be made close to $(\log n/n)^{s/(2s+d)}$ apart from some undersmoothing factor by considering a local polynomial estimator, see Stone (1982). For both series and kernel-type estimators we show below that \bar{a}_n can be bounded by $\sqrt{\log n}$.

3.4. Simulation-Based Construction and Its Theory. Our main and preferred approach is based on the simple idea of simulating quantiles of relevant statistics.

Definition 3 (Simulation Method for Critical Values). *For any compact set $V \subseteq \mathcal{V}$, we set*

$$k_{n,V}(p) = Q_p \left(\sup_{v \in V} Z_n^*(v) \mid \mathcal{D}_n \right). \quad (3.11)$$

We have the following result for simulation inference, analogous to that obtained for analytical inference.

Theorem 2 (Simulation Inference, Estimation, Power under C.1-C.4). *Suppose C.1-C.4 hold. Consider the interval estimator given in Definition 1 with the critical value function specified in Definition 3. Then, for a given $p \in [1/2, 1)$,*

1. *The interval estimator has asymptotic level p :*

$$\mathbb{P}_n \left\{ \theta_{n0} \leq \hat{\theta}_{n0}(p) \right\} \geq p - o(1).$$

2. *The estimation risk is bounded by, $wp \rightarrow 1$ under \mathbb{P}_n ,*

$$\left| \hat{\theta}_{n0}(p) - \theta_{n0} \right| \leq 4\bar{\sigma}_n \left(a_n(\bar{V}_n) + \frac{O_{\mathbb{P}_n}(1)}{a_n(\bar{V}_n)} \right) \lesssim_{\mathbb{P}_n} \bar{\sigma}_n \bar{a}_n.$$

3. *Any, possibly data-dependent, alternative $\theta_{na} > \theta_{n0}$ such that*

$$\theta_{na} \geq \theta_{n0} + 4\bar{\sigma}_n \left(a_n(\bar{V}_n) + \frac{\mu_n}{a_n(\bar{V}_n)} \right), \quad \mu_n \rightarrow_{\mathbb{P}_n} \infty,$$

is rejected with probability converging to 1 under \mathbb{P}_n .

3.5. Properties of the Set Estimator \hat{V}_n . In this section we establish some containment properties for the estimator \hat{V}_n . Moreover, these containment properties imply a useful rate result under the following condition:

Condition V (Degree of Identifiability for V_0). *There exist constants $\rho_n > 0$ and $c_n > 0$, possibly dependent on n , and a positive constant δ , independent of n , such that*

$$\theta_n(v) - \theta_{n0} \geq (c_n d(v, V_0))^{\rho_n} \wedge \delta, \quad \forall v \in \mathcal{V}. \quad (3.12)$$

We say $(c_n, 1/\rho_n)$ characterize the degree of identifiability of V_0 , as these parameters determine the rate at which V_0 can be consistently estimated. Note that if $V_0 = \mathcal{V}$, then this condition holds with $c_n = \infty$ and $\rho_n = 1$, where we adopt the convention that $0 \cdot \infty = 0$.

We have the following result, whose first part we use in the proof of Theorems 1 and 2 above, and whose second part we use below in the proof of Theorem 3.

Lemma 2 (Estimation of V_n and V_0). *Suppose C.1-C.4 hold.*

1. *(Containment). Then $wp \rightarrow 1$, for either analytical or simulation methods,*

$$V_n \subseteq \hat{V}_n \subseteq \bar{V}_n,$$

for V_n defined in (3.3) with $\gamma'_n = \gamma_n - o(1)$, and \bar{V}_n defined in (3.9).

2. *(Rate) If also Condition V holds and $\bar{\kappa}_n \bar{\sigma}_n \rightarrow 0$, then $wp \rightarrow 1$*

$$\begin{aligned} d_H(\hat{V}_n, V_0) &\leq d_H(\hat{V}_n, V_n) + d_H(V_n, V_0) \\ &\leq d_H(\bar{V}_n, V_n) + d_H(V_n, V_0) \leq r_n := 2(\bar{\kappa}_n \bar{\sigma}_n)^{1/\rho_n} / c_n. \end{aligned}$$

3.6. Automatic Sharpness of Simulation Construction. When the penultimate process Z_n^* does not lose equicontinuity too fast, and V_0 is sufficiently well-identified, our simulation-based inference procedure becomes sharp in the sense of not only achieving the right level but in fact automatically achieving the right size. In such cases we typically have some small improvements in the rates of convergence of the estimators. The most important case covered is that where V_0 is singleton¹⁴ (or a finite collection of points) and θ_n is locally quadratic, i.e. $\rho_n \geq 2$ and $c_n \geq c > 0$ for all n . These sharp situations occur when the inferential process concentrates on V_0 and not just on the neighborhood V_n , in the sense described below. For this to happen we impose the following condition.

Condition S (Equicontinuity radii are not smaller than r_n). *Under Condition V holding, the scaled penultimate process $\bar{a}_n Z_n^*$ has an equicontinuity radius φ_n that is no smaller than $r_n := 2(\bar{\kappa}_n \bar{\sigma}_n)^{1/\rho_n}/c_n$:*

$$\sup_{\|v-v'\| \leq \varphi_n} \bar{a}_n |Z_n^*(v) - Z_n^*(v')| = o_{\mathbb{P}_n}(1), \quad r_n \leq \varphi_n.$$

When Z_n^* is Donsker, i.e. asymptotically equicontinuous, this condition holds automatically, since in this case $\bar{a}_n \propto 1$, and for any $o(1)$ term, equicontinuity radii obey $\varphi_n = o(1)$, so that consistency $r_n = o(1)$ is sufficient. When Z_n^* is not Donsker, its finite-sample equicontinuity properties decay as $n \rightarrow \infty$, with radii φ_n characterizing the decay. However, as long as φ_n is not smaller than r_n , we have just enough finite-sample equicontinuity left to achieve the following result.

Lemma 3 (Inference Sometimes Concentrates on V_0). *Suppose C.1-C.4, S, and V hold. Then for any $\gamma_n \nearrow 1$,*

$$\mathbb{P}_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \leq x \right) = \mathbb{P}_n \left(\sup_{v \in V_0} Z_n^*(v) \leq x \right) + o(1).$$

Under the stated conditions, our inference and estimation procedures *automatically* become sharp in terms of size and rates.

Theorem 3 (Sharpness of Simulation Inference). *Suppose C.1-C.4, S, and V hold. Consider the interval estimator given in Definition 1 with the critical value function specified in Definition 3. Then, for a given $p \in [1/2, 1)$,*

1. *The interval estimator has asymptotic size p :*

$$\mathbb{P}_n \left\{ \theta_{n0} \leq \hat{\theta}_{n0}(p) \right\} = p + o(1).$$

¹⁴This case is generic in the sense that if one draws $\theta_n(\cdot)$ from the Sobolev ball equipped with the Wiener measure, then V_0 is singleton with probability one.

2. Its estimation risk is bounded by, $wp \rightarrow 1$ under P_n ,

$$\left| \widehat{\theta}_{n0}(p) - \theta_{n0} \right| \leq 4\bar{\sigma}_n \left(a_n(V_0) + \frac{O_{P_n}(1)}{a_n(V_0)} \right) \lesssim_{P_n} \bar{\sigma}_n a_n(V_0).$$

3. Any, possibly data-dependent, alternative $\theta_{na} > \theta_{n0}$ such that

$$\theta_{na} \geq \theta_{n0} + 4\bar{\sigma}_n \left(a_n(V_0) + \frac{\mu_n}{a_n(V_0)} \right), \quad \mu_n \rightarrow_{P_n} \infty,$$

is rejected with probability converging to 1 under P_n .

4. INFERENCE ON INTERSECTION BOUNDS IN LEADING CASES

4.1. Parametric estimation of bounding function. We now show that the above conditions apply to various parametric estimation methods for $v \mapsto \theta_n(v)$. This is an important practical, and indeed tractable, case. The required conditions cover standard parametric estimators of bounding functions such as least squares, quantile regression, and other estimators.

Condition P (Finite-Dimensional Bounding Function). We have that (i) $\theta_n(v) := \theta_n(v, \beta_n)$, where $\mathcal{V} \times \mathcal{B} \mapsto \theta_n(v, \beta)$ is a known function parameterized by finite-dimensional vector $\beta \in \mathcal{B}$, where \mathcal{V} is a compact subset of \mathbb{R}^d and \mathcal{G} is a subset of \mathbb{R}^k , where the sets do not depend on n . (ii) The function $(v, \beta) \mapsto p_n(v, \beta) := \partial \theta_n(v, \beta) / \partial \beta$ is uniformly Lipschitz with Lipschitz coefficient $L_n \leq L$, where L is a finite constant that does not depend on n . (iii) An estimator $\widehat{\beta}_n$ is available such that

$$\Omega_n^{-1/2} \sqrt{n}(\widehat{\beta}_n - \beta_n) = \mathcal{N}_k + o_{P_n}(1), \quad \mathcal{N}_k =_d N(0, I_k),$$

i.e. \mathcal{N}_k is a random k -vector with the multivariate standard normal distribution. (iv) $\|p_n(v, \beta_n)\|$ is bounded away from zero, uniformly in v and n . The eigenvalues of Ω_n are bounded from above and away from zero, uniformly in n . (v) There is also a consistent estimator $\widehat{\Omega}_n$ such that $\|\widehat{\Omega}_n - \Omega_n\| = O_{P_n}(n^{-b})$ for some constant $b > 0$, independent of n .

Example 1 (A Saturated Model). As a simple, but relevant example we consider the following model. Suppose that v takes on a finite set of values, denoted $1, \dots, k$, so that $\theta_n(v, \beta) = \sum_{j=1}^k \beta_j 1(v = j)$. Suppose first that $P_n = P$ is fixed, so that $\beta_n = \beta_0$, a fixed value. Condition (ii) and the boundedness requirement of (iv) follow from $\partial \theta_n(v, \beta) / \partial \beta_j = 1(v = j)$ for each $j = 1, \dots, k$. Condition (v) applies to many estimators. Then if the estimator $\widehat{\beta}$ satisfies $\Omega^{-1/2} \sqrt{n}(\widehat{\beta} - \beta_0) \rightarrow_d N(0, I_k)$ where Ω is positive definite, the strong

approximation in condition (iii) follows from Skorohod's theorem and Lemma 9.¹⁵ Suppose next that P_n and the true value $\beta_n = (\beta_{n1}, \dots, \beta_{nk})'$ change with n . Then if

$$\Omega_n^{-1/2} \sqrt{n}(\hat{\beta}_n - \beta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n u_{i,n} + o_{P_n}(1),$$

with $\{u_{i,n}\}$ i.i.d. with mean zero, for each n , and $E\|u_{i,n}\|^{2+\delta}$ bounded uniformly in n for some $\delta > 0$, then $\Omega_n^{-1/2} \sqrt{n}(\hat{\beta}_n - \beta_n) \rightarrow_d N(0, I_k)$, then again condition (iii) follows from Skorohod's theorem and Lemma 9. \blacksquare

Lemma 4 (P and V imply C.1-C.4, S). *Condition P implies Conditions C.1-C.4, where, for $p_n(v, \beta) := \frac{\partial \theta_n(v, \beta)}{\partial \beta}$,*

$$\begin{aligned} Z_n(v) &= \frac{\theta_n(v) - \hat{\theta}_n(v)}{\sigma_n(v)}, \quad Z_n^*(v) = \frac{p_n(v, \beta_n)' \Omega_n^{1/2}}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \mathcal{N}_k, \quad Z_n^\star(v) = \frac{p_n(v, \hat{\beta}_n)' \hat{\Omega}_n^{1/2}}{\|p_n(v, \hat{\beta}_n)' \hat{\Omega}_n^{1/2}\|} \mathcal{N}_k, \\ \sigma_n(v) &= \|n^{-1/2} p_n(v, \beta_n)' \Omega_n^{1/2}\|, \quad s_n(v) = \|n^{-1/2} p_n(v, \hat{\beta}_n)' \hat{\Omega}_n^{1/2}\|, \quad \delta_n = o(1), \\ \bar{a}_n &\lesssim 1, \quad \bar{\sigma}_n \lesssim \sqrt{1/n}, \quad a_n(V) = \left(2\sqrt{\log\{C(1 + C' L_n \text{diam}(V))^d\}} \right) \vee (1 + \sqrt{d}), \end{aligned}$$

for some positive constants C and C' , and $P[\mathcal{E} > x] = \exp(-x/2)$. Furthermore, if also Condition V holds and $c_n^{-1}(\ell_n/\sqrt{n})^{1/\rho_n} = o(1)$, then Condition S holds.

The following is an immediate consequence of Lemma 4 and Theorems 1, 2, and 3.

Theorem 4 (Estimation and Inference with Parametrically Estimated Bounding Functions). *Suppose Condition P holds and consider the interval estimator $\hat{\theta}_{n0}(p)$ given in Definition 1 with simulation-based critical values specified in Definition 3 for the simulation process Z_n^\star specified above. (1) Then (i) $P_n[\theta_{n0} \leq \hat{\theta}_{n0}(p)] \geq p - o(1)$, (ii) $|\theta_{n0} - \hat{\theta}_{n0}(p)| = O_{P_n}(\sqrt{1/n})$, (iii) $P_n(\theta_{n0} + \mu_n \sqrt{1/n} \geq \hat{\theta}_{n0}(p)) \rightarrow 1$ for any $\mu_n \rightarrow_{P_n} \infty$. (2) If Condition V holds with $c_n \geq c > 0$ and $\rho_n \leq \rho < \infty$, then $P_n[\theta_{n0} \leq \hat{\theta}_{n0}(p)] = p + o(1)$.*

We next provide two examples that generalize the simple, but well-used, saturated example of Example 1 to more substantive cases. Aside from being practically relevant due to the common use of parametric restriction in applications, these examples offer a natural means of transition to the next section, which deals with series estimation, and which can be viewed as parametric estimation with parameters of increasing dimension and vanishing approximation errors.

¹⁵See Theorem 1.10.3 of van der Vaart and Wellner (1996) on page 58 and the subsequent historical discussion attributing the earliest such results to Skorohod (1956), later generalized by Wichura and Dudley.

Example 2 (Linear Bounding Function). Suppose that $\theta_n(v, \beta_n) = p_n(v)' \beta_n$, where $p_n(v)' \beta : \mathcal{V} \times \mathcal{B} \mapsto \mathbb{R}$. Suppose that (a) $v \mapsto p_n(v)$ is Lipschitz with Lipschitz coefficient $L_n \leq L$, for all n , with the first component equal to 1, (b) there is an estimator available that is asymptotically linear

$$\Omega_n^{-1/2} \sqrt{n}(\hat{\beta}_n - \beta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n u_{i,n} + o_{P_n}(1),$$

with $\{u_{i,n}\}$ i.i.d. with mean zero, for each n , and $E\|u_{i,n}\|^{2+\delta}$ bounded uniformly in n for some $\delta > 0$, and (c) Ω_n has eigenvalues bounded away from zero and from above. These conditions imply Condition P(i)-(iv). Indeed, (i), (ii), and (iv) hold immediately, while (iii) follows from the Lindeberg-Feller CLT, which implies that under P_n

$$\Omega_n^{-1/2} \sqrt{n}(\hat{\beta}_n - \beta_n) \rightarrow_d N(0, I_k),$$

and the strong approximation follows by the Skorohod representation and Lemma 9 by suitably enriching the probability space if needed. Note that if $\theta_n(v, \beta_n)$ is the conditional expectation of Y_i given $V_i = v$, then $\hat{\beta}_n$ can be obtained by the mean regression of Y_i on $p_n(V_i)$, $i = 1, \dots, n$; if $\theta_n(v, \beta_n)$ is the conditional u -quantile of Y_i given $V_i = v$, then $\hat{\beta}_n$ can be obtained by the u -quantile regression of Y_i on $p_n(V_i)$, $i = 1, \dots, n$. Regularity conditions that imply the ones stated above can be found in e.g. White (1984) and Koenker (2005). Finally estimators of Ω_n depend on the estimator of β_n ; for mean regression the standard estimator is the Eicker-Huber-White estimator, and for quantile regression the standard estimator is Powell's estimator, see Powell (1984). For brevity we do not restate sufficient conditions for Condition P(v), but these are readily available for common estimators. ■

Example 3 (Conditional Moment Inequalities). This is a generalization of the previous example where now the bounding function is the minimum of J conditional mean functions. Referring to the conditional moment inequality setting specified in Section 2, suppose we have an i.i.d. sample of (X_i, Z_i) , $i = 1, \dots, n$, with $\text{support}(Z_i) = \mathcal{Z} \subseteq [0, 1]^d$. Let $v = (z, j)$, where j denotes the enumeration index for the conditional moment inequality, $j \in \{1, \dots, J\}$, and suppose $\mathcal{V} \subseteq \mathcal{Z} \times \{1, \dots, J\}$. The parameters J and d do not depend on n . Hence

$$\theta_{n0} = \min_{v \in \mathcal{V}} \theta_n(v) = \min_{(z,j) \in \mathcal{V}} \theta_n(z, j).$$

Suppose that $\theta_n(v) = E_{P_n}[m(X, \mu, j)|z] = b(z)' \chi_n(j)$, for $b : \mathcal{Z} \mapsto \mathbb{R}^m$, denoting some transformation of z , with m independent of n , and where $\chi_n(j)$ are the population regression coefficients in the regression of $Y(j) := m(X, \mu, j)$ on $b(Z)$, $j = 1, \dots, J$, respectively, under P_n . Suppose that the first $J_0/2$ pairs correspond to moment inequalities generated from

moment equalities so that $\theta_n(j) = -\theta_n(j-1)$, $j = 2, 4, \dots, J_0$, and so these functions are replicas of each other up to sign; also note that $\chi_n(j) = -\chi_n(j-1)$, $j = 2, 4, \dots, J_0$. Then we can rewrite

$$\begin{aligned}\theta_n(v) &= E_{P_n}[m(X, \mu, j)|Z = z] = b(z, j)' \chi_n(j) := p_n(v)' \beta_n, \\ \beta_n &= (\chi_n(j)', j \in \mathcal{J})', \quad \mathcal{J} := \{2, 4, \dots, J_0, J_0 + 1, J_0 + 2, \dots, J\}',\end{aligned}$$

where β_n is a K -vector of regression coefficients, and $p_n(v)$ is a K -vector such that $p_n(z, j) = [0'_m, \dots, 0'_m, (-1)^{j+1} b'_m(z), 0'_m, \dots, 0'_m]'$ with $b'_m(z)$ appearing in the j -th block for $1 \leq j \leq J_0$; $p_n(z, j) = [0'_m, \dots, 0'_m, b'_m(z), 0'_m, \dots, 0'_m]'$ with $b(z)$ appearing in the j -th block for $J_0 + 1 \leq j \leq J$, where 0_m is an m -dimensional vector of zeroes.¹⁶ Note that this removal of duplicated regressions is done to simplify the technical arguments; it is not necessary for practical implementation.

We impose the following conditions:

- (a) $b(z)$ includes constant 1, (b) $z \mapsto b(z)$ has Lipschitz coefficient bounded above by L , (c) for $Y_i = (Y_i(j), j \in \mathcal{J})'$ and for $\epsilon_i := Y_i - E_{P_n}[Y_i|Z_i]$, the eigenvalues of $E_{P_n}[\epsilon_i \epsilon_i' | Z_i = z]$ are bounded away from zero and from above, uniformly in $z \in \mathcal{Z}$ and n ; (d) $Q = E_{P_n}[b(Z_i)b(Z_i)']$ has eigenvalues bounded away from zero and from above, uniformly in n , and (e) $E_{P_n}\|b(Z_i)\|^4$ and $E_{P_n}\|\epsilon_i\|^4$ are bounded from above uniformly in n .

Then it follows from e.g. by White (1984) that for $\hat{\chi}_n(j)$ denoting the ordinary least square estimator obtained by regressing $Y_i(j), i = 1, \dots, n$, on $b(Z_i), i = 1, \dots, n$,

$$\sqrt{n}(\hat{\chi}_n(j) - \chi_n(j)) = Q^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n b(Z_i) \epsilon_i(j) + o_{P_n}(1), \quad j \in \mathcal{J},$$

so that

$$\sqrt{n}(\hat{\beta}_n - \beta_n) = (I_{|\mathcal{J}|} \otimes Q)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{(I_{|\mathcal{J}|} \otimes b(Z_i)) \epsilon_i}_{u_i} + o_{P_n}(1).$$

By conditions (c) and (d) $E_{P_n}[u_i u_i']$ and Q have eigenvalues bounded away from zero and from above, so the same is true of $\Omega_n = (I_{|\mathcal{J}|} \otimes Q)^{-1} E_{P_n}[u_i u_i'] (I_{|\mathcal{J}|} \otimes Q)^{-1}$. These conditions verify condition P(i),(ii),(iv). Application of the Lindeberg-Feller CLT, Skorohod's theorem, and Lemma 9 verifies Condition P(iii). By the argument given in Chapter VI of White

¹⁶Note the absence of $\chi_n(j)$ for odd j up to J_0 in the definition of the coefficient vector β_n . This is required to enable non-singularity of $E_{P_n}[\epsilon_i \epsilon_i' | Z_i = z]$. Imposing non-singularity simplifies the proofs, and is not needed for practical implementation, as the removal of these indices is not required for estimation and inference, and the estimated variance matrix $\hat{\Omega}_n$ can be allowed to be singular.

(1984), Condition P(v) holds for the standard analog estimator for Ω_n :

$$\hat{\Omega}_n = (I_{|\mathcal{J}|} \otimes \hat{Q})^{-1} \mathbb{E}_n[\hat{u}_i \hat{u}_i'] (I_{|\mathcal{J}|} \otimes \hat{Q})^{-1},$$

where $\hat{Q} = \mathbb{E}_n[b(Z_i)b(Z_i)']$ and $\hat{u}_i = (I_{|\mathcal{J}|} \otimes b(Z_i))\hat{\epsilon}_i$, with $\hat{\epsilon}_i(j) = Y_i(j) - b(Z_i)'\hat{\chi}_n(j)$, and $\hat{\epsilon}_i = (\hat{\epsilon}_i(j), j \in \mathcal{J})'$. \blacksquare

4.2. Nonparametric Estimation of $\theta_n(v)$ 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, i.e. if we undersmooth, then the analysis closely mimics the parametric case.

Condition NS. *The function $v \mapsto \theta_n(v)$ is continuous in v . The series estimator $\hat{\theta}_n(v)$ has the form $\hat{\theta}_n(v) = p_n(v)'\hat{\beta}_n$, where $p_n(v) := (p_{n,1}(v), \dots, p_{n,K_n}(v))'$ is a collection of K_n continuous series functions mapping $\mathcal{V} \subset \mathcal{K} \subset \mathbb{R}^d$ to \mathbb{R}^{K_n} , and $\hat{\beta}_n$ is a K_n -vector of coefficient estimates, and \mathcal{K} is a fixed compact set. Furthermore,*

NS.1 (a) *The estimator satisfies the following linearization and strong approximation condition:*

$$\frac{\hat{\theta}_n(v) - \theta_n(v)}{\|p_n(v)'\Omega_n^{1/2}\|/\sqrt{n}} = \frac{p_n(v)'\Omega_n^{1/2}}{\|p_n(v)'\Omega_n^{1/2}\|} \mathcal{N}_n + R_n(v),$$

where

$$\mathcal{N}_n =_d N(0, I_{K_n}), \quad \sup_{v \in \mathcal{V}} |R_n(v)| = o_{P_n}(1/\log n).$$

(b) *The matrices Ω_n are positive definite, with eigenvalues bounded from above and away from zero, uniformly in n . Moreover, there are sequences of constants ζ_n and ζ'_n such that $1 \leq \zeta'_n \lesssim \|p_n(v)\| \leq \zeta_n$ uniformly for all $v \in \mathcal{V}$ and $\sqrt{\zeta_n^2 \log n/n} \rightarrow 0$, and $\|p_n(v) - p_n(v')\|/\zeta'_n \leq L_n \|v - v'\|$ for all $v, v' \in \mathcal{V}$, where $\log L_n \lesssim \log n$, uniformly in n .*

NS.2 *There exists $\hat{\Omega}_n$ such that $\|\hat{\Omega}_n - \Omega_n\| = O_{P_n}(n^{-b})$, where $b > 0$ is a constant.*

Condition NS is not primitive, but reflects the function-wise large sample normality of series estimators. It requires that the studentized nonparametric process is approximated by a sequence of Gaussian processes, which take a very simple intuitive form, rather than by a fixed single Gaussian process. Indeed, the latter would be impossible in non-parametric settings, since the sequence of Gaussian processes is not asymptotically tight. Note also that the condition implicitly requires that some undersmoothing takes place so that the approximation error is negligible relative to the sampling error. We provide primitive conditions

that imply condition NS.1 in three examples presented below. In particular, we show that the asymptotic linearization for $\hat{\beta}_n - \beta_n$, which is available from the literature on series regression, e.g. from Andrews (1991) and Newey (1997), and the use of Yurinskii's coupling Yurinskii (1977) imply condition NS.1. This result could be of independent interest, although we only provide sufficient conditions for the strong approximation to hold.

Note that under condition NS, the uniform rate of convergence of $\hat{\theta}_n(v)$ to $\theta_n(v)$ is given by $\sqrt{\zeta_n^2/n}\sqrt{\log n} \rightarrow 0$, where $\zeta_n \propto \sqrt{K_n}$ for standard series terms such as B-splines or trigonometric series.

Lemma 5 (NS implies C.1-C.4). *Condition NS implies Conditions C.1-C.4 with*

$$\begin{aligned} Z_n(v) &= \frac{\theta_n(v) - \hat{\theta}_n(v)}{\sigma_n(v)}, \quad Z_n^*(v) = \frac{p_n(v)' \Omega_n^{1/2}}{\|p_n(v)' \Omega_n^{1/2}\|} \mathcal{N}_n, \quad Z_n^\star(v) = \frac{p_n(v)' \hat{\Omega}_n^{1/2}}{\|p_n(v)' \hat{\Omega}_n^{1/2}\|} \mathcal{N}_n, \\ \sigma_n(v) &= \|n^{-1/2} p_n(v)' \Omega_n^{1/2}\|, \quad s_n(v) = \|n^{-1/2} p_n(v)' \hat{\Omega}_n^{1/2}\|, \quad \delta_n = 1/\log n, \\ \bar{a}_n &\lesssim \sqrt{\log n}, \quad \bar{\sigma}_n \lesssim \sqrt{\zeta_n^2/n}, \quad a_n(\mathbf{V}) = \left(2\sqrt{\log\{C(1 + C' L_n \text{diam}(\mathbf{V}))^d\}} \right) \vee (1 + \sqrt{d}), \end{aligned}$$

for some constants C and C' , where $\text{diam}(\mathbf{V})$ denotes the diameter of the set \mathbf{V} , and $P[\mathcal{E} > x] = \exp(-x/2)$.

Remark 2. Lemma 5 verifies the main conditions C.1-C.4. These conditions enable construction of simulated or analytical critical values. For the latter, the p -th quantile of \mathcal{E} is given by $c(p) = -2\log(1 - p)$, so we can set

$$k_{n,\mathbf{V}}(p) = a_n(\mathbf{V}) - 2\log(1 - p)/a_n(\mathbf{V}), \quad (4.1)$$

where

$$a_n(\mathbf{V}) = \left(2\sqrt{\log\{\ell_n (1 + \ell_n L_n \text{diam}(\mathbf{V}))^d\}} \right), \quad (4.2)$$

is a feasible scaling factor which bounds the scaling factor in the statement of Lemma 5, at least for all large n . Here, all unknown constants have been replaced by slowly growing numbers ℓ_n such that $\ell_n > C \vee C'$ for all large n . Note also that $\mathbf{V} \mapsto k_{n,\mathbf{V}}(p)$ is monotone in \mathbf{V} for all sufficiently large n , as required in the analytical construction given in Definition 2. A sharper analytical approach can be based on Hotelling's tube method; for details we refer to Chernozhukov, Lee, and Rosen (2009). That approach is tractable for the case of $d = 1$ but does not immediately extend to $d > 1$. Note that the simulation-based approach is effectively a numeric version of the exact version of the tube formula, and is less conservative than using simplified tube formulas. ■

Lemma 6 (Condition NS implies S in some cases). *Suppose Condition NS holds. Then, (1) The radius φ_n of equicontinuity of Z_n^* obeys:*

$$\varphi_n \leq o(1) \cdot \left(\frac{1}{L_n \sqrt{\log n}} \right),$$

for any $o(1)$ term. (2) If Condition V holds and

$$\left(\sqrt{\frac{\zeta_n^2}{n} \log n} \right)^{1/\rho_n} c_n^{-1} = o \left(\frac{1}{L_n \sqrt{\log n}} \right), \quad (4.3)$$

then Condition S holds. (3) If V_0 is singleton and (4.3) holds, $\rho_n \leq 2$, and $c_n \geq c > 0$, for all n , $\zeta_n \lesssim \sqrt{K_n}$ and $L_n \lesssim K_n$, we have $a_n(V_0) \propto 1$ and (4.3) reduces to

$$K_n^5 \log^3 n / n \rightarrow 0.$$

The following is an immediate consequence of Lemmas 5 and 6 and Theorems 1, 2, and 3.

Theorem 5 (Estimation and Inference with Series-Estimated Bounding Functions). *Suppose Condition NS holds and consider the interval estimator $\hat{\theta}_{n0}(p)$ given in Definition 1 with either analytical critical value $c(p) = -2 \log(1 - p)$, or simulation-based critical values from Definition 3 for the simulation process Z_n^* above. (1) Then (i) $P_n[\theta_{n0} \leq \hat{\theta}_{n0}(p)] \geq p - o(1)$, (ii) $|\theta_{n0} - \hat{\theta}_{n0}(p)| = O_{P_n}(\sqrt{\log n} \sqrt{\zeta_n^2/n})$, (iii) $P_n(\theta_{n0} + \mu_n \sqrt{\log n} \sqrt{\zeta_n^2/n} \geq \hat{\theta}_{n0}(p)) \rightarrow 1$ for any $\mu_n \rightarrow_{P_n} \infty$. (2) Moreover, for the simulation-based critical values, if Condition V and relation (4.3) hold, then (i) $P_n[\theta_{n0} \leq \hat{\theta}_{n0}(p)] = p - o(1)$, (ii) $|\theta_{n0} - \hat{\theta}_{n0}(p)| = O_{P_n}(\sqrt{\zeta_n^2/n})$, (iii) $P_n(\theta_{n0} + \mu_n \sqrt{\zeta_n^2/n} \geq \hat{\theta}_{n0}(p)) \rightarrow 1$ for any $\mu_n \rightarrow_{P_n} \infty$.*

We next present some examples with primitive conditions that imply Condition NS.

Example 4 (Bounding Function is Conditional Quantile). Suppose that $\theta_n(v) := Q_{Y_i|V_i}[\tau|v]$ is the τ -th conditional quantile of Y_i given V_i under P_n , assumed to be a continuous function in v . Suppose we estimate $\theta_n(v)$ with a series estimator. There is an i.i.d. sample $(Y_i, V_i), i = 1, \dots, n$, with $\text{support}(V_i) \subseteq [0, 1]^d$ for each n , defined on a probability space equipped with probability measure P_n . Suppose that the intersection region of interest is $\mathcal{V} \subseteq \text{support}(V_i)$. Here the index d does not depend on n , but all other parameters, unless stated otherwise, can depend on n . Then $\theta_n(v) = p_n(v)' \beta_n + A_n(v)$, where $p_n : [0, 1]^d \mapsto \mathbb{R}^{K_n}$ are the series functions, β_n is the quantile regression coefficient in the population, $A_n(v)$ is the approximation error, and K_n is the number of series terms that depends on n . Let C be a positive constant.

We impose the following technical conditions to verify NS.1 and NS.2:

Uniformly in n , (i) p_n are either b-splines of a fixed order or trigonometric series terms or any other terms $p_n = (p_{n1}, \dots, p_{nK_n})$ such that $\|p_n(v)\| \lesssim \zeta_n = \sqrt{K_n}$ for all $v \in \text{support}(V_i)$, $E_{P_n}[\|p_n(V_i)\|^3] \lesssim K_n^{3/2}$, $\|p_n(v)\| \gtrsim \zeta'_n \geq 1$ for all $v \in \mathcal{V}$, and $\log \text{lip}(p_n) \lesssim \log K_n$, (ii) the mapping $v \mapsto \theta_n(v)$ is sufficiently smooth, namely $\sup_{v \in \mathcal{V}} |A_n(v)| \lesssim K_n^{-s}$, for some $s > 0$, (iii) $\lim_{n \rightarrow \infty} (\log n)^c K_n^{-s+1} = 0$ and $\lim_{n \rightarrow \infty} (\log n)^c \sqrt{n} K_n / \zeta'_n = 0$, for each $c > 0$, (iv) eigenvalues of $\Lambda_n = E_{P_n}[p_n(V_i)p_n(V_i)']$ are bounded away from zero and from above, (v) $f_{Y_i|V_i}(\theta_n(v)|v)$ is bounded uniformly over $v \in \mathcal{V}$ away from zero and from above, (vi) $\lim_{n \rightarrow \infty} K_n^5 (\log n)^c / n = 0$ for each $c > 0$, and (vii) the restriction on the bandwidth sequence in Powell's estimator \hat{Q}_n of $Q_n = E_{P_n}[f_{Y_i|V_i}(\theta_n(V_i)|V_i)p_n(V_i)p_n(V_i)']$ specified in Belloni, Chernozhukov, and Fernandez-Val (2011) holds.

Suppose that we use the standard quantile regression estimator

$$\hat{\beta}_n = \arg \min_{b \in \mathbb{R}^{K_n}} \mathbb{E}_n[\rho_\tau(Y_i - p_n(V_i)'b)],$$

so that $\hat{\theta}_n(v) = p_n(v)'\hat{\beta}$ for $\rho_\tau(u) = (\tau - 1(u < 0))u$. Then by Belloni, Chernozhukov, and Fernandez-Val (2011), under conditions (i)-(vi), the following asymptotically linear representation holds:

$$\sqrt{n}(\hat{\beta}_n - \beta_n) = Q_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{p_n(V_i)\epsilon_i}_{u_i} + o_{P_n} \left(\frac{1}{\log n} \right),$$

for $\epsilon_i = (\tau - 1(w_i \leq \tau))$, where $(w_i, i = 1, \dots, n)$ are i.i.d. uniform, independent of $(V_i, i = 1, \dots, n)$. Note that by conditions (iv) and (v) $S_n := E_{P_n}[u_i u_i'] = \tau(1 - \tau)\Lambda_n$, and J_n have eigenvalues bounded away from zero and from above uniformly in n , and so the same is also true of $\Omega_n = Q_n^{-1} S_n Q_n^{-1}$. Given other restrictions imposed in condition (i), Condition NS.1(b) is verified. Next using condition (iv) and others the strong approximation required in NS.1(a) follows by invoking Theorem 7, Corollary 1 in Section 5, which is based on Yurinskii's coupling. To verify Condition NS.2, consider the plug-in estimator $\hat{\Omega}_n = \hat{J}_n^{-1} \hat{S}_n \hat{J}_n^{-1}$, where \hat{J}_n is the Powell's estimator for J_n , and $\hat{S}_n = \tau(1 - \tau) \cdot \mathbb{E}_n[p_n(V_i)p_n(V_i)']$. Then under condition (vii) it follows from the proof of Theorem 7 in Belloni, Chernozhukov, and Fernandez-Val (2011) that $\|\hat{\Omega}_n - \Omega_n\| = O_{P_n}(n^{-b})$ for some $b > 0$. ■

Example 5 (Bounding Function is Conditional Mean). Now suppose that $\theta_n(v) = E_{P_n}[Y_i|V_i = v]$, assumed to be a continuous function with respect to $v \in \text{support}(V_i)$, and the intersection region is $\mathcal{V} \subseteq \text{support}(V_i)$. Suppose we are using the series approach to approximating and estimating $\theta_n(v)$. There is an i.i.d. sample $(Y_i, V_i), i = 1, \dots, n$, with

$\text{support}(V_i) \subseteq [0, 1]^d$ for each n . Here d does not depend on n , but all other parameters, unless stated otherwise, can depend on n . Then we have $\theta_n(v) = p_n(v)' \beta_n + A_n(v)$, for $p_n : [0, 1]^d \mapsto \mathbb{R}^{K_n}$ representing the series functions; β_n is the coefficient of the best least squares approximation to $\theta_n(v)$ in the population, and $A_n(v)$ is the approximation error. The number of series terms K_n depends on n .

We impose the following technical conditions:

Uniformly in n , (i) p_n are either b-splines of a fixed order or trigonometric series terms or any other terms $p_n = (p_{n1}, \dots, p_{nK_n})$ such that $\|p_n(v)\| \lesssim \zeta_n = \sqrt{K_n}$ for all $v \in \text{support}(V_i)$, $E_{P_n}[\|p_n(V_i)\|^3] \lesssim K_n^{3/2}$, $\|p_n(v)\| \gtrsim \zeta'_n \geq 1$ for all $v \in \mathcal{V}$, and $\log \text{lip}(p_n) \lesssim \log K_n$, (ii) the mapping $v \mapsto \theta_n(v)$ is sufficiently smooth, namely $\sup_{v \in \mathcal{V}} |A_n(v)| \lesssim K_n^{-s}$, for some $s > 0$, (iii) $\lim_{n \rightarrow \infty} (\log n)^c \sqrt{n} K_n^{-s} = 0$ for each $c > 0$,¹⁷ (iv) for $\epsilon_i = Y_i - E_{P_n}[Y_i|V_i]$, $E_{P_n}[\epsilon_i^2|V_i = v]$ is bounded away from zero uniformly in $v \in \text{support}(V_i)$, and (v) eigenvalues of $Q_n = E_{P_n}[p_n(V_i)p_n(V_i)']$ are bounded away from zero and from above, and (vi) $E_{P_n}[|\epsilon_i|^4|V_i = v]$ is bounded from above uniformly in $v \in \text{support}(V_i)$, (vii) $\lim_{n \rightarrow \infty} (\log n)^c K_n^5/n = 0$ for each $c > 0$.

We use the standard least squares estimator

$$\hat{\beta}_n = \mathbb{E}_n[p_n(V_i)p_n(V_i)']^{-1} \mathbb{E}_n[p_n(V_i)Y_i],$$

so that $\hat{\theta}_n(v) = p_n(v)' \hat{\beta}_n$. Then by Newey (1997), under conditions implied by (i)-(vii), we have the following asymptotically linear representation:

$$\sqrt{n}(\hat{\beta}_n - \beta_n) = Q_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{p_n(V_i) \epsilon_i}_{u_i} + o_{P_n}(1/\log n).$$

For details, see Supplementary Appendix I. Note that $E_{P_n}(u_i u_i')$ and Q_n have eigenvalues bounded away from zero and from above uniformly in n , and so the same is also true of $\Omega_n = Q_n^{-1} E_{P_n}(u_i u_i') Q_n^{-1}$. Thus, under condition (i), Condition NS.1(a) is verified. The strong approximation condition NS.1(a) now follows from invoking Theorem 7 in Section 5. Finally, Newey (1997) verifies that NS.2 holds for the standard analog estimator $\hat{\Omega}_n = \hat{Q}_n^{-1} \mathbb{E}_n(\hat{u}_i \hat{u}_i') \hat{Q}_n^{-1}$ for $\hat{u}_i = p_n(V_i)(Y_i - \hat{\theta}_n(V_i))$ and $\hat{Q}_n = \mathbb{E}_n[p_n(V_i)p_n(V_i)']$ under conditions that are implied by those above.

¹⁷This condition, which is based on Newey (1997) can be relaxed to $(\log n)^c K_n^{-s+1} \rightarrow 0$ and $(\log n)^c \sqrt{n} K_n^{-s}/\zeta'_n \rightarrow 0$, using the recent results of Belloni, Chen, and Chernozhukov (2010) for least squares series estimators.

Finally, note that if we had $\epsilon_i \sim N(0, \sigma^2(V_i))$, conditional on V_i , we could establish Condition NS.1 with a much weaker growth restriction than (vii). Thus, while our use of Yurinskii's coupling provides concrete sufficient conditions for strong approximation, the function-wise large sample normality is likely to hold even under weaker conditions in many situations. \blacksquare

Example 6 (Bounding Function from Conditional Moment Inequalities). Consider now Example C of Section 2, where now the bounding function is the minimum of J conditional mean functions. Suppose we have an i.i.d. sample of $(X_i, Z_i), i = 1, \dots, n$, with $\text{support}(Z_i) = \mathcal{Z} \subseteq [0, 1]^d$, defined on a probability space equipped with probability measure P_n . Let $v = (z, j)$, where j denotes the enumeration index for the conditional moment inequality, $j \in \{1, \dots, J\}$, and $\mathcal{V} \subseteq \mathcal{Z} \times \{1, \dots, J\}$. The parameters J and d do not depend on n . Hence

$$\theta_{n0} = \min_{v \in \mathcal{V}} \theta_n(v),$$

for $\theta_n(v) = E_{P_n}[m(X_i, \mu, j)|Z_i = z]$, assumed to be a continuous function with respect to $z \in \mathcal{Z}$. Suppose we use the series approach to approximate and estimate $\theta_n(z, j)$ for each j . Then $E_{P_n}[m(X, \mu, j)|z] = b_n(z)' \chi_n(j) + A_n(z, j)$, for $b_n : [0, 1]^d \mapsto \mathbb{R}^{m_n}$ denoting a m_n -vector of series functions; $\chi_n(j)$ is the coefficient of the best least squares approximation to $E_{P_n}[m(x, \mu, j)|z]$ in the population, and $A_n(z, j)$ is the approximation error. Let \mathcal{J} be a subset of $\{1, \dots, J\}$ as defined as in the parametric Example 3 (to handle inequalities associated with equalities).

We impose the following conditions:

Uniformly in n , (i) $b_n(z)$ are either b-splines of a fixed order or trigonometric series terms or any other terms $b_n(z) = (b_{n1}(z), \dots, b_{nm_n}(z))'$ such that $\|b_n(z)\| \lesssim \zeta_n = \sqrt{m_n}$ for all $z \in \mathcal{Z}$, $E_{P_n}[\|b_n(Z_i)\|^3] \lesssim m_n^{3/2}$, $\|b_n(z)\| \gtrsim \zeta'_n \geq 1$ for all $z \in \mathcal{Z}$, and $\log \text{lip}(b_n(z)) \lesssim \log m_n$; (ii) the mapping $z \mapsto \theta_n(z, j)$ is sufficiently smooth, namely $\sup_{z \in \mathcal{Z}} |A_n(z, j)| \lesssim m_n^{-s}$, for some $s > 0$, for all $j \in \mathcal{J}$; (iii) $\lim_{n \rightarrow \infty} (\log n)^c \sqrt{n} m_n^{-s} = 0$ for each $c > 0$; ¹⁸ (iv) for $Y(j) := m(X, \mu, j)$ and $Y_i := (Y_i(j), j \in \mathcal{J})'$ and $\epsilon_i := Y_i - E_{P_n}[Y_i|Z_i]$, the eigenvalues of $E_{P_n}[\epsilon_i \epsilon_i' | Z_i = z]$ are bounded away from zero, uniformly in $z \in \mathcal{Z}$; (v) eigenvalues of $Q_n = E_{P_n}[b_n(Z_i) b_n(Z_i)']$ are bounded away from zero and from above; (vi) $E_{P_n}[\|\epsilon_i\|^4 | Z_i = z]$ is bounded above, uniformly in $z \in \mathcal{Z}$; and (vii) $\lim_{n \rightarrow \infty} m_n^5 (\log n)^c / n = 0$ for each $c > 0$.

¹⁸See the previous footnote on a possible relaxation of this condition.

The above construction implies $\theta_n(v) = b_n(z)' \chi_n(j) + A_n(z, j) =: p_n(v)' \beta_n + A_n(v)$, for $\beta_n = (\chi_n'(j), j \in \mathcal{J})'$, where $p_n(v)$ and β_n are vectors of dimension $K_n := m_n \times |\mathcal{J}|$. Consider the standard least squares estimator $\hat{\beta}_n = (\hat{\chi}_n'(j), j \in \mathcal{J})'$ consisting of $|\mathcal{J}|$ least square estimators, where $\hat{\chi}_n(j) = \mathbb{E}_n[b_n(Z_i)b_n(Z_i)']^{-1} \mathbb{E}_n[b_n(Z_i)Y_i(j)]$. Then it follows from Newey (1997) that for $Q_n = E_{P_n}[b_n(Z_i)b_n(Z_i)']^{-1}$

$$\sqrt{n}(\hat{\chi}_n(j) - \chi_n(j)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Q_n^{-1} b_n(Z_i) \epsilon_i(j) + o_{P_n}(1/\log n), \quad j \in \mathcal{J},$$

so that

$$\sqrt{n}(\hat{\beta}_n - \beta_n) = (I_{|\mathcal{J}|} \otimes Q_n)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{(I_{|\mathcal{J}|} \otimes b_n(Z_i)) \epsilon_i}_{u_i} + o_{P_n}(1/\log n).$$

By conditions (iv), (v), and (vi) $E_{P_n}[u_i u_i']$ and Q_n have eigenvalues bounded away from zero and from above, so the same is true of $\Omega_n = (I_{|\mathcal{J}|} \otimes Q_n)^{-1} E_{P_n}[u_i u_i'] (I_{|\mathcal{J}|} \otimes Q_n)^{-1}$. This and condition (i) imply that Condition NS.1(b) holds. Application of Theorem 7, based on Yurinskii's coupling, verifies Condition NS.1(a). Finally, Condition NS.2 holds for the standard plug-in estimator for Ω_n , by the same argument as given in the proof of Theorem 2 of Newey (1997). \blacksquare

4.3. Nonparametric Estimation of $\theta_n(v)$ via Kernel Methods. In this section we provide conditions under which kernel-type estimators satisfy Conditions C.1-C.4. These conditions cover both standard kernel estimators as well as local polynomial estimators.

Condition NK. Let $v = (z, j)$ and $\mathcal{V} \subseteq \mathcal{Z} \times \{1, \dots, J\}$, where \mathcal{Z} is a compact convex set that does not depend on n . The estimator $v \mapsto \hat{\theta}_n(v)$ and the function $v \mapsto \theta_n(v)$ are continuous in v . In what follows, let e_j denote the J -vector with j th element one and all other elements zero. Suppose that (U, Z) is a $(J + d)$ -dimensional random vector, where U is a generalized residual such that $E[U|Z] = 0$ a.s. and Z is a covariate; the density f_n of Z is continuous and bounded away from zero and from above on \mathcal{Z} , uniformly in n ; and the support of U is bounded uniformly in n . \mathbf{K} is a twice continuously differentiable, possibly higher-order, product kernel function with support on $[-1, 1]^d$, $\int \mathbf{K}(u) du = 1$; and h_n is a sequence of bandwidths such that $h_n \rightarrow 0$ and $nh_n^d \rightarrow \infty$ at a polynomial rate in n .

NK.1 We have that uniformly in $v \in \mathcal{V}$,

$$(nh_n^d)^{1/2}(\hat{\theta}_n(v) - \theta_n(v)) = \mathbb{B}_n(g_v) + o_{P_n}(\delta_n), \quad g_v(U, Z) := \frac{e_j' U}{(h_n^d)^{1/2} f_n(z)} \mathbf{K}\left(\frac{z - Z}{h_n}\right),$$

where \mathbb{B}_n is a P_n -Brownian bridge such that $v \mapsto \mathbb{B}_n(g_v)$ has continuous sample paths over \mathcal{V} . Moreover, the latter process can be approximated via the Gaussian multiplier method,

namely there exist sequences $o(\delta_n)$ and $o(1/\ell_n)$ such that

$$P_n \left(\sup_{v \in \mathcal{V}} |\mathbb{G}_n^o(g_v) - \bar{\mathbb{B}}_n(g_v)| > o(\delta_n) \middle| \mathcal{D}_n \right) = o_{P_n}(1/\ell_n),$$

for some independent copy $v \mapsto \bar{\mathbb{B}}_n(g_v)$ of the process $v \mapsto \mathbb{B}_n(g_v)$. Here, $\mathbb{G}_n^o(g_v) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i g_v(U_i, Z_i)$, where η_i are i.i.d. $N(0, 1)$, independent of the data \mathcal{D}_n and of $\{(U_i, Z_i)\}_{i=1}^n$, which are i.i.d. copies of (U, Z) . Covariates $\{Z_i\}_{i=1}^n$ are part of the data.

NK.2 There exists an estimator $z \mapsto \hat{f}_n(z)$, having continuous sample paths, such that $\sup_{z \in \mathcal{Z}} |\hat{f}_n(z) - f_n(z)| = O_{P_n}(n^{-b})$, and there are estimators \hat{U}_i of generalized residuals such that $\max_{1 \leq i \leq n} \|\hat{U}_i - U_i\| = O_{P_n}(n^{-\tilde{b}})$ for some constants $b > 0$ and $\tilde{b} > 0$.

Condition NK.1 is a high-level condition that captures the large sample Gaussianity of the entire estimated function where estimation is done via a kernel or local method. Under some mild regularity conditions, specifically those stated in Appendix G, NK.1 follows from the Rio-Massart coupling and from the Bahadur expansion holding uniformly in $v \in \mathcal{V}$:

$$(nh_n^d)^{1/2}(\hat{\theta}_n(v) - \theta_n(v)) = \mathbb{G}_n(g_v) + o_{P_n}(\delta_n).$$

Uniform Bahadur expansions have been established for a variety of local estimators, see e.g. Masry (1996) and Kong, Linton, and Xia (2010), including higher-order kernel and local polynomial estimators. It is possible to use more primitive sufficient conditions stated in Appendix G based on the Rio-Massart coupling, but these conditions are merely sufficient and other primitive conditions may also be adequate. Our general argument, however, relies only on validity of Condition NK.1.

For simulation purposes, we define

$$\begin{aligned} \mathbb{G}_n^o(\hat{g}_v) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \hat{g}_v(U_i, Z_i), \quad \eta_i \text{ i.i.d. } N(0, 1), \text{ independent of the data } \mathcal{D}_n, \\ \hat{g}_v(U_i, Z_i) &= \frac{e_j' \hat{U}_i}{(h_n^d)^{1/2} \hat{f}_n(z)} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right). \end{aligned}$$

Lemma 7 (Condition NK implies C.1-C.4). *Condition NK implies C.1-C.4 with $v = (z, j) \in \mathcal{V} \subseteq \mathcal{Z} \times \{1, \dots, J\}$,*

$$\begin{aligned} Z_n(v) &= \frac{\theta_n(v) - \hat{\theta}_n(v)}{\sigma_n(v)}, \quad Z_n^*(v) = \frac{\mathbb{B}_n(g_v)}{\sqrt{E_{P_n}[g_v^2]}}, \quad Z_n^\star(v) = \frac{\mathbb{G}_n^\circ(\hat{g}_v)}{\sqrt{\mathbb{E}_n[\hat{g}_v^2]}}, \\ \sigma_n^2(v) &= E_{P_n}[g_v^2]/(nh_n^d), \quad s_n^2(v) = \mathbb{E}_n[\hat{g}_v^2]/(nh_n^d), \quad \delta_n = 1/\log n, \\ \bar{a}_n &\lesssim \sqrt{\log n}, \quad \bar{\sigma}_n \lesssim \sqrt{1/(nh^d)}, \quad \text{and} \\ a_n(\mathbf{V}) &= \left(2\sqrt{\log\{C(1 + C'(1 + h_n^{-1})\text{diam}(\mathbf{V}))^d\}} \right) \vee (1 + \sqrt{d}), \end{aligned}$$

for some constants C and C' , where $\text{diam}(\mathbf{V})$ denotes the diameter of the set \mathbf{V} . Moreover, $P[\mathcal{E} > x] = \exp(-x/2)$.

Remark 3. Lemma 7 verifies the main conditions C.1-C.4. These conditions enable construction of either simulated or analytical critical values. For the latter, the p -th quantile of \mathcal{E} is given by $c(p) = -2\log(1 - p)$, so we can set

$$k_{n,\mathbf{V}}(p) = a_n(\mathbf{V}) - 2\log(1 - p)/a_n(\mathbf{V}), \quad (4.4)$$

where

$$a_n(\mathbf{V}) = \left(2\sqrt{\log\{\ell_n (1 + \ell_n(1 + h_n^{-1})\text{diam}(\mathbf{V}))^d\}} \right), \quad (4.5)$$

is a feasible version of the scaling factor, in which unknown constants have been replaced by the slowly growing sequence ℓ_n . Note that $\mathbf{V} \mapsto k_{n,\mathbf{V}}(p)$ is monotone in \mathbf{V} for large n , as required in the analytical construction given in Definition 2. A sharper analytical approach can be based on Hotelling's tube method or on the use of extreme value theory. For details of the extreme value approach, we refer the reader to Chernozhukov, Lee, and Rosen (2009). Note that the simulation-based approach is effectively a numeric version of the exact version of the tube formula, and is less conservative than using simplified tube formulas. In Chernozhukov, Lee, and Rosen (2009) we established that inference based on extreme value theory is valid, but the asymptotic approximation is accurate only when sets \mathbf{V} are “large”, and does not seem to provide an accurate approximation when \mathbf{V} is small. Moreover, it often requires a very large sample size for accuracy even when \mathbf{V} is large. ■

Lemma 8 (Condition NK implies S in some cases). *Suppose Condition NK holds. Then (1) The radius φ_n of equicontinuity of Z_n^* obeys:*

$$\varphi_n \leq o(1) \cdot \left(\frac{h_n}{\sqrt{\log n}} \right),$$

for any $o(1)$ term. (2) If Condition V holds and

$$\left(\sqrt{\frac{\log n}{nh^d}} \log n \right)^{1/\rho_n} c_n^{-1} = o\left(\frac{h_n}{\sqrt{\log n}} \right), \quad (4.6)$$

then Condition S holds.

The following is an immediate consequence of Lemmas 7 and 8 and Theorems 1, 2, and 3.

Theorem 6 (Estimation and Inference for Bounding Functions Using Local Methods). *Suppose Condition NK holds and consider the interval estimator $\hat{\theta}_{n0}(p)$ given in Definition 1 with either analytical critical values specified in Remark 3 or simulation-based critical values given in Definition 3 for the simulation process Z_n^\star specified above. (1) Then (i) $P_n[\theta_{n0} \leq \hat{\theta}_{n0}(p)] \geq p - o(1)$, (ii) $|\theta_{n0} - \hat{\theta}_{n0}(p)| = O_{P_n}\left(\sqrt{\log n/(nh_n^d)}\right)$, (iii) $P_n(\theta_{n0} + \mu_n \sqrt{\log n/(nh_n^d)} \geq \hat{\theta}_{n0}(p)) \rightarrow 1$ for any $\mu_n \rightarrow_{P_n} \infty$. (2) Moreover, for simulation-based critical values, if condition V and (4.6) hold, then (i) $P_n[\theta_{n0} \leq \hat{\theta}_{n0}(p)] = p - o(1)$, (ii) $|\theta_{n0} - \hat{\theta}_{n0}(p)| = O_{P_n}(\sqrt{1/(nh_n^d)})$, (iii) $P_n(\theta_{n0} + \mu_n \sqrt{1/(nh_n^d)} \geq \hat{\theta}_{n0}(p)) \rightarrow 1$ for any $\mu_n \rightarrow_{P_n} \infty$.*

In Supplementary Appendix F we provide an example where the bounding function is obtained from conditional moment inequalities, and where Condition NK holds under primitive conditions. We provide only one example for brevity, but more examples can be covered as for series estimation in Section 4.2. In Supplementary Appendix G we provide conditions under which the required strong approximation in Condition NK.1 holds.

5. STRONG APPROXIMATION FOR ASYMPTOTICALLY LINEAR SERIES ESTIMATORS

In the following theorem we establish strong approximation for series estimators appearing in the previous section as part of Condition NS.1. In Appendix I of the on-line supplement we demonstrate as a leading example how the required asymptotically linear representation can be achieved from primitive conditions for the case of estimation of a conditional mean function.

Theorem 7 (Strong Approximation For Asymptotically Linear Series Estimators). *Let (A, \mathcal{A}, P_n) be the probability space for each n , and let $n \rightarrow \infty$. Let $\delta_n \rightarrow 0$ be a sequence of constants converging to 0 at no faster than a polynomial rate in n . Assume (a) the series estimator has the form $\hat{\theta}_n(v) = p_n(v)' \hat{\beta}_n$, where $p_n(v) := (p_{n,1}(v), \dots, p_{n,K_n}(v))$ is a collection of K_n -dimensional approximating functions such that $K_n \rightarrow \infty$ and $\hat{\beta}_n$ is a K_n -vector of estimates; (b) The estimator $\hat{\beta}_n$ satisfies an asymptotically linear representation*

around some K_n -dimensional vector β_n

$$\Omega_n^{-1/2} \sqrt{n}(\hat{\beta}_n - \beta_n) = n^{-1/2} \sum_{i=1}^n u_{i,n} + r_n, \quad \|r_n\| = o_{P_n}(\delta_n), \quad (5.1)$$

$$u_{i,n}, i = 1, \dots, n \text{ are independent with } E_{P_n}[u_{i,n}] = 0, E_{P_n}[u_{i,n}u'_{i,n}] = I_{K_n}, \text{ and} \quad (5.2)$$

$$\Delta_n = \sum_{i=1}^n E\|u_{i,n}\|^3 / n^{3/2} \text{ such that } K_n \Delta_n / \delta_n^3 \rightarrow 0, \quad (5.3)$$

where Ω_n is a sequence of $K_n \times K_n$ invertible matrices. (c) The function $\theta_n(v)$ admits the approximation $\theta_n(v) = p_n(v)' \beta_n + A_n(v)$, where the approximation error $A_n(v)$ satisfies $\sup_{v \in \mathcal{V}} \sqrt{n} |A_n(v)| / \|g_n(v)\| = o(\delta_n)$, for $g_n(v) := p_n(v)' \Omega_n^{-1/2}$. Then we can find a random normal vector $\mathcal{N}_n =_d \mathcal{N}(0, I_{K_n})$ such that $\|\Omega_n^{-1/2} \sqrt{n}(\hat{\beta}_n - \beta_n) - \mathcal{N}_n\| = o_{P_n}(\delta_n)$ and

$$\sup_{v \in \mathcal{V}} \left| \frac{\sqrt{n}(\hat{\theta}_n(v) - \theta_n(v))}{\|g_n(v)\|} - \frac{g_n(v)}{\|g_n(v)\|} \mathcal{N}_n \right| = o_{P_n}(\delta_n).$$

The following corollary covers the cases considered in the examples of the previous section.

Corollary 1 (A Leading Case of Influence Function). *Suppose the conditions of Theorem 7 hold with $u_{i,n} := \Omega_n^{-1/2} Q_n^{-1} p_n(V_i) \epsilon_i$, where (V_i, ϵ_i) are i.i.d. with $E_{P_n}[\epsilon_i p_n(V_i)] = 0$, $S_n := E_{P_n}[\epsilon_i^2 p_n(V_i) p_n(V_i)']$, and $\Omega_n := Q_n^{-1} S_n (Q_n^{-1})'$, where Q_n^{-1} is a non-random invertible matrix, and $\|\Omega_n^{-1/2} Q_n^{-1}\| \leq \tau_n$; $E_{P_n}[|\epsilon_i|^3 | V_i = v]$ is bounded above uniformly in $v \in \text{support}(V_i)$, and $E_{P_n}[\|p_n(V_i)\|^3] \leq C_n K_n^{3/2}$. Then, the key growth restriction on the number of series terms $K_n \Delta_n / \delta_n^3 \rightarrow 0$ holds if $\tau_n^6 C_n^2 K_n^5 / (n \delta_n^6) \rightarrow 0$.*

Remark 4 (Applicability). In this paper $\delta_n = 1/\log n$. Sufficient conditions for linear approximation (b) follow from results in the literature on series estimation, e.g. Andrews (1991), Newey (1995), and Newey (1997), and Belloni, Chernozhukov, and Fernandez-Val (2011). See also Chen (2007) and references therein for a general overview of sieve estimation and recent developments. The main text provides several examples, including mean and quantile regression, with primitive conditions that provide sufficient conditions for the linear approximation. ■

6. IMPLEMENTATION

In Section 6.1 we lay out steps for implementation of parametric and series estimation of bounding functions, while in Section 6.2 we provide implementation steps for kernel-type estimation. The end goal in each case is to obtain estimators $\hat{\theta}_{n0}(p)$ that provide bias-corrected estimates or the endpoints of confidence intervals depending on the chosen value

of p , e.g. $p = 1/2$ or $p = 1 - \alpha$. As before, we focus here on the upper bound. If instead $\widehat{\theta}_{n0}(p)$ were the lower bound for θ^* , given by the supremum of a bounding function, the same algorithm could be applied to perform inference on $-\theta^*$, bounded above by the infimum of the negative of the original bounding function, and then any inference statements for $-\theta^*$ could trivially be transformed to inference statements for θ^* . Indeed, any set of lower and upper bounds can be similarly transformed to a collection of upper bounds, and the above algorithm applied to perform inference on θ^* , e.g. according to the methods laid out for inference on parameters bounded by conditional moment inequalities in Section 3.¹⁹ Alternatively, if one wishes to perform inference on the identified set in such circumstances one can use the intersection of upper and lower one-sided intervals each based on $\tilde{p} = (1 + p)/2$ as an asymptotic level- p confidence set for Θ_I , which is valid by Bonferroni's inequality.²⁰

6.1. Parametric and Series Estimators. Let β_n denote the bounding function parameter vector if parametric estimation is used, while β_n denotes the coefficients of the series terms if series estimation is used, as in Section 4.2. K denotes the dimension of β_n and I_K denotes the K -dimensional identity matrix. As in the main text let $p_n(v) = \partial \theta_n(v, \widehat{\beta}_n) / \partial \beta_n$, which are simply the series terms in the case of series estimation.

Algorithm 1 (Implementation for Parametric and Series Estimation). (1) Set $\tilde{\gamma}_n \equiv 1 - .1 / \log n$. Simulate a large number R of draws denoted Z_1, \dots, Z_R from the K -variate standard normal distribution $\mathcal{N}(0, I_K)$. (2) Compute $\widehat{\Omega}_n$, a consistent estimator for the large sample variance of $\sqrt{n}(\widehat{\beta}_n - \beta_n)$. (3) For each $v \in \mathcal{V}$, compute $\widehat{g}(v) = p_n(v)' \widehat{\Omega}_n^{1/2}$ and set $s_n(v) = \|\widehat{g}(v)\| / \sqrt{n}$. (4) Compute

$$k_{n,\mathcal{V}}(\tilde{\gamma}_n) = \tilde{\gamma}_n - \text{quantile of } \left\{ \sup_{v \in \mathcal{V}} (\widehat{g}(v)' Z_r / \|\widehat{g}(v)\|), r = 1, \dots, R \right\}, \text{ and}$$

$$\widehat{V}_n = \{v \in \mathcal{V} : \widehat{\theta}_n(v) \leq \min_{v \in \mathcal{V}} \left(\widehat{\theta}_n(v) + k_{n,\mathcal{V}}(\tilde{\gamma}_n) s_n(v) \right) + 2k_{n,\mathcal{V}}(\tilde{\gamma}_n) s_n(v)\},$$

(5) Compute

$$k_{n,\widehat{V}_n}(p) = p - \text{quantile of } \left\{ \sup_{v \in \widehat{V}_n} (\widehat{g}(v)' Z_r / \|\widehat{g}(v)\|), r = 1, \dots, R \right\}, \text{ and set}$$

¹⁹For example if we have $\theta_n^l(z) \leq \theta_n^* \leq \theta_n^u(z)$ for all $z \in \mathcal{Z}$, then we can equivalently write $\min_{z \in \mathcal{Z}} \min_{j=1,2} g_n(\theta_n^*, z, j) \geq 0$, where $g_n(\theta_n^*, z, 1) = \theta_n^u(z) - \theta_n^*$ and $g_n(\theta_n^*, z, 2) = \theta_n^* - \theta_n^l(z)$. Then we can apply our method through use of the auxiliary function $g_n(\theta_n, z, j)$, in similar fashion as in Example C with multiple conditional moment inequalities.

²⁰In an earlier version of this paper, Chernozhukov, Lee, and Rosen (2009), we provided a different method for inference on a parameter with both lower and upper bounding functions, which can also be used for valid inference on θ^* .

$$\widehat{\theta}_{n0}(p) = \inf_{v \in \mathcal{V}} \left[\widehat{\theta}_n(v) + k_{n, \widehat{V}_n}(p) \|\widehat{g}(v)\| / \sqrt{n} \right].$$

An important special case of the parametric setup is that where the support of v is finite, as in Example 1 of Section 4.1, so that $\mathcal{V} = \{1, \dots, J\}$. In this case the algorithm applies with $\theta_n(v, \beta_n) = \sum_{j=1}^J 1[v = j] \beta_{nj}$, i.e. where for each j , $\theta_n(j, \beta_n) = \beta_{nj}$ and $\widehat{g}(v) = (1[v = 1], \dots, 1[v = J]) \cdot \widehat{\Omega}_n^{1/2}$. Note that this covers the case where the bounding function is a conditional mean or quantile with discrete conditioning variable, such as conditional mean estimation with discrete regressors, in which case $\beta_{nj} = E[Y|V = j]$ can be estimated by a sample mean.

The above algorithm applies when the bounding function is separable in the parameter of interest. When the bounding function is non-separable in this parameter, say μ where $\theta(v) := \theta(\mu, v)$, it can be used to test the hypothesis that any given μ is in the identified set as described in Example C in Section 2. That is, for any fixed μ and any chosen $\alpha \in [1/2, 1)$ it can be used to produce a critical value $\widehat{\theta}_{1-\alpha}$ such that for the true parameter value μ ,

$$P_n \left\{ \inf_{v \in \mathcal{V}} \theta(\mu, v) \geq \widehat{\theta}_{1-\alpha} \right\} \leq \alpha + o(1).$$

A confidence set for μ can then be formed by inverting this test. This is done by first carrying out step (1), computing all components of $\widehat{\Omega}_n$ in step (2) that are not dependent upon μ , such as \widehat{Q} and \widehat{Q}_n in the context of the conditional moment inequality examples 3 and 6, and then performing the rest of step (2) and steps (3)-(5) at every μ in some set of points approximating the parameter space. Note that objective function approaches to inference with set identification construct confidence sets through the inversion of tests in both non-separable and separable cases. Similarly to our procedure, in the non-separable case this requires computing a test statistic and critical value at each of a large grid of points approximating the parameter space. In the separable case our approach produces a critical value that is not parameter-dependent, so that the steps above need only be carried out once to produce the desired confidence set.

Remark 5. In practice one needs to approximate \mathcal{V} over a dense set of grid points, say $\{v_1, \dots, v_M\}$. The asymptotic linear representations of Condition P or Condition NS.1(a) can then be used to approximate the asymptotic distribution of

$$\{\theta_n(v_1) - \widehat{\theta}_n(v_1), \dots, \theta_n(v_M) - \widehat{\theta}_n(v_M)\}$$

directly, rather than iterating over each value of v , which speeds up implementation considerably. For instance, to do this when the bounding function is a conditional mean function as in Example 5, let \mathbf{P}_n denote the $(n \times K)$ matrix whose i th row is $p_n(V_i)'$. Let $\mathbf{\Psi}_n$ denote

the $(M \times K)$ matrix whose m th row is $p_n(v_m)'$, and define

$$\hat{\Sigma}_n = \Psi_n(\mathbf{P}'_n \mathbf{P}_n)^{-1} \mathbf{P}'_n \text{diag}(\hat{U}_n^2) \mathbf{P}_n (\mathbf{P}'_n \mathbf{P}_n)^{-1} \Psi'_n,$$

where $\text{diag}(\hat{U}_n^2)$ is the n -dimensional diagonal matrix whose i th component is

$$\hat{U}_i^2 = \left[Y_i - \hat{\theta}_n(V_i) \right]^2.$$

Then one can simulate $N(0, \hat{\Sigma}_n)$ random draws to approximate the joint distribution of $\{\theta_n(v_1) - \hat{\theta}_n(v_1), \dots, \theta_n(v_M) - \hat{\theta}_n(v_M)\}$.

Remark 6. In the case of series estimation, if desired one can bypass simulation of the stochastic process by instead employing the analytical critical value in step 4, $k_{n,\mathcal{V}}(p) = a_n(\mathbf{V}) - 2 \log(1 - p)/a_n(\mathbf{V})$ from Remark 2 in Section 4.2. This is convenient because it does not involve simulation, though it requires computation of $a_n(\hat{V}_n) = 2\sqrt{\log\{\ell_n(1 + \ell_n L_n \text{diam}(\hat{V}_n))^d\}}$. Moreover, it could be too conservative in some applications. Thus, we recommend using simulation, unless the computational cost is too high.

6.2. Kernel-Type Estimators. In this section we describe the steps for implementation of kernel-type estimators.

Algorithm 2 (Implementation for Kernel Case). (1) Set $\gamma_n \equiv 1 - .1/\log n$. Simulate $R \times n$ independent draws from $N(0, 1)$, denoted by $\{\eta_{ir} : i = 1, \dots, n, r = 1, \dots, R\}$, where n is the sample size and R is the number of simulation repetitions. (2) For each $v \in \mathcal{V}$ and $r = 1, \dots, R$, compute $\mathbb{G}_n^o(\hat{g}_v; r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_{ir} \hat{g}_v(U_i, Z_i)$, where $\hat{g}_v(U_i, Z_i)$ is defined in Section 4.3, that is

$$\hat{g}_v(U_i, Z_i) = \frac{e'_j \hat{U}_i}{(h_n^d)^{1/2} \hat{f}_n(z)} \mathbf{K}\left(\frac{z - Z_i}{h_n}\right).$$

Let $s_n^2(v) = \mathbb{E}_n[\hat{g}_v^2]/(nh_n^d)$ and $\mathbb{E}_n[\hat{g}_v^2] = n^{-1} \sum_{i=1}^n \hat{g}_v^2(U_i, Z_i)$. Here, \hat{U}_i is the kernel-type regression residual and $\hat{f}_n(z)$ is the kernel density estimator of density of Z_i . (3) Compute $k_{n,\mathcal{V}}(\gamma_n) = \gamma_n - \text{quantile of } \left\{ \sup_{v \in \mathcal{V}} \mathbb{G}_n^o(\hat{g}_v; r) / \sqrt{\mathbb{E}_n[\hat{g}_v^2]}, r = 1, \dots, R \right\}$, and $\hat{V}_n = \{v \in \mathcal{V} : \hat{\theta}_n(v) \leq \min_{v \in \mathcal{V}} \left(\hat{\theta}_n(v) + k_{n,\mathcal{V}}(\gamma_n) s_n(v) \right) + 2k_{n,\mathcal{V}}(\gamma_n) s_n(v)\}$, (4) Compute $k_{n,\hat{V}_n}(p) = p - \text{quantile of } \left\{ \sup_{v \in \hat{V}_n} \mathbb{G}_n^o(\hat{g}_v; r) / \sqrt{\mathbb{E}_n[\hat{g}_v^2]}, r = 1, \dots, R \right\}$, and set $\hat{\theta}_{n0}(p) = \inf_{v \in \mathcal{V}} [\hat{\theta}(v) + k_{n,\hat{V}_n}(p) s_n(v)]$.

Remark 7. (1) The researcher also has the option of employing an analytical approximation in place of simulation if desired. This can be done by using $k_{n,\mathcal{V}}(p) = a_n(\mathbf{V}) - 2 \log(1 -$

$p)/a_n(\mathbf{V})$ from Remark 3, but requires computation of

$$a_n(\widehat{V}_n) = 2\sqrt{\log\{\ell_n(1 + \ell_n(1 + h_n^{-1})\text{diam}(\widehat{V}_n)^d)\}}.$$

This approximation could be too conservative in some applications, and thus we recommend using simulation, unless the computational cost is too high. (2) In the case where the bounding function is non-separable in a parameter of interest, a confidence interval for this parameter can be constructed as described in Section 6.1, where step (1) is carried out once and steps (2)-(4) are executed iteratively on a set of parameter values approximating the parameter space. However, the bandwidth, $\hat{f}_n(z)$, and $\mathbf{K}\left(\frac{z-Z_i}{h_n}\right)$, each Z_i , do not vary across iterations and thus only need to be computed once. ■

7. MONTE CARLO EXPERIMENTS

In this section we present results of Monte Carlo experiments to illustrate the finite-sample performance of our method. We consider a Monte Carlo design with bounding function

$$\theta(v) := L\phi(v), \tag{7.1}$$

where L is a constant and $\phi(\cdot)$ is the standard normal density function. Throughout the Monte Carlo experiments, the parameter of interest is $\theta_0 = \sup_{v \in \mathcal{V}} \theta(v)$.²¹

7.1. Data-Generating Processes. We consider four Monte Carlo designs for the sake of illustration.²² In the first Monte Carlo design, labeled DGP1, the bounding function is completely flat so that $V_0 = \mathcal{V}$. In the second design, DGP2, the bounding function is non-flat, but smooth in a neighborhood of its maximizer, which is unique so that V_0 is singleton. In DGP3 and DGP4, the bounding function is also non-flat and smooth in a neighborhood of its (unique) maximizer, though relatively peaked. Illustrations of these bounding functions are provided in Figures 1 and 2 of our on-line supplement. In practice the shape of the bounding function is unknown, and the inference and estimation methods we consider do not make use of this information. As we describe in more detail below, we evaluate the finite sample performance of our approach in terms of coverage probability for the true point θ_0 and coverage for a false parameter value θ that is close to but below θ_0 . We compare the performance of our approach to that of the Cramer Von-Mises statistic

²¹Previous sections focused on $\theta_0 = \inf_{v \in \mathcal{V}} \theta(v)$ rather than $\theta_0 = \sup_{v \in \mathcal{V}} \theta(v)$. This is not a substantive difference as for any function $\theta(\cdot)$, $\sup_{v \in \mathcal{V}} \theta(v) = -\inf_{v \in \mathcal{V}} (-\theta(v))$.

²²We consider some additional Monte Carlo designs in Section L of the on-line supplement.

proposed by AS. DGP1 and DGP2 in particular serve to effectively illustrate the relative advantages of both procedures as we describe below. Neither approach dominates.

For all DGPs we generated 1000 independent samples from the following model:

$$V_i \sim \text{Unif}[-2, 2], U_i = \min\{\max\{-3, \sigma\tilde{U}_i\}, 3\}, \text{ and } Y_i = L\phi(V_i) + U_i,$$

where $\tilde{U}_i \sim N(0, 1)$ and L and σ are constants. We set these constants in the following way:

$$\text{DGP1: } L = 0 \text{ and } \sigma = 0.1; \quad \text{DGP2: } L = 1 \text{ and } \sigma = 0.1;$$

$$\text{DGP3: } L = 5 \text{ and } \sigma = 0.1; \quad \text{DGP4: } L = 5 \text{ and } \sigma = 0.01.$$

We considered sample sizes $n = 500$ and $n = 1000$, and we implemented both series and kernel-type estimators to estimate the bounding function $\theta(v)$ in (7.1). We set \mathcal{V} to be an interval between the .05 and .95 sample quantiles of V_i 's in order to avoid undue influence of outliers at the boundary of the support of V_i . For both types of estimators, we computed critical values via simulation as described in Section 6, and we implemented our method with both the conservative but simple, non-stochastic choice $\hat{V} = \mathcal{V}$ and the set estimate $\hat{V} = \hat{V}_n$ described in Section 3.2.

7.2. Series Estimation. For basis functions we use polynomials and cubic B-splines with knots equally spaced over the sample quantiles of V_i . The number $K = K_n$ of approximating functions was obtained by the following simple rule-of-thumb:

$$K = \underline{\hat{K}}, \quad \hat{K} := \hat{K}_{cv} \times n^{-1/5} \times n^{2/7}, \quad (7.2)$$

where \underline{a} is defined as the largest integer that is smaller than or equal to a , and \hat{K}_{cv} is the minimizer of the leave-one-out least squares cross validation score. If $\theta(v)$ 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 (7.2) ensures that the bias is asymptotically negligible from under-smoothing.²³

²³For B-splines the optimal \hat{K}_{cv} was first selected from the first $5 \times n^{1/5}$ values starting from 5, with $n^{1/5}$ rounded up to the nearest integer. If the upper bound was selected, the cross validation (CV) score of \hat{K}_{cv} was compared to that of $\hat{K}_{cv} + 1$ iteratively, such that \hat{K}_{cv} was increased until further increments resulted in no improvement. This allows $\hat{K}_{cv} \propto n^{1/5}$ and provides a crude check against the upper bound binding in the CV search, though in these DGPs results differed little from those searching over $\{5, 6, 7, 8, 9\}$, reported in Chernozhukov, Lee, and Rosen (2009). For polynomials the CV search was limited to the set $\{3, 4, 5, 6\}$ due to multicollinearity issues that arose when too many terms were used.

7.3. Kernel-Type Estimation. ²⁴ We use 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 rule of thumb bandwidth:

$$h = \hat{h}_{ROT} \times \hat{s}_v \times n^{1/5} \times n^{-2/7}, \quad (7.3)$$

where \hat{s}_v is the square root of the sample variance of the V_i , and \hat{h}_{ROT} is the rule-of-thumb bandwidth for estimation of $\theta(v)$ with studentized V , as prescribed in Section 4.2 of Fan and Gijbels (1996). The exact form of \hat{h}_{ROT} is

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

where \tilde{V}_i 's are studentized V_i 's, $\tilde{\theta}^{(2)}(\cdot)$ is the second-order derivative of the global quartic parametric fit of $\theta(v)$ 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 .10 and .90 sample quantiles of \tilde{V}_i . Again, the factor $n^{1/5} \times n^{-2/7}$ is multiplied in (7.3) to ensure that the bias is asymptotically negligible due to under-smoothing.

7.4. Simulation Results. To evaluate the relative performance of our inference method, we also implemented one of the inference methods proposed by AS, specifically their Cramér-von Mises-type (CvM) statistic with both plug-in asymptotic (PA/Asy) and asymptotic generalized moment selection (GMS/Asy) critical values. For instrument functions we used countable hypercubes and the S -function of AS Section 3.2.²⁵ We set the weight function and tuning parameters for the CvM statistic exactly as in AS (see AS Section 9). These values performed well in their simulations, but our Monte Carlo design differs from theirs, and alternative choices of tuning parameters could perform more or less favorably in our design. We did not examine sensitivity to the choice of tuning parameters for the CvM statistic.

The coverage probability (CP) of confidence intervals with nominal level 95% is evaluated for the true lower bound θ_0 , and false coverage probability (FCP) is reported at $\theta = \theta_0 - 0.02$. There were 1,000 replications for each experiment. Tables 1, 2, and 3 summarize the results. CLR and AS refer to our inference method and that of AS, respectively.

²⁴Appendices G and H of the on-line supplement provide strong approximation results and proofs for kernel-type estimators, including the local linear estimator used here.

²⁵All three S -functions in AS Section 3.2 are equivalent in our design, since there is a single conditional moment inequality.

We first consider the performance of our method for DGP1. In terms of coverage for θ_0 both series estimators and the local linear estimator perform reasonably well, with the series estimators performing best. The polynomial series and local linear estimators perform somewhat better in terms of false coverage probabilities, which decrease with the sample size for all estimators. The argmax set V_0 is the entire set \mathcal{V} , and our set estimator \hat{V}_n detects this. Turning to DGP2 we see that coverage for θ_0 is in all cases roughly .98 to .99. There is non-trivial power against the false parameter θ in all cases, with the series estimators giving the lowest false coverage probabilities. For DGP3 the bounding function is relatively peaked compared to the smooth but non-flat bounding function of DGP2. Consequently the average endpoints of the preliminary set estimator \hat{V}_n become more concentrated around 0, the maximizer of the bounding function. Performance in terms of coverage probabilities improves in nearly all cases, with the series estimators performing significantly better when $n = 1000$ and \hat{V}_n is used. With DGP4 the bounding function remains as in DGP3, but now with the variance of Y_i decreased by a factor of 100, which has roughly the same effect on the precision of estimators as would increasing sample size by a factor of 100. The result is that the bounding function is more accurately estimated at every point. Moreover, the set estimator \hat{V}_n is now a much smaller interval around 0. Coverage frequencies for θ_0 do not change much relative to DGP3, but false coverage probabilities drop to 0. Note that in DGPs 2-4, our method performs better when V_0 is estimated in that it makes the coverage probability more accurate and the false coverage probability smaller. DGPs 3-4 serve to illustrate the convergence of our set estimator \hat{V}_n when the bounding function is peaked and precisely estimated, respectively.

In Table 2 we report the results of using the CvM statistic of AS to perform inference. For DGP1 with a flat bounding function the CvM statistic with both the PA/Asy and GMS/Asy performs well. Coverage frequencies for θ_0 were close to the nominal level, closer than our method using polynomial series or local linear regression. The CvM statistic has a lower false coverage probability than the CLR confidence intervals in this case, although at a sample size of 1000 the difference is not large. For DGP2 the bounding function is non-flat but smooth in a neighborhood of V_0 and the situation is much different. For both PA/Asy and GMS/Asy critical values with the CvM statistic, coverage frequencies for θ_0 were 1. Our confidence intervals also over-covered in this case, with coverage frequencies of roughly .98 to .99. Moreover, the CvM statistic has low power against the false parameter θ , with coverage 1 with PA/Asy and coverage .977 and .933 with sample size 500 and 1000, respectively using GMS/Asy critical values. For DGP3 and DGP4 both critical values for the CvM statistic gave coverage for θ_0 and the false parameter θ equal to one. Thus under

DGPs 2,3, and 4 our confidence intervals perform better by both measures. However, overall neither approach dominates.

Thus, in our Monte Carlo experiments the CvM statistic exhibits better power when the bounding function is flat, while our confidence intervals exhibit better power when the bounding function is non-flat. AS establish that the CvM statistic has power against some $n^{-1/2}$ local alternatives under conditions that are satisfied under DGP1, but that do not hold when the bounding function has a unique minimum.²⁶ We have established local asymptotic power for nonparametric estimators of polynomial order less distant than $n^{-1/2}$ that apply whether the bounding function is flat or non-flat. Our Monte Carlo results accord with these findings.²⁷ In the on-line supplement, we present further supporting Monte Carlo evidence and local asymptotic power analysis to show why our method performs better than the AS method in non-flat cases.

In Table 4 we report computation times for our Monte Carlo experiments.²⁸ The fastest performance in terms of total simulation time was achieved with the CvM statistic of AS, which took roughly 9 minutes to execute a total of 16,000 replications. Simulations using our approach with B-spline series, polynomial series, and local linear polynomials took roughly 58, 19, and 84 minutes, respectively. Based on these times the table shows for each statistic the average time for a single test, and the relative performance of each method to that obtained using the CvM statistic.²⁹

In practice one will not perform Monte Carlo experiments but will rather be interested in computing a single confidence region for the parameter of interest. When the bounding function is separable our approach offers the advantage that the critical value does not vary with the parameter value being tested. As a result, we can compute a confidence region in the same amount of time it takes to compute a single test. On the other hand, to construct

²⁶Specifically Assumptions LA3 and LA3' of AS Theorem 4 do not hold when the sequence of models has a fixed bounding function with a unique minimum. As they discuss after the statement of Assumptions LA3 and LA3', in such cases GMS and plug-in asymptotic tests have trivial power against $n^{-1/2}$ local alternatives.

²⁷We did not do CP-correction in our reported results. Our conclusion will remain valid even with CP-correction as in AS, since our method performs better in DGP2-DGP4 where we have over-coverage.

²⁸These computation times were obtained on a 2011 iMac desktop with a 2.7 GHz processor and 8GB RAM using our implementation. Generally speaking, performance time for both methods will depend on the efficiency of one's code, and more efficient implementation times for both methods may be possible.

²⁹The difference in computation time between polynomial and B-spline series implementation was almost entirely due to the simpler cross-validation search from polynomials. With a search over only 4 values of \hat{K}_{cv} , simulation time for B-splines took roughly 20 minutes. Cross-validation is also in part accountable for slower performance relative to AS. We followed Section 9 of AS in choosing tuning parameters for the CvM statistic, which does not involve cross-validation. Using B-splines with a deterministic bandwidth resulted in a computation time of 12 minutes, roughly 1.4 times the total computation time for the CvM statistic. Nonetheless, we prefer cross-validation for our method in practice.

a confidence region based on the CvM statistic, one must compute the statistic and its associated critical value at a large number of points in the parameter space, where the number of points required will depend on the size of the parameter space and the degree of precision desired. If however the bounding function is not separable in the parameter of interest, then both approaches use parameter-dependent critical values.

8. 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 bounding functions complicates inference. We overcame these difficulties by applying a precision-correction to the estimated bounding 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.

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 estimators 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. DEFINITION OF STRONG APPROXIMATION

The following definitions are used extensively.

Definition 4 (Strong approximation). *Suppose that for each n there are random variables Z_n and Z'_n defined on a probability space (A, \mathcal{A}, P_n) and taking values in the separable metric space (S, d_S) . We say that $Z_n =_d Z'_n + o_{P_n}(\delta_n)$, for $\delta_n \rightarrow 0$, if there are identically distributed copies of Z_n and Z'_n , denoted \bar{Z}_n and \bar{Z}'_n , defined on (A, \mathcal{A}, P_n) (suitably enriched*

if needed), such that

$$d_S(\bar{Z}_n, \bar{Z}'_n) = o_{P_n}(\delta_n).$$

Note that copies \bar{Z}_n and \bar{Z}'_n can be always defined on (A, \mathcal{A}, P_n) by suitably enriching this space by taking product probability spaces. It turns out that for the Polish spaces, this definition implies the following stronger, and much more convenient, form.

Lemma 9 (A Convenient Implication for Polish Spaces via Dudley and Philipp). *Suppose that (S, d_S) is Polish, i.e. complete, separable metric space, and (A, \mathcal{A}, P_n) has been suitably enriched. Suppose that Definition 4 holds, then there is also an identical copy Z_n^* of Z'_n such that $Z_n = Z_n^* + o_{P_n}(\delta_n)$, that is,*

$$d_S(Z_n, Z_n^*) = o_{P_n}(\delta_n)$$

Proof. We start with the original probability space (A', \mathcal{A}', P'_n) that can carry Z_n and (\bar{Z}_n, \bar{Z}'_n) . In order to apply Lemma 2.11 of Dudley and Philipp (1983), we need to carry a standard uniform random variable $U \sim U(0, 1)$ that is independent of Z_n . To guarantee this we can always consider $U \sim U(0, 1)$ on the standard space $([0, 1], \mathcal{F}, \lambda)$, where \mathcal{F} is the Borel sigma algebra on $[0, 1]$ and λ is the usual Lebesgue measure, and then enrich the original space (A', \mathcal{A}', P'_n) by creating formally a new space (A, \mathcal{A}, P_n) as the product of (A', \mathcal{A}', P'_n) and $([0, 1], \mathcal{F}, \lambda)$. Then using Polishness of (S, d_S) , given the joint law of (\bar{Z}_n, \bar{Z}'_n) , we can apply Lemma 2.11 of Dudley and Philipp (1983) to construct Z_n^* such that (Z_n, Z_n^*) has the same law as (\bar{Z}_n, \bar{Z}'_n) , so that $d_S(\bar{Z}_n, \bar{Z}'_n) = o_{P_n}(\delta_n)$ implies $d_S(Z_n, Z_n^*) = o_{P_n}(\delta_n)$. ■

Since in all of our cases the relevant metric spaces are either the space of continuous functions defined on a compact set equipped with the uniform metric or finite-dimensional Euclidian spaces, which are all Polish spaces, we can use Lemma 9 throughout the paper. Using this implication of strong approximation makes our proofs slightly simpler.

APPENDIX B. PROOFS FOR SECTION 3

B.1. Some Useful Facts and Lemmas. A useful result in our case is the anti-concentration inequality derived in Chernozhukov and Kato (2011).

Lemma 10 (Anti-Concentration Inequality, Chernozhukov and Kato (2011)). *Let $X = (X_t)_{t \in T}$ be a separable Gaussian process indexed by a semimetric space T such that $E_P[X_t] = 0$ and $E_P[X_t^2] = 1$ for all $t \in T$. Then*

$$\sup_{x \in \mathbb{R}} P \left(\left| \sup_{t \in T} X_t - x \right| \leq \epsilon \right) \leq C\epsilon \left(E_P \left[\sup_{t \in T} X_t \right] \vee 1 \right), \quad \forall \epsilon > 0, \quad (\text{B.1})$$

where C is an absolute constant.

An immediate consequence of this lemma is the following result:

Corollary 2 (Anti-concentration for $\sup_{v \in V_n} Z_n^*(v)$). *Let V_n be any sequence of compact non-empty subsets in \mathcal{V} . Then under condition C.2-C.3, we have that for $\delta_n \rightarrow 0$ such that $\delta_n = o(1/\bar{a}_n)$*

$$\sup_{x \in \mathbb{R}} \mathbb{P}_n \left(\left| \sup_{v \in V_n} Z_n^*(v) - x \right| \leq \delta_n \right) = o(1).$$

Proof. Continuity in Condition C.2 implies separability of Z_n^* . Condition C.3 implies that $E_{\mathbb{P}_n}[\sup_{v \in \mathcal{V}_n} Z_n^*(v)] \leq E_{\mathbb{P}_n}[\sup_{v \in \mathcal{V}} Z_n^*(v)] \leq K\bar{a}_n$ for some constant K that depends only on η , so that

$$\sup_{x \in \mathbb{R}} \mathbb{P}_n \left(\left| \sup_{v \in \mathcal{V}} Z_n^*(v) - x \right| \leq \delta_n \right) \leq C\delta_n[K\bar{a}_n \vee 1] = o(1).$$

■

Lemma 11 (Closeness in Conditional Probability Implies Closeness of Conditional Quantiles Unconditionally). *Let X_n and Y_n be random variables and \mathcal{D}_n be a random vector. Let $F_{X_n}(x \mid \mathcal{D}_n)$ and $F_{Y_n}(y \mid \mathcal{D}_n)$ denote the conditional distribution functions, and $F_{X_n}^{-1}(p \mid \mathcal{D}_n)$ and $F_{Y_n}^{-1}(p \mid \mathcal{D}_n)$ denote the corresponding conditional quantile functions. If $\mathbb{P}_n(|X_n - Y_n| > \xi_n \mid \mathcal{D}_n) = o_{\mathbb{P}_n}(\tau_n)$ for some sequence $\tau_n \searrow 0$, then with unconditional probability \mathbb{P}_n converging to one, for some $\varepsilon_n = o(\tau_n)$,*

$$F_{X_n}^{-1}(p \mid \mathcal{D}_n) \leq F_{Y_n}^{-1}(p - \varepsilon_n \mid \mathcal{D}_n) + \xi_n \text{ and } F_{Y_n}^{-1}(p \mid \mathcal{D}_n) \leq F_{X_n}^{-1}(p - \varepsilon_n \mid \mathcal{D}_n) + \xi_n, \forall p \in (\varepsilon_n, 1 - \varepsilon_n).$$

Proof. We have that for some $\varepsilon_n = o(\tau_n)$, $\mathbb{P}_n[\mathbb{P}_n\{|X_n - Y_n| > \xi_n \mid \mathcal{D}_n\} \leq \varepsilon_n] \rightarrow 1$, that is, there is a set Ω_n such that $\mathbb{P}_n(\Omega_n) \rightarrow 1$ such that $\mathbb{P}_n\{|X_n - Y_n| > \xi_n \mid \mathcal{D}_n\} \leq \varepsilon_n$ for all $\mathcal{D}_n \in \Omega_n$. So, for all $\mathcal{D}_n \in \Omega_n$

$$F_{X_n}(x \mid \mathcal{D}_n) + \varepsilon_n \geq F_{Y_n + \xi_n}(x \mid \mathcal{D}_n) \text{ and } F_{Y_n}(x \mid \mathcal{D}_n) + \varepsilon_n \geq F_{X_n + \xi_n}(x \mid \mathcal{D}_n), \forall x \in \mathbb{R},$$

which implies the inequality stated in the lemma, by definition of the conditional quantile function and equivariance of quantiles to location shifts. ■

B.2. Proof of Lemma 1. (Concentration of Inference on V_n .) Step 1. Letting

$$\begin{aligned} A_n &:= \sup_{v \in V_n} Z_n(v), \quad B_n := \sup_{v \in \mathcal{V}} Z_n(v), \quad R_n := \left(\sup_{v \in \mathcal{V}} |Z_n(v)| + \kappa_n \right) \sup_{v \in \mathcal{V}} \left| \frac{\sigma_n(v)}{s_n(v)} - 1 \right|, \\ A_n^* &:= \sup_{v \in V_n} Z_n^*(v), \quad B_n^* := \sup_{v \in \mathcal{V}} Z_n^*(v), \quad R_n^* := \left(\sup_{v \in \mathcal{V}} |Z_n^*(v)| + \kappa_n \right) \sup_{v \in \mathcal{V}} \left| \frac{\sigma_n(v)}{s_n(v)} - 1 \right|, \end{aligned}$$

we obtain

$$\begin{aligned}
\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \widehat{\theta}_n(v)}{s_n(v)} &= \sup_{v \in \mathcal{V}} \left\{ \frac{\theta_{n0} - \theta_n(v)}{s_n(v)} + Z_n(v) \frac{\sigma_n(v)}{s_n(v)} \right\} \\
&= \sup_{v \in V_n} \left\{ \frac{(\theta_{n0} - \theta_n(v))}{s_n(v)} + Z_n(v) \frac{\sigma_n(v)}{s_n(v)} \right\} \vee \sup_{v \notin V_n} \left\{ \frac{(\theta_{n0} - \theta_n(v))}{s_n(v)} + Z_n(v) \frac{\sigma_n(v)}{s_n(v)} \right\} \\
&\leq_{(1)} \sup_{v \in V_n} \left\{ Z_n(v) \frac{\sigma_n(v)}{s_n(v)} \right\} \vee \sup_{v \notin V_n} \left\{ \frac{-\kappa_n \sigma_n(v)}{s_n(v)} + Z_n(v) \frac{\sigma_n(v)}{s_n(v)} \right\} \\
&\leq A_n \vee (B_n - \kappa_n) + 2R_n \leq_{(2)} A_n^* \vee (B_n^* - \kappa_n) + 2R_n^* + o_{P_n}(\delta_n),
\end{aligned}$$

where in (1) we used that $\theta_n(v) \geq \theta_{n0}$ and $\theta_{n0} - \theta_n(v) \leq -\kappa_n \sigma_n(v)$ outside V_n , and in (2) we used C.2. Next, since we assumed in the statement of the lemma that $\kappa_n \lesssim \bar{a}_n + \ell \ell_n$, and by C.4: $R_n^* = O_{P_n}(\bar{a}_n + \bar{a}_n + \ell \ell_n) o_{P_n}(\delta_n / (\bar{a}_n + \ell \ell_n)) = o_{P_n}(\delta_n)$. Therefore, there is a deterministic term $o(\delta_n)$ such that $P_n(2R_n^* + o_{P_n}(\delta_n) > o(\delta_n)) = o(1)$.³⁰

Hence uniformly in $x \in [0, \infty)$

$$\begin{aligned}
P_n \left(\sup_{v \in \mathcal{V}} \frac{(\theta_{n0} - \widehat{\theta}_n(v))}{s_n(v)} > x \right) &\leq P_n(A_n^* + o(\delta_n) > x) + P_n(B_n^* - \kappa_n + o(\delta_n) > 0) + o(1) \\
&\leq P_n(A_n^* > x) + P_n(B_n^* - \kappa_n > 0) + o(1) \leq P_n(A_n^* > x) + (1 - \gamma'_n) + o(1),
\end{aligned}$$

where the last two inequalities follow by Corollary 2 and by $\kappa_n = Q_{\gamma'_n}(B_n^*)$.

Step 2. To complete the proof, we must show that there is $\gamma'_n \nearrow 1$ such that $\kappa_n \lesssim \bar{a}_n + \ell \ell_n$. Let $1 - \gamma'_n \searrow 0$ such that $1 - \gamma'_n \geq C/\ell_n$. It suffices to show that

$$\kappa_n \leq \left(\bar{a}_n + \frac{c(\gamma'_n)}{\bar{a}_n} \right) \leq \left(\bar{a}_n + \frac{\eta \ell \ell_n + \eta \log C^{-1}}{\bar{a}_n} \right) \lesssim \bar{a}_n + \ell \ell_n, \quad (\text{B.2})$$

where $c(\gamma'_n) = Q_{\gamma'_n}(\mathcal{E})$. To show the first inequality in (B.2) note

$$P_n \left(\sup_{v \in \mathcal{V}} Z_n^*(v) \leq (\bar{a}_n + c(\gamma'_n)/\bar{a}_n) \right) =_{(1)} P_n(\mathcal{E}_n(\mathcal{V}) \leq c(\gamma'_n)) \geq_{(2)} P_n(\mathcal{E} \leq c(\gamma'_n)) = \gamma'_n,$$

where (1) holds by definition of $\mathcal{E}_n(\mathcal{V})$ and (2) by C.3. To show the second inequality in (B.2) note that by C.3 $P(\mathcal{E} > t) \leq \exp(-t\eta^{-1})$, for some constant $\eta > 0$, so that $c(\gamma'_n) \leq -\eta \log(1 - \gamma'_n) \leq \eta \ell \ell_n + \eta \log C^{-1}$. \blacksquare

³⁰Throughout the paper we use the elementary fact: If $X_n = o_{P_n}(\Delta_n)$, for some $\Delta_n \searrow 0$, then there is $o(\Delta_n)$ term such that $P_n\{|X_n| > o(\Delta_n)\} \rightarrow 0$.

B.3. Proof of Theorem 1 (Analytical Construction). Part 1.(Level) Observe that for sufficiently large n

$$\begin{aligned}
P_n \left(\theta_{n0} \leq \widehat{\theta}_{n0}(p) \right) &= P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \widehat{\theta}_n(v)}{s_n(v)} \leq k_{n, \widehat{V}_n}(p) \right) \\
&\geq_{(1)} P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \widehat{\theta}_n(v)}{s_n(v)} \leq k_{n, V_n}(p) \right) - P_n \left(V_n \not\subseteq \widehat{V}_n \right) \\
&\geq_{(2)} P_n \left(\sup_{v \in V_n} Z_n^*(v) \leq k_{n, V_n}(p) \right) - o(1) \\
&= P_n \left(\mathcal{E}_n(V_n) \leq c(p) - o(1) \right) - o(1) \geq_{(3)} P_n \left(\mathcal{E} \leq c(p) - o(1) \right) - o(1) =_{(4)} p - o(1),
\end{aligned}$$

where (1) follows by monotonicity of $V \mapsto k_{n, V}(p) = a_n(V) + c(p)/a_n(V)$ for large n holding by C.3, (2) by Lemma 1, by $P_n(V_n \not\subseteq \widehat{V}_n) = o(1)$ holding by Lemma 2, and also by the fact that the critical value $k_{n, V_n}(p) \geq 0$ is non-stochastic, and (3) and (4) by the existence of majorizing rv \mathcal{E} with a continuous distribution function (see C.3).

Part 2.(Estimation Risk) We have that under P_n

$$\begin{aligned}
\left| \widehat{\theta}_{n0}(p) - \theta_{n0} \right| &= \left| \inf_{v \in \mathcal{V}} \left[\widehat{\theta}_n(v) + k_{n, \widehat{V}_n}(p) s_n(v) \right] - \theta_{n0} \right| \\
&= \left| \sup_{v \in \mathcal{V}} \left(\left[\frac{\theta_{n0} - \widehat{\theta}_n(v)}{s_n(v)} + k_{n, \widehat{V}_n}(p) \right] \sigma_n(v) \frac{s_n(v)}{\sigma_n(v)} \right) \right| \\
&\leq_{(1)} \left(\left| \sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \widehat{\theta}_n(v)}{s_n(v)} \right| + k_{n, \widehat{V}_n}(p) \right) \bar{\sigma}_n \left(1 + o_{P_n} \left(\frac{\delta_n}{\bar{a}_n + \ell \ell_n} \right) \right) \\
&\leq_{(2)} \left(\left| \sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \widehat{\theta}_n(v)}{\sigma_n(v)} \right| + k_{n, \widehat{V}_n}(p) \right) \bar{\sigma}_n \left(1 + o_{P_n} \left(\frac{\delta_n}{\bar{a}_n + \ell \ell_n} \right) \right)^2 \\
&\leq_{(3)} \left(\sup_{v \in V_n} |Z_n^*(v)| + o_{P_n}(\delta_n) + k_{n, \widehat{V}_n}(p) \right) \bar{\sigma}_n \left(1 + o_{P_n} \left(\frac{\delta_n}{\bar{a}_n + \ell \ell_n} \right) \right)^2 \text{ wp } \rightarrow 1 \\
&\leq_{(4)} \left(\sup_{v \in V_n} |Z_n^*(v)| + o_{P_n}(\delta_n) + k_{n, \bar{V}_n}(p) \right) \bar{\sigma}_n \left(1 + o_{P_n} \left(\frac{\delta_n}{\bar{a}_n + \ell \ell_n} \right) \right)^2 \text{ wp } \rightarrow 1 \\
&\leq_{(5)} 3 \left| a_n(\bar{V}_n) + \frac{O_{P_n}(1)}{a_n(\bar{V}_n)} + o_{P_n}(\delta_n) \right| \bar{\sigma}_n \left(1 + o_{P_n} \left(\frac{\delta_n}{\bar{a}_n + \ell \ell_n} \right) \right)^2 \text{ wp } \rightarrow 1 \\
&\leq_{(6)} 4 \left| a_n(\bar{V}_n) + \frac{O_{P_n}(1)}{a_n(\bar{V}_n)} \right| \bar{\sigma}_n \text{ wp } \rightarrow 1,
\end{aligned}$$

where (1) holds by C.4 and the triangle inequality; (2) holds by C.4; (3) follows because $\text{wp} \rightarrow 1$, for some $o(\delta_n)$

$$\sup_{v \in V_0} Z_n^*(v) - o(\delta_n) \leq_{(a)} \sup_{v \in \hat{V}_0} Z_n(v) \leq_{(b)} \sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{\sigma_n(v)} \leq_{(c)} \left(\sup_{v \in V_n} Z_n^*(v) \right) \vee 0 + o(\delta_n),$$

where (a) is by C.2, (b) by definition of Z_n , while (c) by the proof of Lemma 1, so that $\text{wp} \rightarrow 1$

$$\left| \sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{\sigma_n(v)} \right| \leq \sup_{v \in V_n} |Z_n^*(v)| + o_{\mathbb{P}_n}(\delta_n);$$

(4) follows by Lemma 2 which implies $V_n \subseteq \hat{V}_n \subseteq \bar{V}_n$ $\text{wp} \rightarrow 1$, so that

$$k_{n, \hat{V}_n}(p) \leq k_{n, \bar{V}_n}(p) = a_n(\bar{V}_n) + \frac{c(p)}{a_n(\bar{V}_n)},$$

Condition C.3 gives (5). Inequality (6) follows because $a_n(\bar{V}_n) \geq 1$, $\bar{a}_n \geq 1$, and $\delta_n = o(1)$; this inequality is the claim that we needed to prove.

Part 3. We have that

$$\theta_{na} - \theta_{n0} \geq 4\bar{\sigma}_n \left(a_n(\bar{V}_n) + \frac{\mu_n}{a_n(\bar{V}_n)} \right) > \hat{\theta}_{n0}(p) - \theta_{n0} \text{ wp } \rightarrow 1,$$

with the last inequality occurring by Part 2 since $\mu_n \rightarrow_{\mathbb{P}_n} \infty$. ■

B.4. Proof of Theorem 2 (Simulation Construction). Part 1. (Level Consistency)

Let us compare critical values

$$k_{n, V_n}(p) = Q_p \left(\sup_{v \in V_n} Z_n^*(v) \mid \mathcal{D}_n \right) \text{ and } \kappa_{n, V_n}(p) = Q_p \left(\sup_{v \in V_n} \bar{Z}_n^*(v) \right).$$

The former is data-dependent while the latter is deterministic. Note that $k_{n, V_n}(p) \geq 0$ by C.2(b) for $p \geq 1/2$. By C.2 $\text{wp} \rightarrow 1$ for some deterministic term $o(\delta_n)$,

$$\mathbb{P}_n \left(\left| \sup_{v \in V_n} Z_n^*(v) - \sup_{v \in V_n} \bar{Z}_n^*(v) \right| > o(\delta_n) \mid \mathcal{D}_n \right) = o_{\mathbb{P}_n}(1),$$

which implies by Lemma 11 that for some $\varepsilon_n \searrow 0$, $\text{wp} \rightarrow 1$

$$k_{n, V_n}(p) \geq (\kappa_{n, V_n}(p - \varepsilon_n) - o(\delta_n))_+ \text{ for all } p \in [1/2, 1 - \varepsilon_n]. \quad (\text{B.3})$$

The result follows analogously to the proof in Part 1 of Theorem 1, namely:

$$\begin{aligned}
P_n \left(\theta_{n0} \leq \hat{\theta}_{n0}(p) \right) &= P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \leq k_{n, \hat{V}_n}(p) \right) \\
&\geq_{(1)} P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \leq k_{n, V_n}(p) \right) - o(1) \\
&\geq_{(2)} P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \leq (\kappa_{n, V_n}(p - \varepsilon_n) - o(\delta_n))_+ \right) - o(1) \\
&\geq_{(3)} P_n \left(\sup_{v \in V_n} Z_n^*(v) \leq (\kappa_{n, V_n}(p - \varepsilon_n) - o(\delta_n))_+ \right) - o(1) \\
&\geq P_n \left(\sup_{v \in V_n} Z_n^*(v) \leq \kappa_{n, V_n}(p - \varepsilon_n) - o(\delta_n) \right) - o(1) \geq_{(4)} p - \varepsilon_n - o(1) = p - o(1),
\end{aligned}$$

where (1) follows by monotonicity of $V \mapsto k_{n, V}(p)$ holding for sufficiently large n by C.3 and by $P_n(V_n \not\subseteq \hat{V}_n) = o(1)$ shown in Lemma 2, (2) holds by the comparison of quantiles in equation (B.3), (3) by Lemma 1. (4) holds by anti-concentration Corollary 2.

Parts 2 & 3.(Estimation Risk and Power) By Lemma 2 $\text{wp} \rightarrow 1$, $\hat{V}_n \subseteq \bar{V}_n$, so that $k_{n, \hat{V}_n}(p) \leq k_{n, \bar{V}_n}(p)$. By C.2 for some deterministic term $o(\delta_n)$,

$$P_n \left(\left| \sup_{v \in \bar{V}_n} Z_n^*(v) - \sup_{v \in \bar{V}_n} \bar{Z}_n^*(v) \right| > o(\delta_n) \mid \mathcal{D}_n \right) = o_{P_n}(1/\ell_n), \quad (\text{B.4})$$

which implies by Lemma 11 that for some $\varepsilon_n \searrow 0$, $\text{wp} \rightarrow 1$, for all $p \in (\varepsilon_n, 1 - \varepsilon_n)$

$$k_{n, \bar{V}_n}(p) \leq \kappa_{n, \bar{V}_n}(p + \varepsilon_n) + o(\delta_n) \quad (\text{B.5})$$

where the terms $o(\delta_n)$ are different in different places. By C.3, for any fixed $p \in (0, 1)$,

$$\kappa_{\bar{V}_n}(p + \varepsilon_n) \leq a_n(\bar{V}_n) + c(p + \varepsilon_n)/a_n(\bar{V}_n) = a_n(\bar{V}_n) + O(1)/a_n(\bar{V}_n).$$

Thus, combining inequalities above and $o(\delta_n) = o(\bar{a}_n^{-1}) = o(a_n^{-1}(\bar{V}_n))$ by C.2, $\text{wp} \rightarrow 1$,

$$k_{n, \hat{V}_n}(p) \leq a_n(\bar{V}_n) + O(1)/a_n(\bar{V}_n).$$

Now Parts 2 and 3 follow as in the Proof of Parts 2 and 3 of Theorem 1 using this bound on the simulated critical value instead of the bound on the analytical critical value. \blacksquare

B.5. Proof of Lemma 3 (Concentration on V_0). By Conditions S and V

$$\left| \sup_{v \in V_n} Z_n^*(v) - \sup_{v \in V_0} Z_n^*(v) \right| \leq \sup_{\|v - v'\| \leq O_{P_n}(r_n)} |Z_n^*(v) - Z_n^*(v')| = \sup_{v \in V_0} Z_n^*(v) + o_{P_n}(\delta_n). \quad (\text{B.6})$$

Conclude similarly to the proof of Lemma 1, using anti-concentration Corollary 2

$$\mathbb{P}_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \leq x \right) \geq \mathbb{P}_n \left(\sup_{v \in V_0} Z_n^*(v) + o(\delta_n) \leq x \right) - o(1) \geq \mathbb{P}_n \left(\sup_{v \in V_0} Z_n^*(v) \leq x \right) - o(1)$$

This gives a lower bound. Similarly, using C.3 and C.4 and anti-concentration Corollary 2

$$\begin{aligned} \mathbb{P}_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \leq x \right) &\leq \mathbb{P}_n \left(\sup_{v \in V_0} Z_n(v) \frac{\sigma_n(v)}{s_n(v)} \leq x \right) \\ &\leq \mathbb{P}_n \left(\sup_{v \in V_0} Z_n^*(v) - o(\delta_n) \leq x \right) + o(1) \leq \mathbb{P}_n \left(\sup_{v \in V_0} Z_n^*(v) \leq x \right) + o(1) \end{aligned}$$

where $o(\cdot)$ terms above are different in different places, and the first inequality follows from

$$\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \geq \sup_{v \in V_0} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} = \sup_{v \in V_0} Z_n(v) \frac{\sigma_n(v)}{s_n(v)}.$$

This gives the upper bound. ■

B.6. Proof of Theorem 3 (When Simulation Inference Becomes Sharp). Part 1.

(Size) By Lemma 2 $\text{wp} \rightarrow 1$, $\hat{V}_n \subseteq \bar{V}_n$, so that $k_{n, \hat{V}_n}(p) \leq k_{n, \bar{V}_n}(p)$ $\text{wp} \rightarrow 1$. So let us compare critical values

$$k_{n, \bar{V}_n}(p) = Q_p \left(\sup_{v \in \bar{V}_n} Z_n^*(v) \mid \mathcal{D}_n \right) \text{ and } \kappa_{n, V_0}(p) = Q_p \left(\sup_{v \in V_0} \bar{Z}_n^*(v) \right).$$

The former is data-dependent while the latter is deterministic. Recall that by C.2 $\text{wp} \rightarrow 1$ we have (B.4). By Condition V $d_H(\bar{V}_n, V_0) \leq r_n$, and so by S, we have for some $o(\delta_n)$,

$$\mathbb{P}_n \left(\left| \sup_{v \in \bar{V}_n} \bar{Z}_n^*(v) - \sup_{v \in V_0} \bar{Z}_n^*(v) \right| > o(\delta_n) \mid \mathcal{D}_n \right) = o_{\mathbb{P}_n}(1).$$

Combining (B.4) and this relation, and by monotonicity of $V \mapsto k_{n, V}(p) = a_n(V) + c(p)/a_n(V)$ for sufficiently large n by C.3, we obtain that for some $o(\delta_n)$ term,

$$\mathbb{P}_n \left(\left| \sup_{v \in \bar{V}_n} Z_n^*(v) - \sup_{v \in V_0} \bar{Z}_n^*(v) \right| > o(\delta_n) \mid \mathcal{D}_n \right) = o_{\mathbb{P}_n}(1).$$

This implies by Lemma 11 that for some $\varepsilon_n \searrow 0$, and any $p \in (\varepsilon_n, 1 - \varepsilon_n)$, $\text{wp} \rightarrow 1$,

$$k_{n, \hat{V}_n}(p) \leq k_{n, \bar{V}_n}(p) \leq \kappa_{n, V_0}(p + \varepsilon_n) + o(\delta_n). \quad (\text{B.7})$$

Hence, for any fixed p ,

$$\begin{aligned}
P_n \left(\theta_{n0} \leq \widehat{\theta}_{n0}(p) \right) &= P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \widehat{\theta}_n(v)}{s_n(v)} \leq k_{n, \widehat{V}_n}(p) \right) \\
&\leq_{(1)} P_n \left(\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \widehat{\theta}_n(v)}{s_n(v)} \leq \kappa_{n, V_0}(p + \varepsilon_n) + o(\delta_n) \right) + o(1) \\
&\leq_{(2)} P_n \left(\sup_{v \in V_0} Z_n^*(v) \leq \kappa_{n, V_0}(p + \varepsilon_n) + o(\delta_n) \right) + o(1) \leq_{(3)} p + \varepsilon_n + o(1) = p + o(1),
\end{aligned}$$

where (1) is by the quantile comparison (B.7), (2) is by Lemma 3, and (3) is by anti-concentration Corollary 2. Combining this with the lower bound of Theorem 2, we have the result.

Parts 2 & 3. (Estimation Risk and Power) We have that by C.3

$$\kappa_{n, V_0}(p + \varepsilon_n) \leq a_n(V_0) + c(p + \varepsilon_n)/a_n(V_0) = a_n(V_0) + O(1)/a_n(V_0).$$

Hence combining this with equation (B.7) we have $\text{wp} \rightarrow 1$

$$k_{n, \widehat{V}_n}(p) \leq a_n(V_0) + O(1)/a_n(V_0) + o(\bar{a}_n^{-1}) = a_n(V_0) + O(1)/a_n(V_0),$$

where $o(\delta_n) = o(\bar{a}_n^{-1}) = o(a_n^{-1}(V_0))$ by C.2. Then Parts 2 and 3 follow identically to the Proof of Parts 2 and 3 of Theorem 1 using this bound on the simulated critical value instead of the bound on the analytical critical value. \blacksquare

APPENDIX C. PROOFS FOR SECTION 4

C.1. Tools and Auxiliary Lemmas. We shall heavily rely on the Talagrand-Samorodnitsky Inequality, which was obtained by Talagrand sharpening earlier results by Samorodnitsky. Here it is restated from van der Vaart and Wellner (1996) Proposition A.2.7, page 442:

Talagrand-Samorodnitsky Inequality: Let X be a separable zero-mean Gaussian process indexed by a set T . Suppose that for some $\Gamma > \sigma(X) = \sup_{t \in T} \sigma(X_t)$, $0 < \epsilon_0 \leq \sigma(X)$,

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

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 exists a universal constant D such that for every $\lambda \geq \sigma^2(X)(1 + \sqrt{\nu})/\epsilon_0$ we have

$$P \left(\sup_{t \in T} X_t > \lambda \right) \leq \left(\frac{D\Gamma\lambda}{\sqrt{\nu}\sigma^2(X)} \right)^\nu (1 - \Phi(\lambda/\sigma(X))), \quad (\text{C.1})$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

The following lemma is an application of this inequality that we use:

Lemma 12 (Concentration Inequality via Talagrand-Samorodnitsky). *Let Z_n be a separable zero-mean Gaussian process indexed by a set V such that $\sup_{v \in V} \sigma(Z_n(v)) = 1$. Suppose that for some $\Gamma_n(V) > 1$, and $d \geq 1$*

$$N(\varepsilon, V, \rho) \leq \left(\frac{\Gamma_n(V)}{\varepsilon} \right)^d, \quad \text{for } 0 < \varepsilon < 1,$$

where $N(\varepsilon, V, \rho)$ is the covering number of V by ε -balls w.r.t. the standard deviation metric $\rho(v, v') = \sigma(Z_n(v) - Z_n(v'))$. Then for

$$a_n(V) = (2\sqrt{\log L_n(V)}) \vee (1 + \sqrt{d}), \quad L_n(V) := C'_n \left(\frac{\Gamma_n(V)}{\sqrt{d}} \right)^d,$$

where for D denoting Talagrand's constant in (C.1), and C'_n such that

$$C'_n \geq D^d C_d \frac{1}{\sqrt{2\pi}}, \quad C_d := \max_{\lambda \geq 0} \lambda^{d-1} e^{-\lambda^2/4},$$

we have for $z \geq 0$

$$P \left(a_n(V) \left(\sup_{v \in V} Z_n(v) - a_n(V) \right) > z \right) \leq \exp \left(-\frac{z}{2} - \frac{z^2}{4a_n^2(V)} \right) \leq \exp(-z/2).$$

Proof. We apply the TS inequality by setting $t = v$, $X = Z$, $\sigma(X) = 1$, $\epsilon_0 = 1$, $\nu = d$, with $\lambda \geq (1 + \sqrt{d})$, so that

$$\begin{aligned} P \left(\sup_{v \in V} Z_n(v) > \lambda \right) &\leq \left(\frac{D\kappa_n(V)\lambda}{\sqrt{d}} \right)^d (1 - \Phi(\lambda)) \\ &\leq \left(\frac{D\Gamma_n(V)\lambda}{\sqrt{d}} \right)^d \frac{1}{\sqrt{2\pi}} \frac{1}{\lambda} e^{-\lambda^2/2} \leq L_n(V) e^{-\lambda^2/4}. \end{aligned}$$

Setting for $z \geq 0$, $\lambda = \frac{z}{a_n(V)} + a_n(V) \geq (1 + \sqrt{d})$, we obtain

$$L_n(V) \exp \left(-\frac{\lambda^2}{4} \right) \leq \exp \left(-\frac{z}{2} - \frac{z^2}{4a_n^2(V)} \right).$$

■

The following lemma is an immediate consequence of Corollary 2.2.8 of van der Vaart and Wellner (1996).

Lemma 13 (Maximal Inequality for a Gaussian Process). *Let X be a separable zero-mean Gaussian process indexed by a set T . Then for every $\delta > 0$*

$$E \sup_{\rho(s,t) \leq \delta} |X_s - X_t| \lesssim \int_0^\delta \sqrt{\log N(\varepsilon, T, \rho)} d\varepsilon, \quad E \sup_{t \in T} |X_t| \leq \sigma(X) + \int_0^{2\sigma(X)} \sqrt{\log N(\varepsilon, T, \rho)} d\varepsilon,$$

where $\sigma(X) = \sup_{t \in T} \sigma(X_t)$, and $N(\varepsilon, T, \rho)$ is the covering number of T with respect to the semi-metric $\rho(s, t) = \sigma(X_s - X_t)$.

Proof. The first conclusion follows from Corollary 2.2.8 of van der Vaart and Wellner (1996) since covering and packing numbers are related by $N(\varepsilon, T, \rho) \leq D(\varepsilon, T, \rho) \leq N(\varepsilon/2, T, \rho)$. The second conclusion follows from the special case of the first conclusion: for any $t_0 \in T$, $E \sup_{t \in T} |X_t| \leq E|X_{t_0}| + \int_0^{\text{diam}(T)} \sqrt{\log N(\varepsilon, T, \rho)} d\varepsilon \leq \sigma(X) + \int_0^{2\sigma(X)} \sqrt{\log N(\varepsilon, T, \rho)} d\varepsilon$. ■

C.2. Proof of Lemma 5. Step 1. Verification of C.1. This condition holds by inspection, in view of continuity of $v \mapsto p_n(v)$ and by Ω_n and $\widehat{\Omega}_n$ being positive definite.

Step 2. Verification of C.2 Set $\delta_n = 1/\log n$. Condition NS.1 directly assumes C.2(a).

In order to show C.2(b), we employ the maximal inequality stated in Lemma 13. Set $X_t = Z_n^*(v) - Z_n^*(v)$, $t = v$, $T = \mathcal{V}$ and note that for some absolute constant C , conditional on \mathcal{D}_n ,

$$N(\varepsilon, T, \rho) \leq \left(\frac{1 + C\Upsilon_n \text{diam}(T)}{\varepsilon} \right)^d, \quad 0 < \varepsilon < 1,$$

since $\sigma(X_t - X_{t'}) \lesssim \Upsilon_n \|t - t'\|$, $T \subset \mathbb{R}^d$, where Υ_n is an upper bound on the Lipschitz constant of the function

$$v \mapsto \frac{p_n(v)' \Omega_n^{1/2}}{\|p_n(v)' \Omega_n^{1/2}\|} - \frac{p_n(v)' \widehat{\Omega}_n^{1/2}}{\|p_n(v)' \widehat{\Omega}_n^{1/2}\|},$$

where $\text{diam}(T)$ is the diameter of set T under the Euclidian metric. Using inequality (E.6) we can bound

$$\Upsilon_n \leq 2L_n \frac{\lambda_{\max}(\Omega_n^{1/2})}{\lambda_{\min}(\Omega_n^{1/2})} + 2L_n \frac{\lambda_{\max}(\widehat{\Omega}_n^{1/2})}{\lambda_{\min}(\widehat{\Omega}_n^{1/2})} = O_{P_n}(L_n),$$

where L_n is the constant defined in NS.1, and by assumption $\log L_n \lesssim \log n$. Here we use the fact the eigenvalues of Ω_n and $\widehat{\Omega}_n$ are bounded away from zero and from above by NS.1 and NS.2. Therefore, $\log N(\varepsilon, T, \rho) \lesssim \log n + \log(1/\varepsilon)$.

Using (E.6) again,

$$\begin{aligned} \sigma(X) &\lesssim \sup_{v \in \mathcal{V}} \left\| \frac{p_n(v)' \Omega_n^{1/2}}{\|p_n(v)' \Omega_n^{1/2}\|} - \frac{p_n(v)' \widehat{\Omega}_n^{1/2}}{\|p_n(v)' \widehat{\Omega}_n^{1/2}\|} \right\| \\ &\leq \sup_{v \in \mathcal{V}} 2 \frac{\|p_n(v)' (\widehat{\Omega}_n^{1/2} - \Omega_n^{1/2})\|}{\|p_n(v)' \Omega_n^{1/2}\|} \leq \sup_{v \in \mathcal{V}} 2 \frac{\|p_n(v)' \Omega_n^{1/2} (\Omega_n^{1/2} \widehat{\Omega}_n^{-1/2} - I)\|}{\|p_n(v)' \Omega_n^{1/2}\|} \\ &\leq \|\Omega_n^{1/2} \widehat{\Omega}_n^{-1/2} - I\| \leq \|\Omega_n^{-1/2}\| \|\widehat{\Omega}_n^{1/2} - \Omega_n^{1/2}\| = O_{P_n}(n^{-b}) \end{aligned}$$

for some constant $b > 0$, where we have used that the eigenvalues of Ω_n and $\widehat{\Omega}_n$ are bounded away from zero and from above under NS.1 and NS.2, and the assumption $\|\widehat{\Omega}_n - \Omega_n\| = O_{P_n}(n^{-b})$. Hence

$$E \left(\sup_{t \in T} |X_t| \mid \mathcal{D}_n \right) \lesssim \sigma(X) + \int_0^{2\sigma(X)} \sqrt{\log(n/\varepsilon)} d\varepsilon = O_{P_n}(n^{-b} \sqrt{\log n}).$$

Hence for each $C > 0$

$$P_n \left(\sup_{v \in \mathcal{V}} |Z_n^*(v) - Z_n^*(\tilde{v})| > C\delta_n \mid \mathcal{D}_n \right) \lesssim \frac{1}{C\delta_n} O_{P_n}(n^{-b} \sqrt{\log n}) = o_{P_n}(1/\ell_n),$$

which verifies C.2(b).

Step 3. Verification of C.3. We shall employ Lemma 12, which has the required notation in place. We only need to compute an upper bound on the covering numbers $N(\varepsilon, \mathbf{V}, \rho)$ for the process Z_n^* . We have that

$$\begin{aligned} \sigma(Z_n^*(v) - Z_n^*(\tilde{v})) &\leq \left\| \frac{p_n(v)' \Omega_n^{1/2}}{\|p_n(v)' \Omega_n^{1/2}\|} - \frac{p_n(\tilde{v})' \Omega_n^{1/2}}{\|p_n(\tilde{v})' \Omega_n^{1/2}\|} \right\| \leq 2 \left\| \frac{(p_n(v) - p_n(\tilde{v}))' \Omega_n^{1/2}}{\|p_n(v)' \Omega_n^{1/2}\|} \right\| \\ &\leq 2L_n \frac{\lambda_{\max}(\Omega_n^{1/2})}{\lambda_{\min}(\Omega_n^{1/2})} \|v - \tilde{v}\| \leq CL_n \|v - \tilde{v}\|, \end{aligned}$$

where C is some constant that does not depend on n , by the eigenvalues of Ω_n bounded away from zero and from above. Hence it follows that

$$N(\varepsilon, \mathbf{V}, \rho) \leq \left(\frac{1 + CL_n \text{diam}(\mathbf{V})}{\varepsilon} \right)^d, \quad 0 < \varepsilon < 1,$$

where the diameter of \mathbf{V} is measured by the Euclidian metric. Condition C.3 now follow by Lemma 12, with $a_n(\mathbf{V}) = (2\sqrt{\log L_n(\mathbf{V})}) \vee (1 + \sqrt{d})$, $L_n(\mathbf{V}) = C' (1 + CL_n \text{diam}(\mathbf{V}))^d$, where C' is a constant from Lemma 12.

Step 4. Verification of C.4. Under Condition NS, we have that

$$a_n(\mathbf{V}) \leq \bar{a}_n := a_n(\mathcal{V}) \lesssim \sqrt{\log \ell_n + \log n} \lesssim \sqrt{\log n},$$

so that C.4(a) follows if $\sqrt{\log n} \sqrt{\xi^2(K)/n} \rightarrow 0$.

To verify C.4(b) note that uniformly in $v \in \mathcal{V}$,

$$\begin{aligned} \left| \frac{\|p_n(v)' \widehat{\Omega}_n^{1/2}\|}{\|p_n(v)' \Omega_n^{1/2}\|} - 1 \right| &\leq \left| \frac{\|p_n(v)' \widehat{\Omega}_n^{1/2}\| - \|p_n(v)' \Omega_n^{1/2}\|}{\|p_n(v)' \Omega_n^{1/2}\|} \right| \\ &\leq \frac{\|p_n(v)' (\widehat{\Omega}_n^{1/2} - \Omega_n^{1/2})\|}{\|p_n(v)' \Omega_n^{1/2}\|} \leq \frac{\|p_n(v)' \Omega_n^{1/2} (\Omega_n^{1/2} \widehat{\Omega}_n^{-1/2} - I)\|}{\|p_n(v)' \Omega_n^{1/2}\|} \\ &\leq \|\Omega_n^{1/2} \widehat{\Omega}_n^{-1/2} - I\| \leq \|\Omega_n^{-1/2}\| \|\widehat{\Omega}_n^{1/2} - \Omega_n^{1/2}\| = o_{P_n}(\delta_n / \bar{a}_n), \end{aligned}$$

by $\|\widehat{\Omega}_n^{1/2} - \Omega_n^{1/2}\| = O_{P_n}(n^{-b})$ and $\|\Omega_n^{-1/2}\|$ bounded, both implied by the assumptions. ■

C.3. Proof of Lemma 6. To show claim (1), we need to establish that for $\varphi_n = o(1) \cdot \left(\frac{1}{L_n \sqrt{\log n}}\right)$, with any $o(1)$ term, we have that $\sup_{\|v - \tilde{v}\| \leq \varphi_n} |Z_n^*(v) - Z_n^*(\tilde{v})| = o_{P_n}(1)$.

Consider the stochastic process $X = \{Z_n(v), v \in \mathcal{V}\}$. We shall use the standard maximal inequality stated in Lemma 13. From the proof of Lemma 5 we have $\sigma(Z_n^*(v) - Z_n^*(\tilde{v})) \leq CL_n \|v - \tilde{v}\|$, where C is some constant that does not depend on n , and $\log N(\varepsilon, \mathbf{V}, \rho) \lesssim \log n + \log(1/\varepsilon)$. Since $\|v - \tilde{v}\| \leq \varphi_n \implies \sigma(Z_n^*(v) - Z_n^*(\tilde{v})) \leq C \frac{o(1)}{\sqrt{\log n}}$ we have

$$E \sup_{\|v - \tilde{v}\| \leq \varphi_n} |X_s - X_t| \lesssim \int_0^{C \frac{o(1)}{\sqrt{\log n}}} \sqrt{\log(n/\varepsilon)} d\varepsilon \lesssim \frac{o(1)}{\sqrt{\log n}} \sqrt{\log n} = o(1).$$

Hence the conclusion follows from Markov's Inequality.

Under Condition V by Lemma 2 $r_n \lesssim \left(\sqrt{\log n \frac{\zeta_n}{n}}\right)^{1/\rho_n} c_n^{-1}$, so $r_n = o(\varphi_n)$ if

$$\left(\sqrt{\log n \frac{\zeta_n}{n}}\right)^{1/\rho_n} c_n^{-1} = o\left(\frac{1}{L_n \sqrt{\log n}}\right). \quad (\text{C.2})$$

Thus, Condition S holds. The remainder of the lemma follows by direct calculation. ■

APPENDIX D. PROOFS FOR SECTION 5

D.1. Proof of Theorem 7 and Corollary 1. The first step of our proof uses Yurinskii's (1977) coupling. For completeness we now state the formal result from Pollard (2002), page 244.

Yurinskii's Coupling: Consider a sufficiently rich probability space (A, \mathcal{A}, P) . Let ξ_1, \dots, ξ_n be independent K_n -vectors with $E\xi_i = 0$ for each i , and $\Delta := \sum_i E\|\xi_i\|^3$ finite. Let $S = \xi_1 + \dots + \xi_n$. For each $\delta > 0$ there exists a random vector T with $N(0, \text{var}(S))$

distribution such that

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

for some universal constant C_0 .

The proof has two steps: in the first, we couple the estimator $\sqrt{n}(\widehat{\beta}_n - \beta_n)$ with the normal vector; in the second, we establish the strong approximation.

STEP 1. In order to apply the coupling, consider

$$\sum_{i=1}^n \xi_i, \quad \xi_i = u_{i,n}/\sqrt{n} \sim (0, I_{K_n}/n),$$

Then we have that $\sum_{i=1}^n E\|\xi_i\|^3 = \Delta_n$. Therefore, by Yurinskii's coupling,

$$P_n \left\{ \left\| \sum_{i=1}^n \xi_i - \mathcal{N}_n \right\| \geq 3\delta_n \right\} \lesssim K_n \Delta_n / \delta_n^3 \rightarrow 0,$$

Combining this with the assumption on the linearization error r_n , we obtain

$$\begin{aligned} \|\Omega_n^{-1/2} \sqrt{n}(\widehat{\beta}_n - \beta_n) - \mathcal{N}_n\| &\leq \left\| \sum_{i=1}^n \xi_i - \mathcal{N}_n \right\| + \|\Omega_n^{-1/2} \sqrt{n}(\widehat{\beta}_n - \beta_n) - \sum_{i=1}^n \xi_i\| \\ &= o_{P_n}(\delta_n) + r_n = o_{P_n}(\delta_n). \end{aligned}$$

STEP 2. Using the result of Step 1 and that

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

we conclude that

$$\begin{aligned} |S_n(v)| &:= \left| \frac{\sqrt{n} g_n(v)' \Omega_n^{-1/2}(\widehat{\beta}_n - \beta_n)}{\|g_n(v)\|} - \frac{g_n(v)' \mathcal{N}_n}{\|g_n(v)\|} \right| \\ &\leq \left\| \sqrt{n} \Omega_n^{-1/2}(\widehat{\beta}_n - \beta_n) - \mathcal{N}_n \right\| = o_{P_n}(\delta_n), \end{aligned} \tag{D.1}$$

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

$$\begin{aligned} &\sup_{v \in \mathcal{V}} \left| \frac{\sqrt{n}(\widehat{\theta}_n(v) - \theta_n(v))}{\|g_n(v)\|} - \frac{g_n(v)' \mathcal{N}_n}{\|g_n(v)\|} \right| \\ &\leq \sup_{v \in \mathcal{V}} \left| \frac{\sqrt{n}(\widehat{\theta}_n(v) - \theta_n(v))}{\|g_n(v)\|} - \frac{\sqrt{n} g_n(v)' \Omega_n^{-1/2}(\widehat{\beta}_n - \beta_n)}{\|g_n(v)\|} \right| \\ &\quad + \sup_{v \in \mathcal{V}} \left| \frac{\sqrt{n} g_n(v)' \Omega_n^{-1/2}(\widehat{\beta}_n - \beta_n)}{\|g_n(v)\|} - \frac{g_n(v)' \mathcal{N}_n}{\|g_n(v)\|} \right| \\ &= \sup_{v \in \mathcal{V}} |\sqrt{n} A_n(v) / \|g_n(v)\|| + \sup_{v \in \mathcal{V}} |S_n(v)| = o(\delta_n) + o_{P_n}(\delta_n), \end{aligned}$$

using the assumption on the approximation error $A_n(v) = \theta(v) - p_n(v)' \beta_n$ and (D.1). This proves the theorem.

STEP 3. To show the corollary note that

$$E_{P_n} \|u_{i,n}\|^3 \leq \|\Omega_n^{-1/2} Q_n^{-1}\|^3 \cdot E_{P_n} \|p_n(V_i) \epsilon_i\|^3 \lesssim \tau_n^3 K_n^{3/2} C_n,$$

using the boundedness assumptions stated in the corollary. ■

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 P. GUGGENBERGER (2009): “Validity of Subsampling and Plug-In Asymptotic Inference for Parameters Defined by Moment Inequalities,” *Econometric Theory*, 25, 669–709.
- 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 (forthcoming): “Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure,” *Econometrica*.
- ANDREWS, D. W. K., AND X. SHI (forthcoming): “Inference for Parameters Defined by Conditional Moment Inequalities,” *Econometrica*.
- ANDREWS, D. W. K., AND G. SOARES (2010): “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” *Econometrica*, 78(1), 119–157.
- ARMSTRONG, T. B. (2011a): “Asymptotically Exact Inference in Conditional Moment Inequality Models,” working paper, Stanford University.
- (2011b): “Weighted KS Statistics for Inference on Conditional Moment Inequalities,” working paper, Stanford University.
- BELLONI, A., X. CHEN, AND V. CHERNOZHUKOV (2010): “(Some) New Asymptotic Theory for Linear Series Estimators,” working paper, MIT.
- BELLONI, A., V. CHERNOZHUKOV, AND I. FERNANDEZ-VAL (2011): “Conditional Quantile Processes under Increasing Dimensions,” CeMMAP working paper CWP19/11.
- 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.
- BICKEL, P. J., AND M. ROSENBLATT (1973): “On Some Global Measures of the Deviations of Density Function Estimates,” *Annals of Statistics*, 1(6), 1071–1095.
- BIERENS, H. (1982): “Consistent Model Specification Tests,” *Journal of Econometrics*, 20, 105–134.
- 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. (2010): “Bootstrap Inference for Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set,” *Econometrica*, 78(2), 735–753.
- CANAY, I. (2010): “EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity,” *Journal of Econometrics*, 156(2), 408–425.
- 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.
- CHEN, X. (2007): “Large Sample Sieve Estimation of Semi-Nonparametric Models (Chapter 76),” vol. 6, Part 2 of *Handbook of Econometrics*, pp. 5549 – 5632. Elsevier.
- 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., AND K. KATO (2011): “Anti-Concentration and Honest Adaptive Confidence Bands,” working paper, Hiroshima University.
- CHERNOZHUKOV, V., S. LEE, AND A. ROSEN (2009): “Intersection Bounds: Estimation and Inference,” CeMMAP working paper CWP19/09.
- CHERNOZHUKOV, V., R. RIGOBON, AND T. STOKER (2010): “Set Identification and Sensitivity Analysis with Tobin Regressors,” *Quantitative Economics*, 1(2), 255–277.
- CHESHER, A. D. (2010): “Instrumental Variable Models for Discrete Outcomes,” *Econometrica*, 78(2), 575–601.
- CHETVERIKOV, D. (2011): “Adaptive rate optimal test of conditional moment inequalities,” working paper, MIT.
- DUDLEY, R., AND W. PHILIPP (1983): “Invariance Principles for Sums of Banach Space Valued Random Elements and Empirical Processes,” *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 62, 509–552.
- FAN, J., AND I. GIJBELS (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London, UK.
- FAN, Y., AND S. PARK (2011): “Bias-Correction and Confidence Sets Under IV and MIV Assumptions,” Vanderbilt University working paper.
- GALICHON, A., AND M. HENRY (2009): “A Test of Non-identifying Restrictions and Confidence Regions for Partially Identified Parameters,” *Journal of Econometrics*, 159(2), 186–196.
- 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 E. MAMMEN (1993): “Comparing Nonparametric Versus Parametric Regression Fits,” *Annals of Statistics*, 21(4), 1926–1947.
- HECKMAN, J. J., J. SMITH, AND N. CLEMENTS (1997): “Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts,” *Review of Economic Studies*, 64, 487–537.

- HECKMAN, J. J., AND E. J. VYTLACIL (1999): “Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects,” *Proceedings of the National Academy of Sciences*, 96, 4730–4734.
- HIRANO, K., AND J. R. PORTER (2012): “Impossibility Results for Nondifferentiable Functionals,” *Econometrica*, 80(4), 1769–1790.
- HOROWITZ, J., AND V. G. SPOKOINY (2001): “An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative,” *Econometrica*, 69(3), 599–631.
- HOROWITZ, J. L., AND E. MAMMEN (2004): “Nonparametric estimation of an additive model with a link function,” *Ann. Statist.*, 32(6), 2412–2443.
- 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.
- KOENKER, R. (2005): *Quantile Regression*. Cambridge University Press, Cambridge.
- KONG, E., O. LINTON, AND Y. XIA (2010): “Uniform Bahadur representation for local polynomial estimates of m-regression and its application to the additive model,” *Econometric Theory*, 26, 1529–1564.
- 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., K. SONG, AND Y.-J. WHANG (forthcoming): “Testing Functional Inequalities,” *Journal of Econometrics*.
- 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.
- (1997): “Monotone Treatment Response,” *Econometrica*, 65(6), 1311–1334.
- (2003): *Partial Identification of Probability Distributions*. Springer-Verlag, New York.
- 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.
- MASRY, E. (1996): “Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates,” *Journal of Time Series Analysis*, 17, 571–599.
- MASSART, P. (1989): “Strong Approximation for Multivariate Empirical and Related Processes, via KMT Constructions,” *Annals of Probability*, 17(1), 266–291.
- MENZEL, K. (2009): “Estimation and Inference with Many Weak Moment Inequalities,” working paper, MIT.
- NEVO, A., AND A. M. ROSEN (2012): “Identification with Imperfect Instruments,” *Review of Economics and Statistics*, 94(3), 659–671.
- NEWWEY, 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. Srinivasan, pp. 254–275. Blackwell, Cambridge, U.S.A.

- (1997): “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, 79, 147–168.
- NICOLETTI, C., F. FOLIANO, AND F. PERACCHI (2011): “Estimating Income Poverty in The Presence Of Missing Data and Measurement Error Problems,” *Journal of Business and Economic Statistics*, 29(1), 61–72.
- PAKES, A., J. PORTER, K. HO, AND J. ISHII (2005): “The Method of Moments with Inequality Constraints,” working paper, Harvard University.
- POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer, New York, NY.
- (2002): *A User’s Guide to Measure Theoretic Probability*. Cambridge University Press, Cambridge.
- PONOMAREVA, M. (2010): “Inference in Models Defined by Conditional Moment Inequalities with Continuous Covariates,” working paper, University of Western Ontario.
- POWELL, J. L. (1984): “Least Absolute Deviations Estimation for the Censored Regression Model,” *Journal of Econometrics*, 25(3), 303–325.
- 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.
- (2010): “Inference for the Identified Set in Partially Identified Econometric Models,” *Econometrica*, 78(1), 169–211.
- 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. (1999): “Random vectors in the isotropic position,” *Journal of Functional Analysis*, 164(1), 60–72.
- SHAIKH, A. M., AND E. J. VYTLACIL (2011): “Partial Identification in Triangular Systems of Equations with Binary Dependent Variables,” *Econometrica*, 79(3), 949–955.
- SKOROHOD, A. (1956): “Limit Theorems for Stochastic Processes,” *Theory of Probability and Its Applications*, 1, 261–290.
- STONE, C. J. (1982): “Optimal Global Rates of Convergence for Nonparametric Regression,” *Annals of Statistics*, 10(4), 1040–1053.
- STOYE, J. (2009): “More on Confidence Regions for Partially Identified Parameters,” *Econometrica*, 77(4), 1299–1315.
- TSYBAKOV, A. B. (2009): *Introduction to Nonparametric Estimation*. Springer, New York, NY.
- VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY.
- WHITE, H. (1984): *Asymptotic Theory for Econometricians*. Academic Press, New York, NY.
- YURINSKII, V. (1977): “On the Error of the Gaussian Approximation for Convolutions,” *Theory of Probability and Its Applications*, 22, 236–247.

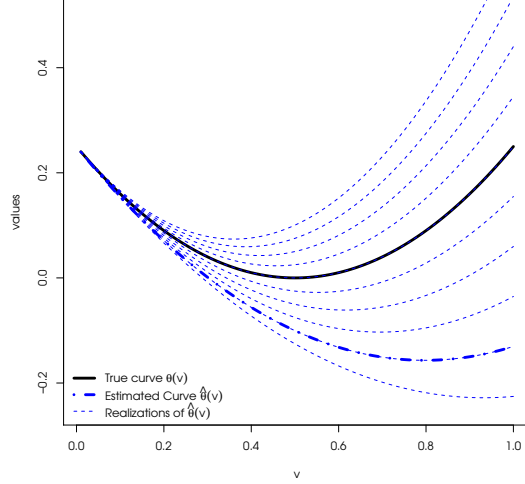


FIGURE 1. This figure illustrates how variation in the precision of the analog estimator at different points may impede inference. The solid curve is the true bounding function $\theta(v)$, while the dash-dot curve is a single realization of its estimator, $\hat{\theta}(v)$. The lighter dashed curves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

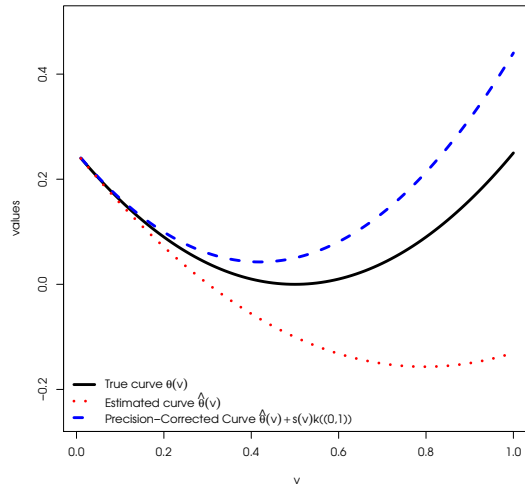


FIGURE 2. This figure depicts a precision-corrected curve (dashed curve) that adjusts the boundary estimate $\hat{\theta}(v)$ (dotted curve) by an amount proportional to its point-wise standard error. The minimum of the precision-corrected curve is closer to the minimum of the true curve (solid) than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

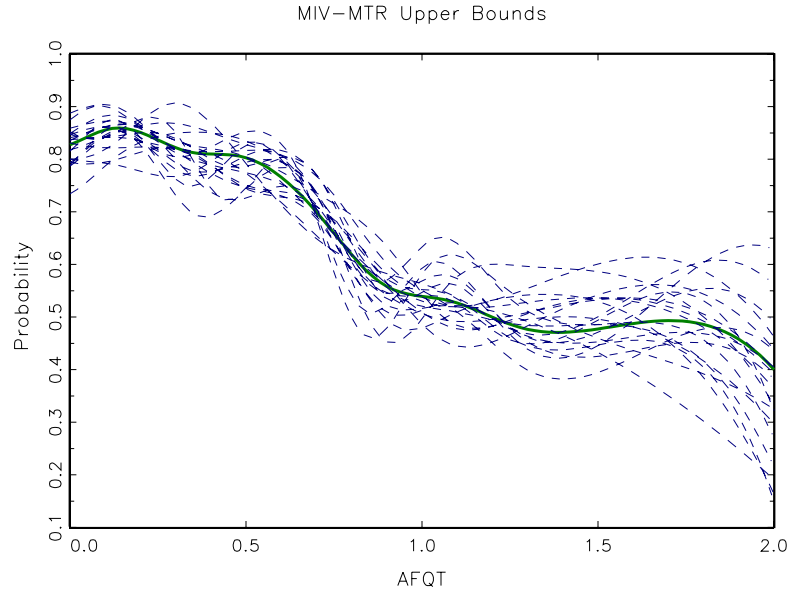


FIGURE 3. This figure, based on NLSY data and the application in Section 5 of Chernozhukov, Lee, and Rosen (2009), depicts an estimate of the bounding function (solid curve) and 20 bootstrapped estimates (dashed curves).

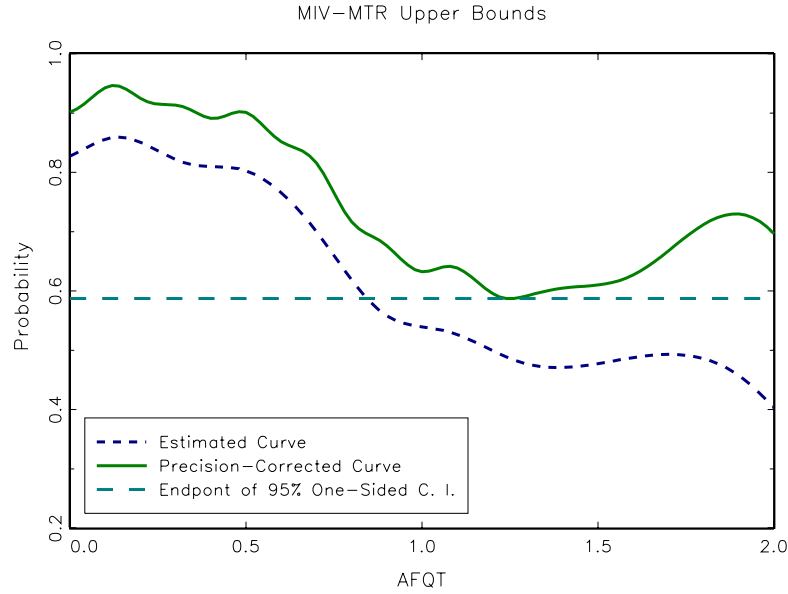


FIGURE 4. This figure, based on NLSY data and the application in Section 5 of Chernozhukov, Lee, and Rosen (2009), depicts a precision-corrected curve (solid curve) that adjusts the boundary estimate $\hat{\theta}(v)$ (dashed curve) by an amount proportional to its point-wise standard error. The horizontal dashed line shows the end point of 95% one-sided confidence interval.

TABLE 1. Results for Monte Carlo Experiments (Series Estimation using B-splines)

DGP	Sample Size	Critical Value	Ave. Smoothing Parameter	Cov. Prob.	False Cov. Prob.	Ave. Argmax Set Min.	Max.
CLR with Series Estimation using B-splines							
Estimating V_n ?							
1	500	No	9.610	0.944	0.149	-1.800	1.792
1	500	Yes	9.610	0.944	0.149	-1.800	1.792
1	1000	No	10.490	0.947	0.013	-1.801	1.797
1	1000	Yes	10.490	0.947	0.013	-1.801	1.797
2	500	No	9.680	0.992	0.778	-1.800	1.792
2	500	Yes	9.680	0.982	0.661	-0.762	0.759
2	1000	No	10.578	0.997	0.619	-1.801	1.797
2	1000	Yes	10.578	0.982	0.470	-0.669	0.670
3	500	No	11.584	0.995	0.903	-1.800	1.792
3	500	Yes	11.584	0.984	0.765	-0.342	0.344
3	1000	No	13.378	0.994	0.703	-1.801	1.797
3	1000	Yes	13.378	0.971	0.483	-0.290	0.290
4	500	No	18.802	0.996	0.000	-1.800	1.792
4	500	Yes	18.802	0.974	0.000	-0.114	0.114
4	1000	No	20.572	1.000	0.000	-1.801	1.797
4	1000	Yes	20.572	0.977	0.000	-0.098	0.091

TABLE 2. Results for Monte Carlo Experiments (AS)

DGP	Sample Size	Critical Value	Cov. Prob.	False Cov. Prob.
AS with CvM (Cramér-von Mises-type statistic)				
1	500	PA/Asy	0.959	0.007
1	500	GMS/Asy	0.955	0.007
1	1000	PA/Asy	0.958	0.000
1	1000	GMS/Asy	0.954	0.000
2	500	PA/Asy	1.000	1.000
2	500	GMS/Asy	1.000	0.977
2	1000	PA/Asy	1.000	1.000
2	1000	GMS/Asy	1.000	0.933
3	500	PA/Asy	1.000	1.000
3	500	GMS/Asy	1.000	1.000
3	1000	PA/Asy	1.000	1.000
3	1000	GMS/Asy	1.000	1.000
4	500	PA/Asy	1.000	1.000
4	500	GMS/Asy	1.000	1.000
4	1000	PA/Asy	1.000	1.000
4	1000	GMS/Asy	1.000	1.000

TABLE 3. Results for Monte Carlo Experiments (Other Estimation Methods)

DGP	Sample Size	Critical Value	Ave. Smoothing Parameter	Cov. Prob.	False Cov. Prob.	Ave. Argmax Set Min.	Max.
CLR with Series Estimation using Polynomials							
		Estimating V_n ?					
1	500	No	5.524	0.954	0.086	-1.800	1.792
1	500	Yes	5.524	0.954	0.086	-1.800	1.792
1	1000	No	5.646	0.937	0.003	-1.801	1.797
1	1000	Yes	5.646	0.937	0.003	-1.801	1.797
2	500	No	8.340	0.995	0.744	-1.800	1.792
2	500	Yes	8.340	0.989	0.602	-0.724	0.724
2	1000	No	9.161	0.996	0.527	-1.801	1.797
2	1000	Yes	9.161	0.977	0.378	-0.619	0.620
3	500	No	8.350	0.998	0.809	-1.800	1.792
3	500	Yes	8.350	0.989	0.612	-0.300	0.301
3	1000	No	9.155	0.996	0.560	-1.801	1.797
3	1000	Yes	9.155	0.959	0.299	-0.253	0.252
4	500	No	8.254	1.000	0.000	-1.800	1.792
4	500	Yes	8.254	0.999	0.000	-0.081	0.081
4	1000	No	9.167	0.998	0.000	-1.801	1.797
4	1000	Yes	9.167	0.981	0.000	-0.069	0.069
CLR with Local Linear Estimation							
		Estimating V_n ?					
1	500	No	0.606	0.923	0.064	-1.799	1.792
1	500	Yes	0.606	0.923	0.064	-1.799	1.792
1	1000	No	0.576	0.936	0.003	-1.801	1.796
1	1000	Yes	0.576	0.936	0.003	-1.801	1.796
2	500	No	0.264	0.995	0.871	-1.799	1.792
2	500	Yes	0.264	0.989	0.808	-0.890	0.892
2	1000	No	0.218	0.996	0.779	-1.801	1.796
2	1000	Yes	0.218	0.990	0.675	-0.776	0.776
3	500	No	0.140	0.995	0.943	-1.799	1.792
3	500	Yes	0.140	0.986	0.876	-0.426	0.424
3	1000	No	0.116	0.992	0.907	-1.801	1.796
3	1000	Yes	0.116	0.986	0.816	-0.380	0.377
4	500	No	0.078	0.991	0.000	-1.799	1.792
4	500	Yes	0.078	0.981	0.000	-0.142	0.142
4	1000	No	0.064	0.997	0.000	-1.801	1.796
4	1000	Yes	0.064	0.991	0.000	-0.127	0.127

TABLE 4. Computation Times of Monte Carlo Experiments

	AS	Series (B-splines)	Series (Polynomials)	Local Linear
Total minutes for simulations	8.67	57.99	19.44	83.75
Average seconds for each test	0.03	0.22	0.07	0.31
Relative Ratio relative to AS	1.00	6.69	2.24	9.66

ON-LINE SUPPLEMENT TO “INTERSECTION BOUNDS: ESTIMATION AND INFERENCE”

VICTOR CHERNOZHUKOV, SOKBAE LEE, AND ADAM M. ROSEN

ABSTRACT. This supplement provides appendices to Chernozhukov, Lee, and Rosen (2012) not included in the main text. Appendix E provides proofs omitted from the main text. Appendices F-H concern kernel-type estimators, providing primitive conditions for their application to conditional moment inequalities, strong approximation results, and proofs. Appendix I provides additional details on the use of primitive conditions to verify an asymptotic linear expansion needed for strong approximation of series estimators and Appendix J gives some detailed arguments for local polynomial estimation of conditional moment inequalities. Appendix K provides local asymptotic power analysis that supports the findings of our Monte Carlo experiments. Appendix L provides further Monte Carlo evidence.

KEY WORDS. Bound analysis, conditional moments, partial identification, strong approximation, infinite dimensional constraints, linear programming, concentration inequalities, anti-concentration inequalities, non-Donsker empirical process methods, moderate deviations, adaptive moment selection.

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

Date: 1 September 2012.

Victor Chernozhukov: Department of Economics, Massachusetts Institute of Technology, vchern@mit.edu.

Sokbae Lee: Department of Economics, Seoul National University, sokbae@gmail.com.

Adam Rosen: Department of Economics, University College London and CeMMAP, adam.rosen@ucl.ac.uk.

APPENDIX E. PROOFS OMITTED FROM THE MAIN TEXT.

E.1. Proof of Lemma 2 (Estimation of V_n). There is a single proof for both analytical and simulation methods, but it is convenient for clarity to split the first step of the proof into separate cases. There are four steps in total.

Step 1a. (Bounds on $k_{n,\mathcal{V}}(\gamma_n)$ in Analytical Case) We have that for some constant $\eta > 0$

$$k_{n,\mathcal{V}}(\gamma_n) := \left(\bar{a}_n + \frac{c(\gamma_n)}{\bar{a}_n} \right),$$

$$\kappa_n := \kappa_n(\gamma'_n) := Q_{\gamma'_n} \left(\sup_{v \in \mathcal{V}} Z_n^*(v) \right), \quad \bar{\kappa}_n := 4 \left(\bar{a}_n + \frac{\eta \ell \ell_n}{\bar{a}_n} \right).$$

The claim of this step is that given the sequence γ_n we have for all large n :

$$k_{n,\mathcal{V}}(\gamma_n) \geq \kappa_n(\gamma_n) \tag{E.1}$$

$$3k_{n,\mathcal{V}}(\gamma_n) < \bar{\kappa}_n \tag{E.2}$$

Inequality (E.2) follows from (B.2) in step 2 of the proof of Lemma 1; (E.1) follows immediately from Condition C.3 (with γ_n in place of γ'_n).

Step 1b. (Bounds on $\kappa_{n,\mathcal{V}}(\gamma_n)$ in Simulation Case) We have

$$k_{n,\mathcal{V}}(\gamma_n) := Q_{\gamma_n} \left(\sup_{v \in \mathcal{V}} Z_n^*(v) \mid \mathcal{D}_n \right),$$

$$\kappa_n = \kappa_n(\gamma'_n) := Q_{\gamma'_n} \left(\sup_{v \in \mathcal{V}} \bar{Z}_n^*(v) \right), \quad \bar{\kappa}_n := 4 \left(\bar{a}_n + \frac{\eta \ell \ell_n}{\bar{a}_n} \right).$$

The claim of this step is that given γ_n there is $\gamma'_n = \gamma_n - o(1)$ such that, wp $\rightarrow 1$

$$k_{n,\mathcal{V}}(\gamma_n) \geq \kappa_n(\gamma'_n) \tag{E.3}$$

$$3k_{n,\mathcal{V}}(\gamma_n) < \bar{\kappa}_n \tag{E.4}$$

To show inequality (E.3), note that by C.2 and Lemma 11 wp $\rightarrow 1$

$$\kappa_{n,\mathcal{V}}(\gamma_n + o(1/\ell_n)) + o(\delta_n) \geq k_{n,\mathcal{V}}(\gamma_n) \geq \kappa_{n,\mathcal{V}}(\gamma_n - o(1/\ell_n)) - o(\delta_n). \tag{E.5}$$

Hence (E.3) follows from

$$\begin{aligned} & \mathbb{P}_n \left(\sup_{v \in \mathcal{V}} \bar{Z}_n^*(v) \leq x \right) \Big|_{x=k_{n,\mathcal{V}}(\gamma_n)} \\ & \geq_{(1)} \mathbb{P}_n \left(\sup_{v \in \mathcal{V}} \bar{Z}_n^*(v) \leq \kappa_{n,\mathcal{V}}(\gamma_n - o(1/\ell_n)) - o(\delta_n) \right) - o(1) \text{ wp } \rightarrow 1 \\ & \geq_{(2)} \mathbb{P}_n \left(\sup_{v \in \mathcal{V}} \bar{Z}_n^*(v) \leq \kappa_{n,\mathcal{V}}(\gamma_n - o(1/\ell_n)) \right) - o(1) = \gamma_n - o(1/\ell_n) - o(1) =: \gamma'_n, \end{aligned}$$

where (1) holds by (E.5) and by C.2 (b) and (2) holds by anti-concentration Corollary 2.

To show inequality (E.4) note that by C.2(b) and Lemma 11 we have $\text{wp} \rightarrow 1$

$$\kappa_{n,\mathcal{V}}(\gamma_n + o(1/\ell_n)) + o(\delta_n) \leq \bar{a}_n + \frac{c(\gamma_n + o(1/\ell_n))}{\bar{a}_n} + o(\delta_n) \leq \bar{a}_n + \frac{\eta\ell\ell_n + \eta\log 10}{\bar{a}_n} + o(\delta_n),$$

where the last inequality relies on

$$c(\gamma_n + o(1/\ell_n)) \leq -\eta \log((1 - \gamma_n - o(1/\ell_n))) = \eta o(\ell\ell_n) + \eta \log 10,$$

holding for large n by C.3. From this we deduce (E.4).

Step 2. (Lower Containment) We have that for all $v \in V_n$,

$$\begin{aligned} A_n(v) &:= \hat{\theta}_n(v) - \theta_{n0} - \inf_{v \in \mathcal{V}} (\hat{\theta}_n(v) + k_{n,\mathcal{V}} s_n(v)) \\ &\leq -Z_n(v) \sigma_n(v) + \kappa_n \sigma_n(v) + \sup_{v \in \mathcal{V}} \{\theta_{n0} - \hat{\theta}_n(v) - k_{n,\mathcal{V}}(\gamma_n) s_n(v)\} := B_n(v) \end{aligned}$$

since $\theta_n(v) \leq \theta_{n0} + \kappa_n \sigma_n(v)$, $\forall v \in V_n$ and $\hat{\theta}_n(v) - \theta_n(v) = -Z_n(v) \sigma_n(v)$. Therefore,

$$\begin{aligned} \mathbb{P}_n\{V_n \subseteq \hat{V}_n\} &= \mathbb{P}_n\{A_n(v) \leq 2k_{n,\mathcal{V}}(\gamma_n) s_n(v), \forall v \in V_n\} \\ &\geq \mathbb{P}_n\{B_n(v) \leq 2k_{n,\mathcal{V}}(\gamma_n) s_n(v), \forall v \in V_n\} \\ &\geq \mathbb{P}_n\{-Z_n(v) \sigma_n(v) \leq 2k_{n,\mathcal{V}}(\gamma_n) s_n(v) - \kappa_n \sigma_n(v), \forall v \in V_n\} \\ &= \mathbb{P}_n\left\{\sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \geq k_{n,\mathcal{V}}(\gamma_n)\right\} \\ &:= a - b = \gamma'_n - o(1) = 1 - o(1), \end{aligned}$$

where $b = o(1)$ follows similarly to the proof of Theorems 1 (analytical case) and Theorem 2 (simulation case), using the observation that $k_{n,\mathcal{V}}(\gamma_n) \geq k_{n,V_n}(\gamma_n)$ for sufficiently large n by C.3, and $a = o(1)$ follows from the following argument:

$$\begin{aligned} a &\geq_{(1)} \mathbb{P}_n \left(\sup_{v \in V_n} -Z_n(v) \leq 2k_{n,\mathcal{V}}(\gamma_n) [1 - o_{\mathbb{P}_n}(\delta_n/(\bar{a}_n + \ell\ell_n))] - \kappa_n \right) \\ &\geq_{(2)} \mathbb{P}_n \left(\sup_{v \in V_n} -Z_n^*(v) \leq 2k_{n,\mathcal{V}}(\gamma_n) - \kappa_n - o_{\mathbb{P}_n}(\delta_n) \right) - o(1) \\ &\geq_{(3)} \mathbb{P}_n \left(\sup_{v \in V_n} -Z_n^*(v) \leq \kappa_n - o_{\mathbb{P}_n}(\delta_n) \right) - o(1) \\ &\geq_{(4)} \mathbb{P}_n \left(\sup_{v \in V_n} -Z_n^*(v) \leq \kappa_{n,V_n}(\gamma'_n) - o(\delta_n) \right) - o(1) \\ &\geq_{(5)} \gamma'_n - o(1) = 1 - o(1), \end{aligned}$$

where terms $o(\delta_n)$ are different in different places; where (1) follows by C.4, (2) is by C.2 and by Step 1, namely by $k_{n,\mathcal{V}}(\gamma'_n) \leq \bar{\kappa}_n \lesssim \bar{a}_n + \ell\ell_n$ wp $\rightarrow 1$, (3) follows by Step 1, (4) follows by monotonicity of $\mathcal{V} \mapsto \kappa_{n,\mathcal{V}}(\gamma'_n)$ and $V_n \subseteq \mathcal{V}$, (5) follows by the anti-concentration Corollary 2.

Step 3. (Upper Containment). We have that for all $v \notin \bar{V}_n$,

$$\begin{aligned} A_n(v) &:= \hat{\theta}_n(v) - \theta_{n0} - \inf_{v \in \mathcal{V}} \left(\hat{\theta}_n(v) + k_{n,\mathcal{V}} s_n(v) \right) \\ &> -Z_n(v) \sigma_n(v) + \bar{\kappa}_n \sigma_n(v) + \sup_{v \in \mathcal{V}} \{ \theta_{n0} - \hat{\theta}_n(v) - k_{n,\mathcal{V}}(\gamma_n) s_n(v) \} := C_n(v), \end{aligned}$$

since $\theta_n(v) > \theta_{n0} + \bar{\kappa}_n \sigma_n(v)$, $\forall v \notin \bar{V}_n$, $\hat{\theta}_n(v) - \theta_n(v) = -Z_n(v) \sigma_n(v)$. Hence

$$\begin{aligned} \mathbb{P}_n \left(\hat{V}_n \not\subseteq \bar{V}_n \right) &= \mathbb{P}_n \{ A_n(v) \leq 2k_{n,\mathcal{V}}(\gamma_n) s_n(v), \exists v \notin \bar{V}_n \} \\ &\leq \mathbb{P}_n \{ C_n(v) \leq 2k_{n,\mathcal{V}}(\gamma_n) s_n(v), \exists v \notin \bar{V}_n \} \\ &\leq \mathbb{P}_n \{ -Z_n(v) \sigma_n(v) < 2k_{n,\mathcal{V}}(\gamma_n) s_n(v) - \bar{\kappa}_n \sigma_n(v), \exists v \notin \bar{V}_n \} \\ &+ \mathbb{P}_n \left\{ \sup_{v \in \mathcal{V}} \frac{\theta_{n0} - \hat{\theta}_n(v)}{s_n(v)} \geq k_{n,\mathcal{V}}(\gamma_n), \exists v \notin \bar{V}_n \right\} \\ &=: c - b \leq (1 - \gamma'_n) + o(1) = o(1), \end{aligned}$$

where $b = o(1)$ from the Step 2, and $c \leq (1 - \gamma'_n) + o(1)$ follows from the following:

$$\begin{aligned} c &\leq_{(1)} \mathbb{P}_n \left(-Z_n(v) < 2k_{n,\mathcal{V}}(\gamma_n) [1 + o_{\mathbb{P}_n}(\delta_n/(\bar{a}_n + \ell\ell_n))] - \bar{\kappa}_n, \exists v \in \mathcal{V} \right) \\ &\leq_{(2)} \mathbb{P}_n \left(-Z_n^*(v) < 2k_{n,\mathcal{V}}(\gamma_n) - \bar{\kappa}_n + o(\delta_n), \exists v \in \mathcal{V} \right) + o(1) \\ &\leq_{(3)} \mathbb{P}_n \left(\sup_{v \in \mathcal{V}} Z_n^*(v) > k_{n,\mathcal{V}}(\gamma_n) - o(\delta_n) \right) + o(1) \\ &\leq_{(4)} \mathbb{P}_n \left(\sup_{v \in \mathcal{V}} Z_n^*(v) > \kappa_n - o(\delta_n) \right) + o(1) \leq_{(5)} (1 - \gamma'_n) + o(1), \end{aligned}$$

where (1) follows by C.4, (2) follows by C.2 and Step 1, namely by $k_{n,\mathcal{V}}(\gamma'_n) \leq \bar{\kappa}_n \lesssim \bar{a}_n + \ell\ell_n$ wp $\rightarrow 1$, (3) follows by Step 1, (4) holds by Step 1, (5) holds by the definition of κ_n and the anti-concentration Corollary 2.

Step 4. (Rate). We have that wp $\rightarrow 1$

$$d_H(\hat{V}_n, V_0) \leq_{(1)} d_H(\hat{V}_n, V_n) + d_H(V_n, V_0) \leq_{(2)} 2d_H(\bar{V}_n, V_0) \leq_{(3)} (\bar{\sigma}_n \bar{\kappa}_n)^{1/\rho_n} / c_n$$

where (1) holds by the triangle inequality, (2) follows by the containment $V_0 \subseteq V_n \subseteq \hat{V}_n \subseteq \bar{V}_n$ holding wp $\rightarrow 1$, and (3) follows from $\bar{\kappa}_n \bar{\sigma}_n \rightarrow 0$ holding by assumption, and from the

following relation holding by Condition V:

$$\begin{aligned}
d_H(\bar{V}_n, V_0) &= \sup_{v \in \bar{V}_n} d(v, V_0) \leq \sup\{d(v, V_0) : \theta_n(v) - \theta_{n0} \leq \bar{\kappa}_n \sigma_n(v)\} \\
&\leq \sup\{d(v, V_0) : (c_n d(v, V_0))^{\rho_n} \wedge \delta \leq \bar{\kappa}_n \bar{\sigma}_n\} \\
&\leq \sup\{t : (c_n t)^{\rho_n} \wedge \delta \leq \bar{\kappa}_n \bar{\sigma}_n\} \leq c_n^{-1} (\bar{\kappa}_n \bar{\sigma}_n)^{1/\rho_n} \text{ for all } 0 \leq \bar{\kappa}_n \bar{\sigma}_n \leq \delta.
\end{aligned}$$

■

E.2. Proof of Lemma 4. Step 1. Verification of C.1. This condition holds by inspection in view of continuity of $v \mapsto p_n(v, \beta_n)$ and $v \mapsto p_n(v, \hat{\beta})$ implied by Condition P(ii) and by Ω_n and $\hat{\Omega}_n$ being positive definite.

Step 2. Verification of C.2. Part (a). By Condition P, uniformly in $v \in \mathcal{V}$,

$$\begin{aligned}
Z_n(v) &= \frac{p_n(v, \beta_n^*(v))'}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \sqrt{n}(\hat{\beta}_n - \beta_n) \\
&= \frac{p_n(v, \beta_n)'}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \sqrt{n}(\hat{\beta}_n - \beta_n) + \frac{L_n \sqrt{n} \|\hat{\beta}_n - \beta_n\|^2}{\min_{v \in \mathcal{V}} \|p_n(v, \beta_n)\|} \frac{\lambda_{\max}(\Omega_n^{1/2})}{\lambda_{\min}(\Omega_n^{1/2})} \\
&= \frac{p_n(v, \beta_n)' \Omega_n^{1/2}}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \mathcal{N}_k + o_{P_n}(\delta'_n) + O_{P_n}(n^{-1/2}).
\end{aligned}$$

Part (b). First note using the inequality

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \leq \left(2 \frac{\|a - b\|}{\|a\|} \right) \wedge \left(2 \frac{\|a - b\|}{\|b\|} \right), \quad (\text{E.6})$$

we have

$$\begin{aligned}
M_n &= \left\| \frac{p_n(v, \beta_n)' \Omega_n^{1/2}}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} - \frac{p_n(v, \hat{\beta}_n)' \hat{\Omega}_n^{1/2}}{\|p_n(v, \hat{\beta}_n)' \hat{\Omega}_n^{1/2}\|} \right\| \leq 2 \frac{\|p_n(v, \beta_n)' \Omega_n^{1/2} - p_n(v, \hat{\beta}_n)' \hat{\Omega}_n^{1/2}\|}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \\
&\leq 2 \frac{\|p_n(v, \beta_n)' \Omega_n^{1/2} (I - \Omega_n^{-1/2} \hat{\Omega}_n^{1/2})\|}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} + 2 \frac{L_n \|\hat{\beta}_n - \beta_n\|}{\min_{v \in \mathcal{V}} \|p_n(v, \beta_n)\|} \frac{\lambda_{\max}(\Omega_n^{1/2})}{\lambda_{\min}(\Omega_n^{1/2})} \\
&\leq 2 \|\Omega_n^{-1/2}\| \|\hat{\Omega}_n^{1/2} - \Omega_n^{1/2}\| + O_{P_n}(n^{-1/2}) \leq O_{P_n}(n^{-b}) + O_{P_n}(n^{-1/2}) = O_{P_n}(n^{-b}),
\end{aligned}$$

for some $b > 0$. We have that

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} |Z_n^*(v) - Z_n^{\star}(v)| \mid \mathcal{D}_n \right) \leq M_n E_{P_n} \|\mathcal{N}_k\| \leq M_n \sqrt{k}.$$

Hence for any $\delta_n'' \propto n^{-b'}$ with a constant $0 < b' < b$, we have by Markov's Inequality that

$$\mathbb{P}_n \left(\sup_{v \in \mathcal{V}} |Z_n^*(v) - Z_n^*(\tilde{v})| > \delta_n \ell_n \mid \mathcal{D}_n \right) \leq \frac{O_{\mathbb{P}_n}(n^{-b})}{\delta_n'' \ell_n} = o_{\mathbb{P}_n}(1/\ell_n).$$

Now select $\delta_n = \delta_n' \vee \delta_n''$.

Step 3. Verification of C.3. We shall employ Lemma 12, which has the required notation in place. We only need to compute an upper bound on the covering numbers $N(\varepsilon, \mathbf{V}, \rho)$ for the process Z_n . We have that

$$\begin{aligned} \sigma(Z_n^*(v) - Z_n^*(\tilde{v})) &\leq \left\| \frac{p_n(v, \beta_n)' \Omega_n^{1/2}}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} - \frac{p_n(\tilde{v}, \beta_n)' \Omega_n^{1/2}}{\|p_n(\tilde{v}, \beta_n)' \Omega_n^{1/2}\|} \right\| \\ &\leq 2 \left\| \frac{(p_n(v, \beta_n) - p_n(\tilde{v}, \beta_n))' \Omega_n^{1/2}}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \right\| \leq 2 \frac{L_n}{\min_{v \in \mathcal{V}} \|p_n(v, \beta_n)\|} \frac{\lambda_{\max}(\Omega_n^{1/2})}{\lambda_{\min}(\Omega_n^{1/2})} \|v - \tilde{v}\| \leq CL \|v - \tilde{v}\|, \end{aligned}$$

where C is some constant that does not depend on n , by the eigenvalues of Ω_n bounded away from zero and from above. Hence by the standard volumetric argument

$$N(\varepsilon, \mathbf{V}, \rho) \leq \left(\frac{1 + CL \text{diam}(\mathbf{V})}{\varepsilon} \right)^d, \quad 0 < \varepsilon < 1,$$

where the diameter of \mathbf{V} is measured by the Euclidian metric. Condition C.3 now follows by Lemma 12, with $a_n(\mathbf{V}) = (2\sqrt{\log L_n(\mathbf{V})}) \vee (1 + \sqrt{d})$, $L_n(\mathbf{V}) = C' (1 + CL \text{diam}(\mathbf{V}))^d$, where C' is a constant from Lemma 12.

Step 4. Verification of C.4. Under Condition P, we have that $1 \leq a_n(\mathbf{V}) \leq \bar{a}_n := a_n(\mathcal{V}) \lesssim 1$, so that C.4(a) follows since by Condition P

$$\bar{\sigma}_n = \sqrt{\max_{v \in \mathcal{V}} \|p_n(v, \beta_n) \Omega_n^{1/2}\|/n} \leq \sqrt{\max_{v \in \mathcal{V}} \|p_n(v, \beta_n)\| \|\Omega_n^{1/2}\|/n} \lesssim \sqrt{1/n}$$

To verify C.4(b) note that uniformly in $v \in \mathcal{V}$,

$$\begin{aligned} \left| \frac{\|p_n(v, \beta_n)' \hat{\Omega}_n^{1/2}\|}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} - 1 \right| &\leq \left| \frac{\|p_n(v, \beta_n)' \hat{\Omega}_n^{1/2}\| - \|p_n(v, \beta_n)' \Omega_n^{1/2}\|}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \right| \\ &\leq \frac{\|p_n(v, \beta_n)' (\hat{\Omega}_n^{1/2} - \Omega_n^{1/2})\|}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \leq \frac{\|p_n(v, \beta_n)' \Omega_n^{1/2} (\hat{\Omega}_n^{1/2} \Omega_n^{-1/2} - I)\|}{\|p_n(v, \beta_n)' \Omega_n^{1/2}\|} \\ &\leq \|\hat{\Omega}_n^{1/2} \Omega_n^{-1/2} - I\| \leq \|\Omega_n^{-1/2}\| \|\hat{\Omega}_n^{1/2} - \Omega_n^{1/2}\| = o_{\mathbb{P}_n}(\delta_n), \end{aligned}$$

since $\|\hat{\Omega}_n^{1/2} - \Omega_n^{1/2}\| = O_{\mathbb{P}_n}(n^{-b})$ for some $b > 0$, and since $\|\Omega_n^{-1/2}\|$ is uniformly bounded, both implied by the assumptions.

Step 5. Verification of S. Then, since under Condition V with for large enough n , $r_n \lesssim c_n^{-1}(1/\sqrt{n})^{\ell_n/\rho_n} = o(1)$, we have that $r_n \leq \varphi_n$ for large n for some $\varphi_n = o(1)$. Condition

It then follows by noting that for any positive $o(1)$ term, $\sup_{\|v-\tilde{v}\|\leq o(1)} |Z_n(v) - Z_n(\tilde{v})| \leq \Upsilon o(1) \|\mathcal{N}_k\| = o_{P_n}(1)$. \blacksquare

APPENDIX F. KERNEL-TYPE ESTIMATION OF BOUNDING FUNCTION FROM CONDITIONAL MOMENT INEQUALITIES

In this section we provide primitive conditions that justify application of kernel-type estimation methods covered in Section 4.3 for models characterized by conditional moment inequalities.

Example 7 (Bounding Function from Conditional Moment Inequalities). Suppose that we have an i.i.d. sample of $(X_i, Z_i), i = 1, \dots, n$ defined on the probability space (A, \mathcal{A}, P) , where we take P fixed in this example. Suppose that $\text{support}(Z_i) = \mathcal{Z} \subseteq [0, 1]^d$, and

$$\theta_{n0} = \min_{v \in \mathcal{V}} \theta_n(v),$$

for $\theta_n(v) = E_P[m(X_i, \mu, j) | Z_i = z]$, $v = (z, j)$, where $\mathcal{V} \subseteq \mathcal{Z} \times \{1, \dots, J\}$ be the set of interest. Suppose the first J_0 functions correspond to equalities treated as inequalities, so that $m(X_i, \mu, j) = -m(X_i, \mu, j+1)$, for $j \in \mathcal{J}_0 = \{1, 3, \dots, J_0 - 1\}$. Hence $\theta_n(z, j) = -\theta_n(z, j+1)$ for $j \in \mathcal{J}_0$, and we only need to estimate functions $\theta_n(z, j)$ with the index $j \in \mathcal{J} := \mathcal{J}_0 \cup \{J_0 + 1, J_0 + 2, \dots, J\}$. Suppose we use the local polynomial approach to approximating and estimating $\theta_n(z, j)$. For $u \equiv (u_1, \dots, u_d)$, a d -dimensional vector of nonnegative integers, let $[u] = u_1 + \dots + u_d$. Let A_p be the set of all d -dimensional vectors u such that $[u] \leq p$ for some integer $p \geq 0$ and let $|A_p|$ denote the number of elements in A_p . For $z \in \mathbb{R}^d$ with $u \in A_p$, let $z^u = \prod_{i=1}^d z_i^{u_i}$. Now define

$$\mathbf{p}(b, z) = \sum_{u \in A_p} b_u z^u, \tag{F.1}$$

where $b = (b_u)_{u \in A_p}$ is a vector of dimension $|A_p|$. For each $v = (z, j)$ and $Y_i(j) := m(X_i, \mu, j)$, define

$$S_n(b) := \sum_{i=1}^n \left[Y_i(j) - \mathbf{p}\left(b, \frac{Z_i - z}{h_n}\right) \right]^2 K_{h_n}(Z_i - z),$$

where $K_h(u) := K(u/h)$, $K(\cdot)$ is a d -dimensional kernel function and h_n is a sequence of bandwidths. The local polynomial estimator $\hat{\theta}_n(v)$ of the regression function is the first element of $\hat{b}(z, j) := \arg \min_{b \in \mathbb{R}^{|A_p|}} S_n(b)$.

We impose the following conditions:

(i) for each $j \in \mathcal{J}$, $\theta(z, j)$ is $(p+1)$ times continuously differentiable with respect to $z \in \mathcal{Z}$, where \mathcal{Z} is convex. (ii) the probability density function f of Z_i is bounded above and bounded below from zero with continuous derivatives on \mathcal{Z} ; (iii) for $Y_i(j) := m(X_i, \mu, j)$, $Y_i := (Y_i(j), j \in \mathcal{J})'$, and $U_i := Y_i - E_P[Y_i|Z_i]$; and U_i is a bounded random vector; (iv) for each j , the conditional on Z_i density of U_i exists and is uniformly bounded from above and below, or, more generally, condition R stated in Appendix G holds; (v) $K(\cdot)$ has support on $[-1, 1]^d$, is twice continuously differentiable, $\int uK(u)du = 0$, and $\int K(u)du = 1$; (vi) $h_n \rightarrow 0$, $nh_n^{d+|\mathcal{J}|+1} \rightarrow \infty$, $nh_n^{d+2(p+1)} \rightarrow 0$, $\sqrt{n^{-1}h_n^{-2d}} \rightarrow 0$ at polynomial rates in n .

These conditions are imposed to verify Assumptions A1-A7 in Kong, Linton, and Xia (2010). Details of verification are given in Supplementary Appendix J. Note that $p > |\mathcal{J}|/2 - 1$ is necessary to satisfy bandwidth conditions in (vi). Conditions (i)-(vi) above are sufficient conditions to check Assumptions A1-A7 in Kong, Linton, and Xia (2010). The assumption that U_i is bounded is technical and is made to simplify exposition and proofs.

Let $\delta_n = 1/\log n$. Then it follows from Corollary 1 and Lemmas 8 and 10 of Kong, Linton, and Xia (2010) that

$$\hat{\theta}_n(z, j) - \theta(z, j) = \frac{1}{nh_n^d f(z)} \mathbf{e}_1' S_p^{-1} \sum_{i=1}^n (e_j' U_i) K_h(Z_i - z) \mathbf{u}_p \left(\frac{Z_i - z}{h_n} \right) + B_n(z, j) + R_n(z, j), \quad (\text{F.2})$$

where \mathbf{e}_1 is an $|A_p| \times 1$ vector whose first element is one and all others are zeros, S_p is an $|A_p| \times |A_p|$ matrix such that $S_p = \{\int z^u (z^v)' du : u \in A_p, v \in A_p\}$, $\mathbf{u}_p(z)$ is an $|A_p| \times 1$ vector such that $\mathbf{u}_p(z) = \{z^u : u \in A_p\}$,

$$B_n(z, j) = O(h_n^{p+1}) \text{ and } R_n(z, j) = o_p \left(\frac{\delta_n}{(nh_n^d)^{1/2}} \right),$$

uniformly in $(z, j) \in \mathcal{Z} \times \{1, \dots, J\}$. The exact form of $B_n(z, j)$ is given in equation (12) of Kong, Linton, and Xia (2010). The result that $B_n(z, j) = O(h_n^{p+1})$ uniformly in (z, j) follows from the standard argument based on Taylor expansion given in Fan and Gijbels (1996), Kong, Linton, and Xia (2010), or Masry (1996). The condition that $nh_n^{d+2(p+1)} \rightarrow 0$ at a polynomial rate in n corresponds to the undersmoothing condition.

Now set $\mathbf{K}(z/h) \equiv \mathbf{e}_1' S_p^{-1} K_h(z) \mathbf{u}_p(z/h)$, which is a kernel of order $(p+1)$ (See Section 3.2.2 of Fan and Gijbels (1996)). Let

$$g_v(U, Z) := \frac{e_j' U}{(h_n^d)^{1/2} f(z)} \mathbf{K} \left(\frac{Z - z}{h_n} \right).$$

Then it follows from (J.1) that uniformly in $v \in \mathcal{V}$

$$(nh_n^d)^{1/2}(\hat{\theta}_n(z, j) - \theta_n(z, j)) = \mathbb{G}_n(g_v) + o_P(\delta_n).$$

Application of Theorems 8 and 9 in Appendix G, based on the Rio-Massart coupling, verifies condition NK.1 (a) and NK.1 (b). Finally, Condition NK.2 holds if we take $\hat{f}_n(z)$ to be the standard kernel density estimator with kernel K and let $e'_j \hat{U}_i = Y_i(j) - \hat{\theta}_n(z, j)$. ■

APPENDIX G. STRONG APPROXIMATION FOR KERNEL-TYPE METHODS

To establish our strong approximation for kernel-type estimators we use Theorem 1.1 in Rio (1994), stated below, which builds on the earlier results of Massart (1989). After the statement of the Rio-Massart coupling we provide our strong approximation result, which generalizes the previous results to kernel-type estimators for regression models with multivariate outcomes. We then provide a novel multiplier method to approximate the distribution of such estimators. The proofs for these results are provided in Appendix H.

G.1. Rio-Massart Coupling. Consider a sufficiently rich probability space (A, \mathcal{A}, P) . If not, then we can always enrich the original space by taking the product with $[0, 1]$ equipped with the uniform measure over Borel sets of $[0, 1]$. Consider a suitably measurable, namely image admissible Suslin, function class \mathcal{F} containing functions $f : I^d \rightarrow I$ for $I = (-1, 1)$. A function class \mathcal{F} is of uniformly bounded variation of at most $K(\mathcal{F})$ if

$$TV(\mathcal{F}) := \sup_{f \in \mathcal{F}} \sup_{g \in \mathcal{D}_c(I^d)} \left(\int_{\mathbb{R}^d} f(x) \operatorname{div} g(x) / \|g\|_\infty dx \right) \leq K(\mathcal{F}),$$

where $\mathcal{D}_c(I^d)$ is the space of C^∞ functions taking values in \mathbb{R}^d with compact support included in I^d , and where $\operatorname{div} g(x)$ is the divergence of $g(x)$. Suppose the function class \mathcal{F} obeys the following uniform L_1 covering condition

$$\sup_Q N(\epsilon, \mathcal{F}, L_1(Q)) \leq C(\mathcal{F}) \epsilon^{d(\mathcal{F})},$$

where sup is taken over probability measures with finite support, and $N(\epsilon, \mathcal{F}, L_1(Q))$ is the covering number under the $L_1(Q)$ norm on \mathcal{F} . Let X_1, \dots, X_n be an i.i.d. sample on the probability space (A, \mathcal{A}, P) from density f_X with support on I^d , bounded from above and away from zero. Let P_X be the measure induced by f_X . Then there exists a P_X -Brownian Bridge \mathbb{B}_n with a.s. continuous paths with respect to the $L^1(P_X)$ metric such that for any positive $t \geq C \log n$,

$$P \left(\sqrt{n} \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f) - \mathbb{B}_n(f)| \geq C \sqrt{tn^{\frac{d-1}{d}} K(\mathcal{F})} + Ct \sqrt{\log n} \right) \leq e^{-t},$$

where constant C depends only on d , $C(\mathcal{F})$, and $d(\mathcal{F})$.

G.2. Strong Approximation for Kernel-Type Estimators. We shall use the following technical condition in what follows.

Condition R. *The random $(J + d)$ -vector (U_i, Z_i) obeys $U_i = (U_{i,1}, \dots, U_{i,J}) = \varphi_n(X_{i,1})$, and $Z_i = \tilde{\varphi}_n(X_{2i})$, where $X_i = (X'_{1i}, X'_{2i})'$ is a $(d_1 + d)$ -vector with $1 \leq d_1 \leq J$, which has density bounded away from zero by \underline{f} and above by \bar{f} on the support I^{d_1+d} , where $\varphi_n : I^{d_1} \mapsto I^J$ and $\sum_{l=1}^{d_1} \int_{I^{d_1}} |D_{x_{1l}} \varphi_n(x_1)| dx_1 \leq B$, where $D_{x_{1l}} \varphi_n(x_1)$ denotes the weak derivative with respect to the l -th component of x_1 , and $\tilde{\varphi}_n : I^d \mapsto I^d$ is continuously differentiable such that $\max_{k \leq d} \sup_{x_2} |\partial \tilde{\varphi}_n(x_2) / \partial x_{2k}| \leq B$ and $|\det \partial \tilde{\varphi}_n(x_2) / \partial x_2| \geq c > 0$, where $\partial \tilde{\varphi}_n(x_2) / \partial x_{2k}$ denotes the partial derivative with respect to the k -th component of x_2 . The constants $J, B, \underline{f}, \bar{f}, c$ and vector dimensions do not depend on n . ($|\cdot|$ denotes ℓ_1 norm.)*

A simple example of (U_i, Z_i) satisfying this condition is given in Corollary 3 below.

Theorem 8 (Strong Approximation for Kernel-Type Estimators). *Consider a suitably enriched probability space (A, \mathcal{A}, P_n) for each n . Let $n \rightarrow \infty$. Assume the following conditions hold for each n : (a) There are n i.i.d. $(J + d)$ -dimensional random vectors of the form (U_i, Z_i) that obey Condition R, and the density f_n of Z is bounded from above and away from zero on the set \mathcal{Z} , uniformly in n . (b) Let $v = (z, j)$ and $\mathcal{V} = \mathcal{Z} \times \{1, \dots, J\}$, where $\mathcal{Z} \subseteq I^d$. The kernel estimator $v \mapsto \hat{\theta}_n(v)$ of some target function $v \mapsto \theta_n(v)$ has an asymptotic linear expansion uniformly in $v \in \mathcal{V}$*

$$(nh_n^d)^{1/2}(\hat{\theta}_n(v) - \theta_n(v)) = \mathbb{G}_n(g_v) + o_{P_n}(\delta_n), \quad g_v(U_i, Z_i) := \frac{1}{(h_n^d)^{1/2} f_n(z)} e'_j U_i \mathbf{K} \left(\frac{z - Z_i}{h_n} \right),$$

where $e'_j U_i \equiv U_{ij}$, \mathbf{K} is twice continuously differentiable product kernel function with support on I^d , $\int \mathbf{K}(u) du = 1$, and h_n is a sequence of bandwidths that converges to zero, (c) for a given $\delta_n \searrow 0$, the bandwidth sequence obeys: $(n^{-1/(d+d_1)} h_n^{-1} \log n)^{1/2} + (nh_n^d)^{-1/2} \log^{3/2} n = o(\delta_n)$. Then there exists a sequence of centered P_n -Gaussian Bridges \mathbb{B}_n such that

$$\sup_{v \in \mathcal{V}} |(nh_n^d)^{1/2}(\hat{\theta}_n(v) - \theta_n(v)) - \mathbb{B}_n(g_v)| = o_{P_n}(\delta_n).$$

Moreover, the paths of $v \mapsto \mathbb{B}_n(g_v)$ can be chosen to be continuous a.s.

Remark 8. Conditions (a) and (b) cover standard conditions in the literature, imposing a uniform Bahadur expansion for kernel-type estimators, which have been shown in Masry (1996) and Kong, Linton, and Xia (2010) for kernel mean regression estimators and also local polynomial estimators under fairly general conditions. Implicit in the expansion above

is that the asymptotic bias is negligible, which can be achieved by undersmoothing, i.e. choosing the bandwidth to be smaller than the rate-optimal bandwidth.

Corollary 3 (A Simple Leading Case for Moment Inequalities Application). *Suppose that (U_i, Z_i) has bounded support, which we then take to be a subset of I^{J+d} without loss of generality. Suppose that $U_i = (U_{ij}, j = 1, \dots, J)$ where for the first $J_0/2$ pairs of terms, we have $U_{ij} = -U_{ij+1}, j = 1, 3, \dots, J_0 - 1$. Let $\mathcal{J} = \{1, 3, \dots, J_0 - 1, J_0 + 1, J_0 + 2, \dots\}$. Suppose that $(U_{ij}, Z_i, j \in \mathcal{J})$ have joint density bounded from above and below by some constants \bar{f} and \underline{f} . Suppose these constants and d, J , and $d_1 = |\mathcal{J}|$ do not depend on n . Then Condition R holds, and the conclusions of Theorem 8 then hold under the additional conditions imposed in the theorem.*

Note that Condition R allows for much more general error terms and regressors. For example, it allows error terms U_i not to have a density at all, and Z_i need only have density bounded from above.

The next theorem shows that the Brownian bridge $\mathbb{B}_n(g_v)$ can be approximately simulated via the Gaussian multiplier method. That is, consider the following symmetrized process

$$\mathbb{G}_n^o(g_v) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i g_v(U_i, Z_i) = \mathbb{G}_n(\xi g_v), \quad (\text{G.1})$$

where ξ_1, \dots, ξ_n are i.i.d $N(0, 1)$, independent of the data \mathcal{D}_n and of $\{(U_i, Z_i)\}_{i=1}^n$, which are i.i.d. copies of (U, Z) . Conditional on the data this is a Gaussian process with a covariance function which is a consistent estimate of the covariance function of $v \mapsto \mathbb{B}_n(g_v)$. The theorem below shows that the uniform distance between $\mathbb{B}_n(g_v)$ and $\mathbb{G}_n^o(g_v)$ is small with an explicit probability bound. Note that if the function class $\{g_v, v \in \mathcal{V}\}$ were Donsker, then such a result would follow from the multiplier functional central limit theorem. In our case, this function class is not Donsker, so we require a different argument.

For the following theorem, consider now a sufficiently rich probability space (A, \mathcal{A}, P_n) . Note that we can always enrich the original space if needed by taking the product with $[0, 1]$ equipped with the uniform measure over Borel sets of $[0, 1]$.

Theorem 9 (Multiplier Method for Kernels). *Let $v = (z, j)$ and $\mathcal{V} \subseteq \mathcal{Z} \times \{1, \dots, J\}$, where \mathcal{Z} is a compact convex set that does not depend on n . The estimator $v \mapsto \hat{\theta}_n(v)$ and the function $v \mapsto \theta_n(v)$ are continuous in v . In what follows, let e_j denote the J -vector with j th element one and all other elements zero. Suppose that (U, Z) is a $(J + d)$ -dimensional random vector, where U is a generalized residual such that $E[U|Z] = 0$ a.s. and Z is a covariate; the density f_n of Z is continuous and bounded away from zero and from above*

on \mathcal{Z} , uniformly in n ; and the support of U is bounded uniformly in n . \mathbf{K} is a twice continuously differentiable, possibly higher-order, product kernel function with support on $[-1, 1]^d$, $\int \mathbf{K}(u) du = 1$; and h_n is a sequence of bandwidths such that $h_n \rightarrow 0$ and $nh_n^d \rightarrow \infty$ such that $\sqrt{n^{-1}h_n^{-2d}} = o((\delta_n/[\ell_n\sqrt{\log n}])^{d+1})$. Let $\{(U_i, Z_i)\}_{i=1}^n$ be i.i.d. copies of (U, Z) , and $\{Z_i\}_{i=1}^n$ be a part of the data \mathcal{D}_n . Let \mathbb{B}_n denote the P_n -Brownian bridge, and

$$g_v(U, Z) := \frac{e'_j U}{(h_n^d)^{1/2} f_n(z)} \mathbf{K}\left(\frac{z - Z}{h_n}\right).$$

Then there exists an independent, identically distributed copy $v \mapsto \bar{\mathbb{B}}_n(g_v)$ of the process $v \mapsto \mathbb{B}_n(g_v)$, such that for some $o(\delta_n)$ and $o(1/\ell_n)$ sequences,

$$P_n \left(\sup_{v \in \mathcal{V}} |\mathbb{G}_n^o(g_v) - \bar{\mathbb{B}}_n(g_v)| > o(\delta_n) \middle| \mathcal{D}_n \right) = o_{P_n}(1/\ell_n). \quad (\text{G.2})$$

APPENDIX H. PROOFS FOR NONPARAMETRIC ESTIMATION OF $\theta(v)$ VIA KERNEL-TYPE METHODS

H.1. Proof of Lemma 7. There are six steps, with the first four verifying conditions C.1-C.4, and the last two providing auxiliary calculations. Let $U_{ij} \equiv e'_j U_i$.

Step 1. Verification of C.1 and C.2. Condition C.1 holds by inspection, in view of continuity of $v \mapsto \hat{\theta}_n(v)$, $v \mapsto \theta_n(v)$, $v \mapsto \sigma_n(v)$, and $v \mapsto s_n(v)$. Condition C.2 is assumed directly.

Step 2. Verification of C.3. Note that

$$\frac{g_v(U_i, Z_i)}{\sigma_n(v) \sqrt{nh_n^d}} = \frac{\frac{1}{h_n^{d/2}} \mathbf{K}\left(\frac{z - Z_i}{h_n}\right) U_{ij}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K}\left(\frac{z - Z_i}{h_n}\right) U_{ij} \right\|_{P_{n,2}}}$$

We shall employ Lemma 12, which has the required notation in place. We only need to compute an upper bound on the covering numbers $N(\varepsilon, \mathbf{V}, \rho)$ of \mathbf{V} under the metric $\rho(v, \bar{v}) = \sigma(Z_n^*(v) - Z_n^*(\bar{v}))$. We have that for $v = (z, j)$ and $\bar{v} = (\bar{z}, j)$

$$\sigma(Z_n^*(v) - Z_n^*(\bar{v})) \leq \Upsilon_n \|z - \bar{z}\|,$$

$$\Upsilon_n := \sup_{v \in \mathbf{V}, 1 \leq k \leq d} \left\| \nabla_{z_k} \frac{\frac{1}{h_n^{d/2}} \mathbf{K}\left(\frac{z - Z_i}{h_n}\right) U_{ij}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K}\left(\frac{z - Z_i}{h_n}\right) U_{ij} \right\|_{P_{n,2}}} \right\|_{P_{n,2}}.$$

We have that

$$\Upsilon_n \leq \sup_{v \in \mathcal{V}, 1 \leq k \leq d} \frac{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} + \frac{\left| \nabla_{z_k} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}} \right|}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}},$$

which is bounded by $C(1 + h_n^{-1})$ for large n by Step 6. Since J is finite, it follows that for all large $n > n_0$ for all non-empty subsets of $\mathcal{V} \subseteq \mathcal{V}$,

$$N(\varepsilon, \mathcal{V}, \rho) \leq \left(\frac{J^{1/d} (1 + C(1 + h_n^{-1}) \text{diam}(\mathcal{V}))}{\varepsilon} \right)^d, \quad 0 < \varepsilon < 1.$$

Condition C.3 now follows for all $n > n_0$ by Lemma 12, with

$$a_n(\mathcal{V}) = (2\sqrt{\log L_n(\mathcal{V})}) \vee (1 + \sqrt{d}), \quad L_n(\mathcal{V}) = C' (1 + C(1 + h_n^{-1}) \text{diam}(\mathcal{V}))^d,$$

where C' is a constant from Lemma 12.

Step 3. Verification of C.4. Under Condition NK, we have that

$$a_n(\mathcal{V}) \leq \bar{a}_n := a_n(\mathcal{V}) \lesssim \sqrt{\log \ell_n + \log n} \lesssim \sqrt{\log n},$$

so that C.4(a) follows if $\sqrt{\log n / (nh_n^d)} \rightarrow 0$.

To verify C.4(b) note that

$$\left| \frac{s_n(v)}{\sigma_n(v)} - 1 \right| = \left| \underbrace{\left(\frac{f_n(z)}{\hat{f}_n(z)} \right)}_a \underbrace{\left(\frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \hat{U}_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} \right)}_{b/c} - 1 \right|.$$

Since $|a(b/c) - 1| \leq 2|a - 1| + |(b - c)/c|$ when $|(b - c)/c| \leq 1$, the result follows from $|a - 1| = O_{\mathbb{P}_n}(n^{-b}) = o_p(\delta_n/\bar{a}_n)$ holding by NK.2 for some $b > 0$ and from

$$\begin{aligned} |(b - c)/c| &\leq \max_{1 \leq i \leq n} \|\hat{U}_i - U_i\| \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} + \left| \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} - 1 \right|, \\ &\leq O_{\mathbb{P}_n}(n^{-b}) O_{\mathbb{P}_n}(1) + O_{\mathbb{P}_n} \left(\sqrt{\frac{\log n}{nh_n^d}} \right) = O_{\mathbb{P}_n}(n^{-b}) = o_{\mathbb{P}_n}(\delta_n/\bar{a}_n) \end{aligned}$$

for some $b > 0$ where we used NK.2, the results of Step 6, and the condition that $nh_n^d \rightarrow \infty$ at a polynomial rate.

Step 4. Verification of C.2. By NK.1 and $1 \lesssim E_{P_n}[g_v^2] \lesssim 1$ uniformly in $v \in \mathcal{V}$ holding by Step 6 give

$$\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n(g_v)}{E_{P_n}[g_v^2]} - \frac{\mathbb{B}_n(g_v)}{E_{P_n}[g_v^2]} \right| = O_{P_n}(\delta_n),$$

where $v \mapsto \mathbb{B}_n(g_v)$ is zero-mean P_n -Brownian bridge, with a.s. continuous sample paths. This and the condition on the remainder term in NK.1 in turn imply C.2(a).

To show C.2(b) we need to show that for any $C > 0$

$$P_n \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(\hat{g}_v)}{\mathbb{E}_n[\hat{g}_v^2]} - \frac{\bar{\mathbb{B}}_n(g_v)}{E_{P_n}[g_v^2]} \right| > C\delta_n \mid \mathcal{D}_n \right) = o_{P_n}(1/\ell_n),$$

where $\bar{\mathbb{B}}_n$ is a copy of \mathbb{B}_n , which is independent of the data. First, Condition NK.1 with the fact that $1 \lesssim E_{P_n}[g_v^2] \lesssim 1$ uniformly in $v \in \mathcal{V}$ implies that

$$P_n \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]} - \frac{\bar{\mathbb{B}}_n(g_v)}{E_{P_n}[g_v^2]} \right| > C\delta_n \mid \mathcal{D}_n \right) = o_{P_n}(1/\ell_n).$$

Therefore, in view of the triangle inequality and the union bound, it remains to show that

$$P_n \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(\hat{g}_v)}{\mathbb{E}_n[\hat{g}_v^2]} - \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]} \right| > C\delta_n \mid \mathcal{D}_n \right) = o_{P_n}(1/\ell_n). \quad (\text{H.1})$$

We have that

$$\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(\hat{g}_v)}{\mathbb{E}_n[\hat{g}_v^2]} - \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]} \right| \leq \sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(\hat{g}_v - g_v)}{E_{P_n}[g_v^2]} \right| + \sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]} \right| \sup_{v \in \mathcal{V}} \left| \frac{\sigma_n(v)}{s_n(v)} - 1 \right|.$$

We observe that

$$\begin{aligned} & E_{P_n} \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]} \right| \sup_{v \in \mathcal{V}} \left| \frac{\sigma_n(v)}{s_n(v)} - 1 \right| \mid \mathcal{D}_n \right) \\ &= E_{P_n} \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]} \right| \mid \mathcal{D}_n \right) \sup_{v \in \mathcal{V}} \left| \frac{\sigma_n(v)}{s_n(v)} - 1 \right| = O_{P_n} \left(\sqrt{\log nn^{-b}} \right) = O_{P_n}(1/\ell_n), \end{aligned}$$

where the last equality follows from Steps 5 and Step 3. Also we note that

$$\begin{aligned} & E_{P_n} \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(\hat{g}_v - g_v)}{E_{P_n}[g_v^2]} \right| \mid \mathcal{D}_n \right) \\ & \leq_{(1)} O_{P_n}(\sqrt{\log n}) \sup_{v \in \mathcal{V}} \frac{\left\| \left(\frac{U_{ij}}{f_n(z)} - \frac{\hat{U}_{ij}}{\hat{f}_n(z)} \right) \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\frac{1}{f_n(z)} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}} \\ & \lesssim_{(2)} O_{P_n}(\sqrt{\log n}) \sup_{v \in \mathcal{V}} \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) (1 + |U_{ij}|) \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}} \left(\left| \frac{f_n(z)}{\hat{f}_n(z)} - 1 \right| \vee \max_{1 \leq i \leq n} \|\hat{U}_i - U_i\| \right) \\ & \leq_{(3)} O_{P_n}(\sqrt{\log n}) O_{P_n}(1) O_{P_n}(n^{-b}) = o_{P_n}(1/\ell_n), \end{aligned}$$

where (1) follows from Step 5, (2) by elementary inequalities, and (3) by Step 6 and NK.2. It follows that (H.1) holds by Markov's Inequality.

Step 5. This step shows that

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]} \right| \mid \mathcal{D}_n \right) = O_{P_n}(\sqrt{\log n}) \quad (\text{H.2})$$

$$\begin{aligned} E_{P_n} \left(\sup_{v \in \mathcal{V}} \left| \frac{\mathbb{G}_n^o(\hat{g}_v - g_v)}{E_{P_n}[g_v^2]} \right| \mid \mathcal{D}_n \right) &\leq O_{P_n}(\sqrt{\log n}) \times \\ &\times \sup_{v \in \mathcal{V}} \frac{\left\| \left(\frac{U_{ij}}{f_n(z)} - \frac{\hat{U}_{ij}}{\hat{f}_n(z)} \right) \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\frac{1}{f_n(z)} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}}. \end{aligned} \quad (\text{H.3})$$

To show (H.2) we use Lemma 13 applied to $X_v = \frac{\mathbb{G}_n^o(g_v)}{E_{P_n}[g_v^2]}$ conditional on \mathcal{D}_n . First, we compute

$$\sigma(X) = \sup_{v \in \mathcal{V}} E_{P_n}(X_v^2 | \mathcal{D}_n) = \sup_{v \in \mathcal{V}} \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}} = 1 + o_{P_n}(1),$$

where the last equality holds by Step 6. Second, we observe that for $v = (z, j)$ and $\bar{v} = (\bar{z}, j)$

$$\sigma(X_v - X_{\bar{v}}) \leq \Upsilon_n \|z - \bar{z}\|, \quad \Upsilon_n := \sup_{v \in \mathcal{V}, 1 \leq k \leq d} \left\| \nabla_{z_k} \frac{\frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}} \right\|_{\mathbb{P}_{n,2}}.$$

We have that

$$\begin{aligned} \Upsilon_n &\leq \sup_{v \in \mathcal{V}, 1 \leq k \leq d} \frac{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}} \\ &+ \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}} \cdot \frac{\left| \nabla_{z_k} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}} \right|}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}}}. \end{aligned}$$

which is bounded with probability converging to one by $C(h_n^{-1} + 1)$ for large n by Step 6 and NK.2. Since J is finite, it follows that for all large $n > n_0$, the covering number for \mathcal{V} under $\rho(v, \bar{v}) = \sigma(X_v - X_{\bar{v}})$ obeys with probability converging to 1,

$$N(\varepsilon, \mathcal{V}, \rho) \leq \left(\frac{J^{1/d} (1 + C(1 + h_n^{-1}) \text{diam}(\mathcal{V}))}{\varepsilon} \right)^d, \quad 0 < \varepsilon < \sigma(X),$$

Hence $\log N(\varepsilon, \mathcal{V}, \rho) \lesssim \log n + \log(1/\varepsilon)$. Hence by Lemma 13, we have that

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} |X_v| \mid \mathcal{D}_n \right) \leq \sigma(X) + \int_0^{2\sigma(X)} \sqrt{\log(n/\varepsilon)} d\varepsilon = O_{P_n}(\sqrt{\log n}).$$

To show (H.3) we use Lemma 13 applied to $X_v = \frac{\mathbb{G}_n^o(\hat{g}_v - g_v)}{E_{P_n}[g_v^2]}$ conditional on \mathcal{D}_n . First, we compute

$$\sigma(X) = \sup_{v \in \mathcal{V}} E_{P_n}(X_v^2 | \mathcal{D}_n) = \sup_{v \in \mathcal{V}} \frac{\left\| \left(\frac{U_{ij}}{f_n(z)} - \frac{\hat{U}_{ij}}{\hat{f}_n(z)} \right) \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\frac{1}{f_n(z)} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}.$$

Second, we observe that for $v = (z, j)$ and $\bar{v} = (\bar{z}, j)$

$$\sigma(X_v - X_{\bar{v}}) \leq (\Upsilon_n + \hat{\Upsilon}_n) \|z - \bar{z}\|,$$

where

$$\hat{\Upsilon}_n := \sup_{v \in \mathcal{V}, 1 \leq k \leq d} \left\| \nabla_{z_k} \frac{\frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \hat{U}_{ij}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} \right\|_{\mathbb{P}_{n,2}},$$

and Υ_n is the same as defined above.

We have that

$$\begin{aligned} \hat{\Upsilon}_n &\leq \sup_{v \in \mathcal{V}, 1 \leq k \leq d} \frac{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \hat{U}_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} \\ &\quad + \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \hat{U}_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} \cdot \frac{\left| \nabla_{z_k} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}} \right|}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} \end{aligned}$$

The first term is bounded by

$$\sup_{v \in \mathcal{V}, 1 \leq k \leq d} \frac{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} + \max_{1 \leq i \leq n} \|\hat{U}_i - U_i\| \sup_{v \in \mathcal{V}, 1 \leq k \leq d} \frac{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}$$

which is bounded by $C(1 + h_n^{-1}) + O_{P_n}(n^{-b})O_{P_n}(1)$ for large n by Step 6 and NK.2. In the second term, the left term of the product is bounded by

$$\sup_{v \in \mathcal{V}, 1 \leq k \leq d} \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} + \max_{1 \leq i \leq n} \|\hat{U}_i - U_i\| \sup_{v \in \mathcal{V}, 1 \leq k \leq d} \frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}$$

which is bounded by $C(1 + o_{P_n}(1)) + O_{P_n}(n^{-b})O_{P_n}(1)$ for large n by Step 6 and NK.2; the right term of the product is bounded by $C(1 + o_{P_n}(1))$ by Step 6. Conclude that $\hat{\Upsilon}_n \leq C(1 + h_n^{-1})$ for some constant $C > 0$ with probability converging to one.

Since J is finite, it follows that for all large $n > n_0$, the covering number for \mathcal{V} under $\rho(v, \bar{v}) = \sigma(X_v - X_{\bar{v}})$ obeys with probability converging to 1,

$$N(\varepsilon, \mathcal{V}, \rho) \leq \left(\frac{J^{1/d}(1 + C(1 + h_n^{-1})\text{diam}(\mathcal{V}))}{\varepsilon} \right)^d, 0 < \varepsilon < \sigma(X),$$

Hence

$$\log N(\varepsilon, \mathcal{V}, \rho) \lesssim \log n + \log(1/\varepsilon).$$

Hence by Lemma 13, we have that

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} |X_v| \mid \mathcal{D}_n \right) \leq \sigma(X) + \int_0^{2\sigma(X)} \sqrt{\log(n/\varepsilon)} d\varepsilon = O_{P_n}(\sqrt{\log n})\sigma(X).$$

Step 6. The claim of this step are the following relations: uniformly in $v \in \mathcal{V}, 1 \leq k \leq d$

$$1 \lesssim \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}} \lesssim 1 \quad (\text{H.4})$$

$$1 \lesssim \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{P_{n,2}} \lesssim 1 \quad (\text{H.5})$$

$$h_n^{-1} \lesssim \left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}} \lesssim h_n^{-1} \quad (\text{H.6})$$

$$h_n^{-1} \lesssim \left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{P_{n,2}} \lesssim h_n^{-1} \quad (\text{H.7})$$

$$h_n^{-1} \lesssim \left| \nabla_{z_k} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{P_{n,2}} \right| \lesssim h_n^{-1} \quad (\text{H.8})$$

and

$$\frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} = 1 + O_{\mathbb{P}_n} \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \quad (\text{H.9})$$

$$\frac{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}} = 1 + O_{\mathbb{P}_n} \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \quad (\text{H.10})$$

$$\frac{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}}{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}} = 1 + O_{\mathbb{P}_n} \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \quad (\text{H.11})$$

$$\frac{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}}{\left\| \nabla_{z_k} \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right\|_{\mathbb{P}_{n,2}}} = 1 + O_{\mathbb{P}_n} \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \quad (\text{H.12})$$

The proofs of (H.4)-(H.8) are all similar to one another, as are those of (H.9)-(H.12), and are standard in the kernel estimator literature. We therefore prove only (H.4) and (H.9) to demonstrate the argument. To establish (H.4) we have

$$\begin{aligned} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}^2 &= h_n^{-d} \int \mathbf{K}^2((z - \bar{z})/h_n) E[U_{ij}^2 | \bar{z}] f_n(\bar{z}) d\bar{z} \\ &\leq_{(1)} h_n^{-d} \int \mathbf{K}^2((z - \bar{z})/h_n) C d\bar{z} \leq_{(2)} \int \mathbf{K}^2(u) C, \end{aligned}$$

for some constant $0 < C < \infty$, where in (1) we use the assumption that $E[U_{ij}^2 | z]$ and $f_n(z)$ are bounded uniformly from above and in (2) we use the assumption that \mathcal{Z} is bounded away from the boundary of the support of Z_i by at least h_n . On the other hand,

$$\begin{aligned} \left\| \frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) U_{ij} \right\|_{\mathbb{P}_{n,2}}^2 &= h_n^{-d} \int \mathbf{K}^2((z - \bar{z})/h_n) E[U_{ij}^2 | \bar{z}] f_n(\bar{z}) d\bar{z} \\ &\geq_{(1)} h_n^{-d} \int \mathbf{K}^2((z - \bar{z})/h_n) C d\bar{z} \geq_{(2)} \int \mathbf{K}^2(u) C, \end{aligned}$$

for some constant $0 < C < \infty$, where in (1) we use the assumption that $E[U_{ij}^2 | z]$ and $f_n(z)$ are bounded away from zero uniformly in n , and in (2) we use the assumption that \mathcal{Z} is bounded away from the boundary of the support of Z_i by at least h_n .

Moving to (H.9), it suffices to show that uniformly in $v \in \mathcal{V}$,

$$\mathbb{E}_n \left(\left(\frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right)^2 U_{ij}^2 \right) - E_{P_n} \left(\left(\frac{1}{h_n^{d/2}} \mathbf{K} \left(\frac{z - Z_i}{h_n} \right) \right)^2 U_{ij}^2 \right) = O_{P_n} \left(\sqrt{\frac{\log n}{n h_n^d}} \right),$$

or equivalently

$$\mathbb{E}_n \left(\mathbf{K}^2 \left(\frac{z - Z_i}{h_n} \right) U_{ij}^2 \right) - E_{P_n} \left(\mathbf{K}^2 \left(\frac{z - Z_i}{h_n} \right) U_{ij}^2 \right) = O_{P_n} \left(\sqrt{\frac{h_n^d \log n}{n}} \right). \quad (\text{H.13})$$

Given the boundedness of U_{ij} imposed by Condition R, this is in fact a standard result on local empirical processes, using Pollard's empirical process methods. Specifically, (H.13) follows by the application of Theorem 37 in chapter II of Pollard (1984). \blacksquare

H.2. Proof of Lemma 8. To show claim (1), we need to establish that for

$$\varphi_n = o(1) \cdot \left(\frac{h_n}{\sqrt{\log n}} \right),$$

for any $o(1)$ term, we have that

$$\sup_{\|v-v'\| \leq \varphi_n} |Z_n^*(v) - Z_n^*(v')| = o_{P_n}(1).$$

Consider the stochastic process $X = \{Z_n(v), v \in \mathcal{V}\}$. We shall use the standard maximal inequality stated in Lemma 13. From the proof of Lemma 7 we have that for $v = (z, j)$ and $v' = (z', j)$, $\sigma(Z_n^*(v) - Z_n^*(v')) \leq C(1 + h_n^{-1})\|z - z'\|$, where C is some constant that does not depend on n , and $\log N(\varepsilon, \mathbf{V}, \rho) \lesssim \log n + \log(1/\varepsilon)$. Since

$$\|v - v'\| \leq \varphi_n \implies \sigma(Z_n^*(v) - Z_n^*(v')) \leq C \frac{o(1)}{\sqrt{\log n}},$$

we have

$$E \sup_{\|v-v'\| \leq \varphi_n} |X_s - X_t| \lesssim \int_0^C \frac{o(1)}{\sqrt{\log n}} \sqrt{\log(n/\varepsilon)} d\varepsilon \lesssim \frac{o(1)}{\sqrt{\log n}} \sqrt{\log n} = o(1).$$

Hence the conclusion follows from Markov's Inequality.

Under Condition V by lemma 2

$$r_n \lesssim \left(\sqrt{\frac{\log n}{n h_n^d} \log n} \right)^{1/\rho_n} c_n^{-1},$$

so $r_n = o(\varphi_n)$ if

$$\left(\sqrt{\frac{\log n}{n h_n^d} \log n} \right)^{1/\rho_n} c_n^{-1} = o \left(\frac{h_n}{\sqrt{\log n}} \right).$$

Thus, Condition S is satisfied. ■

H.3. Proof of Theorem 8. To prove this theorem, we use the Rio-Massart coupling. First we note that

$$\mathcal{M} = \{h_n^{d/2} f_n(z) g_v(U_i, Z_i) = e'_j U_i \mathbf{K}((z - Z_i)/h_n), z \in \mathcal{Z}, j \in \{1, \dots, J\}\}$$

is the product of $\{e'_j U_i, j \in 1, \dots, J\}$ with covering number trivially bounded above by J and $\mathcal{K} := \{\mathbf{K}((z - Z_i)/h_n), z \in \mathcal{Z}\}$ obeys $\sup_Q N(\epsilon, \mathcal{K}, L_1(Q)) \lesssim \epsilon^{-\nu}$ for some finite constant ν ; see Lemma 4.1 of Rio (1994). Therefore, By Lemma A.1 in Ghosal, Sen, and van der Vaart (2000), we have that

$$\sup_Q N(\epsilon, \mathcal{M}, L_1(Q)) \lesssim J(\epsilon/2)^{-\nu} \lesssim \epsilon^{-\nu}. \quad (\text{H.14})$$

Next we bound, for $\mathbf{K}_l(u) = \partial \mathbf{K}(u)/\partial u_l$

$$\begin{aligned} TV(\mathcal{M}) &\leq \sup_{f \in \mathcal{M}} \int |D_{(x'_1, x'_2)'} f(x_1, x_2)| dx_1 dx_2 \\ &\leq \sup_{v \in \mathcal{V}} \int_{I^d} \int_{I^{d_1}} \left(\sum_{l=1}^{d_1} |e'_j D_{x_{1l}} \varphi_n(x_1) \mathbf{K}((z - \tilde{\varphi}_n(x_2))/h_n)| \right. \\ &\quad \left. + \sum_{k=1}^d |e'_j \varphi_n(x_1) \nabla \mathbf{K}((z - \tilde{\varphi}_n(x_2))/h_n) h_n^{-1} \partial \tilde{\varphi}(x_2)/\partial x_{2k}| \right) dx_1 dx_2 \\ &\leq C \sup_{v \in \mathcal{V}} \int_{I^d} \left(|\mathbf{K}((z - \tilde{\varphi}_n(x_2))/h_n)| + h_n^{-1} |\mathbf{K}_l((z - \tilde{\varphi}_n(x_2))/h_n)| B \right) dx_2 \\ &\leq C h_n^d + C h_n^{-1} h_n^d \leq C h_n^{d-1} =: K(\mathcal{M}) \end{aligned}$$

where C is a generic constant, possibly different in different places, and where we rely on

$$\int_{I^{d_1}} |D_{x_{1l}} \varphi_n(x_1)| dx_1 \leq B, \quad \sup_{x_1} |e'_j \varphi_n(x_1)| \leq 1, \quad \sup_{x_2} |\partial \tilde{\varphi}(x_2)/\partial x_{2k}| \leq B$$

as well as on

$$\int_{I^d} |\mathbf{K}((z - \tilde{\varphi}_n(x_2))/h_n)| dx_2 \leq C h^d, \quad \int_{I^d} |\mathbf{K}_l((z - \tilde{\varphi}_n(x_2))/h_n)| dx_2 \leq C h^d.$$

To see how the latter relationships holds, note that $Y = \tilde{\varphi}_n(v)$ when $v \sim U(I^d)$ has a density bounded uniformly from above: $f_Y(y) \lesssim 1/|\det \partial \tilde{\varphi}_n(v)/\partial v| \lesssim 1/c$. Moreover, the functions $|\mathbf{K}((z - y)/h_n)|$ and $|\mathbf{K}_l((z - y)/h_n)|$ are bounded above by some constant \bar{K} and are non-zero only over a y belonging to cube centered at z of volume $(2h)^d$. Hence

$$\int_{I^d} |\mathbf{K}((z - \tilde{\varphi}_n(x_2))/h_n)| dx_2 \leq \int_{I^d} |\mathbf{K}((z - y)/h_n)| f_Y(y) dy \leq \bar{K} (2h)^d (1/c) \leq C h^d,$$

and similarly for the second term.

By the Rio-Massart coupling we have that for some constant C and $t \geq C \log n$:

$$\mathbb{P}_n \left(\sqrt{n} \sup_{f \in \mathcal{M}} |\mathbb{G}_n(f) - \mathbb{B}_n(f)| \geq C \sqrt{tn^{\frac{d+d_1-1}{d+d_1}} K(\mathcal{M})} + Ct \sqrt{\log n} \right) \leq e^{-t},$$

which implies that

$$\mathbb{P}_n \left(\sup_{v \in \mathcal{V}} |\mathbb{G}_n(g_v) - \mathbb{B}_n(g_v)| \geq n^{-1/2} C \sqrt{tn^{\frac{d+d_1-1}{d+d_1}} h_n^{d-1} h_n^{-d/2}} + n^{-1/2} h_n^{-d/2} Ct \sqrt{\log n} \right) \leq e^{-t},$$

which upon inserting $t = C \log n$ gives

$$\mathbb{P}_n \left(\sup_{v \in \mathcal{V}} |\mathbb{G}_n(g_v) - \mathbb{B}_n(g_v)| \geq C \left[n^{-1/2(d+d_1)} (h_n^{-1} \log n)^{1/2} + (n h_n^d)^{-1/2} \log^{3/2} n \right] \right) \lesssim 1/n.$$

This implies the required conclusion. Note that $g_v \mapsto \mathbb{B}_n(g_v)$ is continuous under the $L_1(f_X)$ metric by the Rio-Massart coupling, which implies continuity of $v \mapsto \mathbb{B}_n(g_v)$, since $v - v' \rightarrow 0$ implies $g_v - g_{v'} \rightarrow 0$ in the $L_1(f_X)$ metric. \blacksquare

H.4. Proof of Theorem 9. Step 1. First we note that (G.2), is implied by

$$E_{\mathbb{P}_n} \left(\sup_{v \in \mathcal{V}} |\mathbb{G}_n^o(g_v) - \bar{\mathbb{B}}_n(g_v)| \mid \mathcal{D}_n \right) = o_{\mathbb{P}_n}(\delta_n/\ell_n) \quad (\text{H.15})$$

in view of the Markov inequality. Using calculations similar to those in Step 5 in the proof of Lemma 7, we can conclude that for $X_v := \mathbb{G}_n^o(g_v)$, $v = (z, j)$, and $\bar{v} = (\bar{z}, j)$,

$$\sigma(X_v - X_{\bar{v}}) \leq \|v - \bar{v}\| O_{\mathbb{P}_n}(1 + h_n^{-1}),$$

where $\sigma(X_v - X_{\bar{v}}) := E_{\mathbb{P}_n}((X_v - X_{\bar{v}})^2 \mid \mathcal{D}_n)$. Application of the Gaussian maximal inequality quoted in Lemma 13, similarly Step 6 in the proof of Lemma 7, then gives:

$$E_{\mathbb{P}_n} \left(\sup_{\|v - \bar{v}\| \leq \varepsilon} |X_v - X_{\bar{v}}| \mid \mathcal{D}_n \right) = \varepsilon O_{\mathbb{P}_n} \left((1 + h_n^{-1}) \sqrt{\log n} \right). \quad (\text{H.16})$$

If we set

$$\varepsilon \propto o \left(\frac{\delta_n/\ell_n}{h_n^{-1} \sqrt{\log n}} \right)$$

then

$$E_{\mathbb{P}_n} \left(\sup_{\|v - \bar{v}\| \leq \varepsilon} |X_v - X_{\bar{v}}| \mid \mathcal{D}_n \right) = o_{\mathbb{P}_n}(\delta_n/\ell_n). \quad (\text{H.17})$$

Next we setup a regular mesh $\mathcal{V}_0 \subset \mathcal{V}$ with meshwidth ε . The cardinality of the mesh is given by

$$K_n \propto (1/\varepsilon)^d \propto h_n^{-d} \lambda_n^d, \quad \lambda_n = \frac{\sqrt{\log n}}{o(\delta_n/\ell_n)}.$$

With such mesh selection, we have that

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} |X_v - X_{\pi(v)}| \mid \mathcal{D}_n \right) \leq E_{P_n} \left(\sup_{\|v - \bar{v}\| \leq \varepsilon} |X_v - X_{\bar{v}}| \mid \mathcal{D}_n \right) \leq o_{P_n}(\delta_n/\ell_n), \quad (\text{H.18})$$

where $\pi(v)$ denotes a point in \mathcal{V}_0 that is closest to v .

The steps given below will show that there is a Gaussian process $\{Z_v, v \in \mathcal{V}\}$, which is independent of \mathcal{D}_n , having the same law as $\{\mathbb{B}_n(g_v), v \in \mathcal{V}\}$, and having the following two key properties:

$$E_{P_n} \left(\sup_{v \in \mathcal{V}_0} |X_v - Z_v| \mid \mathcal{D}_n \right) = o_{P_n}(\delta_n/\ell_n), \quad (\text{H.19})$$

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} |Z_v - Z_{\pi(v)}| \right) = o(\delta_n/\ell_n). \quad (\text{H.20})$$

The claim of the lemma then follows by setting $\{\bar{\mathbb{B}}_n(g_v), v \in \mathcal{V}\} = \{Z_v, v \in \mathcal{V}\}$, and then noting that

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} |X_v - Z_v| \mid \mathcal{D}_n \right) = o_{P_n}(\delta_n/\ell_n)$$

holds by the triangle inequality for the sup norm and (H.18)-(H.20). Note that the last display is equivalent to (H.15). We now prove these assertions in the followings steps.

Step 2. In this step we construct the process Z_v on points $v \in \mathcal{V}_0$, and show that (H.19) holds. In what follows, we use the notation $(X_v)_{v \in \mathcal{V}_0}$ to denote a K_n vector collecting X_v with indices $v \in \mathcal{V}_0$. We have that conditional on the data \mathcal{D}_n ,

$$(X_v)_{v \in \mathcal{V}_0} = \hat{\Omega}_n^{1/2} \mathcal{N}, \quad \mathcal{N} \sim N(0, I),$$

where \mathcal{N} is independent of \mathcal{D}_n , and

$$\hat{\Omega} = \mathbb{E}_n[p_i p_i'] \quad \text{and} \quad p_i = (g_v(U_i, Z_i))_{v \in \mathcal{V}_0}.$$

We then set $(Z_v)_{v \in \mathcal{V}_0} = \Omega_n^{1/2} \mathcal{N}$ for $\Omega = E_{P_n}[p_i p_i']$ and the same \mathcal{N} as defined above.

Before proceeding further, we note that by construction the process $\{Z_v, v \in \mathcal{V}_0\}$ is independent of the data \mathcal{D}_n . This is facilitated by suitably enlarging the the probability space as needed.³¹

³¹Given the space (A', \mathcal{A}', P'_n) that carries \mathcal{D}_n and given a different space $(A'', \mathcal{A}'', P''_n)$ that carries $\{Z_v, v \in \mathcal{V}_0\}$ as well as its complete version $\{Z_v, v \in \mathcal{V}\}$, we can take (A, \mathcal{A}, P_n) as the product of the two spaces, thereby maintaining independence between the data and the constructed process. Since $\{Z_v, v \in \mathcal{V}\}$ constructed below takes values in a separable metric space, it suffices to take $(A'', \mathcal{A}'', P''_n)$ as the canonical probability space, as noted in Appendix A.

Since the support of \mathbf{K} is compact and points of the grid \mathcal{V}_0 are equally spaced, we have that

$$N_i := |\{v \in \mathcal{V}_0 : g_v(U_i, Z_i) \neq 0\}| \lesssim (h_n/\varepsilon_n)^d \lesssim \lambda_n^d.$$

Using the boundedness assumptions of the lemma, we have that

$$\|p_i\| \leq (\bar{U}/\underline{f})\sqrt{N_i}/h_n^{d/2} \lesssim (\lambda_n/h_n)^{d/2},$$

where \bar{U} is the upper bound on U and \underline{f} is the lower bound on the density f_n , both of which do not depend on n .

The application of Rudelson's Law of Large Numbers for operators, Rudelson (1999), yields

$$E_{P_n} \|\hat{\Omega}_n - \Omega_n\| \lesssim \sqrt{\log n / (n(h_n/\lambda_n)^d)}.$$

The application of the Gaussian maximal inequality quoted in Lemma 13 gives:

$$E_{P_n} \left(\sup_{v \in \mathcal{V}_0} |X_v - Z_v| \mid \mathcal{D}_n \right) \lesssim \sqrt{\log K_n} \max_{v \in \mathcal{V}_0} \sigma(X_v - Z_v).$$

Since $(X_v)_{v \in \mathcal{V}_0} - (Z_v)_{v \in \mathcal{V}_0} = (\hat{\Omega}_n^{1/2} - \Omega_n^{1/2})' \mathcal{N}$, we have that:

$$\max_{v \in \mathcal{V}_0} \sigma(X_v - Z_v)^2 \leq \|(\hat{\Omega}_n^{1/2} - \Omega_n^{1/2})^2\| \leq \|\hat{\Omega}_n^{1/2} - \Omega_n^{1/2}\|^2 \leq K_n \|\hat{\Omega}_n - \Omega_n\|^2,$$

where the last inequality follows by a useful matrix inequality derived in Chetverikov (2011).

Putting bounds together and using $\log K_n \lesssim \log n$

$$E_{P_n} \left(\sup_{v \in \mathcal{V}_0} |X_v - Z_v| \mid \mathcal{D}_n \right) = O_{P_n} \left(\sqrt{\log n} \sqrt{K_n} \sqrt{\log n / (n(h_n/\lambda_n)^d)} \right) \quad (\text{H.21})$$

$$= O_{P_n}(\log n / \sqrt{n(h_n/\lambda_n)^{2d}}) = o_{P_n}(\delta_n/\ell_n), \quad (\text{H.22})$$

where the last condition holds by the conditions on the bandwidth.

Step 3. In this step we complete the construction of the process $\{Z_v, v \in \mathcal{V}\}$. We have defined the process Z_v for all $v \in \mathcal{V}_0$. We want to embed these random variables into a path of a Gaussian process $\{Z_v, v \in \mathcal{V}\}$, whose covariance function is given by $(v, \bar{v}) \mapsto E_{P_n}[g_v g_{\bar{v}}]$. We want to maintain the independence of the process from \mathcal{D}_n . The construction follows by Lemma 11 in Belloni, Chernozhukov, and Fernandez-Val (2011). This lemma requires that a version of $\{Z_v, v \in \mathcal{V}\}$ has a.s. uniformly continuous sample paths, which follows from the Gaussian maximal inequalities and entropy calculations similar to those given in Step 3 in the proof of Lemma 7. Indeed, we can conclude that

$$\sigma(Z_v - Z_{\bar{v}}) \leq \|v - \bar{v}\| C(1 + h_n^{-1}),$$

which establishes boundedness of \mathcal{V} under the standard deviation pseudometric. Moreover, application of the Gaussian maximal inequality Lemma 13 to $X_v := \mathbb{G}_n^o(g_v)$ gives:

$$E_{P_n} \left(\sup_{\|v-\bar{v}\| \leq \varepsilon} |Z_v - Z_{\bar{v}}| \right) \leq C\varepsilon \left((1 + h_n^{-1}) \sqrt{\log n} \right).$$

By a standard argument, e.g. van der Vaart and Wellner (1996), these facts imply that the paths of Z_v are a.s. uniformly continuous.³²

The last claim (H.20) follows from the preceding display, the choice of meshwidth ε , and the inequality:

$$E_{P_n} \left(\sup_{v \in \mathcal{V}} |Z_v - Z_{\pi(v)}| \right) \leq E_{P_n} \left(\sup_{\|v-\bar{v}\| \leq \varepsilon} |Z_v - Z_{\bar{v}}| \right) = o(\delta_n/\ell_n).$$

■

APPENDIX I. ASYMPTOTIC LINEAR REPRESENTATION FOR SERIES ESTIMATOR OF A CONDITIONAL MEAN

In this section we use the primitive conditions set out in Example 5 of the main text to verify the required asymptotically linear representation for $\sqrt{n}(\hat{\beta}_n - \beta_n)$ using Newey (1997). This representation is also Condition (b) of Theorem 7. We now reproduce the imposed conditions from the example for clarity. We note that it is also possible to develop similar conditions for nonlinear estimators, see for example Theorem 1(d) of Horowitz and Mammen (2004).

We have that $\theta_n(v) = E_{P_n}[Y_i|V_i = v]$, assumed to be a continuous function. There is an i.i.d. sample $(Y_i, V_i), i = 1, \dots, n$, with $\mathcal{V} \subseteq \text{support}(V_i) \subseteq [0, 1]^d$ for each n . Here d does not depend on n , but all other parameters, unless stated otherwise, can depend on n . Then we have $\theta_n(v) = p_n(v)' \beta_n + A_n(v)$, for $p_n : \mathcal{V} \mapsto \mathbb{R}^{K_n}$ representing the series functions; β_n is the coefficient of the best least squares approximation to $\theta_n(v)$ in the population, and $A_n(v)$ is the approximation error. The number of series terms K_n depends on n .

Recall that we have imposed the following technical conditions in the main text:

Uniformly in n , (i) p_n are either b-splines of a fixed order or trigonometric series terms or any other terms $p_n = (p_{n1}, \dots, p_{nK_n})$ such that $\|p_n(v)\| \lesssim \zeta_n = \sqrt{K_n}$ for all $v \in \text{support}(V_i)$, $E_{P_n}[\|p_n(V_i)\|^3] \lesssim K_n^{3/2}$, $\|p_n(v)\| \gtrsim \zeta'_n \geq 1$

³²Note however that the process depends on n , and the statement here is a non-asymptotic statement, holding for any n . This property should not be confused with asymptotic equicontinuity, which does not hold here.

for all $v \in \mathcal{V}$, and $\log \text{lip}(p_n) \lesssim \log K_n$, (ii) the mapping $v \mapsto \theta_n(v)$ is sufficiently smooth, namely $\sup_{v \in \mathcal{V}} |A_n(v)| \lesssim K_n^{-s}$, for some $s > 0$, (iii) $\lim_{n \rightarrow \infty} (\log n)^c \sqrt{n} K_n^{-s} = 0$ for each $c > 0$,³³ (iv) for $\epsilon_i = Y_i - E_{P_n}[Y_i|V_i]$, $E_{P_n}[\epsilon_i^2|V_i = v]$ is bounded away from zero uniformly in $v \in \text{support}(V_i)$, and (v) eigenvalues of $Q_n = E_{P_n}[p_n(V_i)p_n(V_i)']$ are bounded away from zero and from above, and (vi) $E_{P_n}[|\epsilon_i|^4|V_i = v]$ is bounded from above uniformly in $v \in \text{support}(V_i)$, (vii) $\lim_{n \rightarrow \infty} (\log n)^c K_n^5/n = 0$ for each $c > 0$.

We impose Condition (i) directly through the choice of basis functions. Condition (ii) is a standard condition on the error of the series approximation, and is the same as Assumption A3 of Newey (1997), also used by Chen (2007). Condition (v) is Assumption 2(i) of Newey (1997). The constant s will depend on the choice of basis functions. For example, if splines or power series are used, then $\alpha = s/d$, where s is the number of continuous derivatives of $\theta_n(v)$ and d is the dimension of v . Restrictions on K_n in conditions (iii) and (vii) require that $s > 5d/2$. Conditions (i), (vi), and (vii) and Theorem 7, namely Corollary 1, ensure that the constraint on the growth rate for the number of series terms is satisfied.

Define $S_n \equiv E[\epsilon_i^2 p_n(V_i)p_n(V_i)']$ and $\Omega_n \equiv Q_n^{-1} S_n Q_n^{-1}$. Arguments based on Newey (1997) give the following lemma, which verifies the linear expansion required in condition (b) of Theorem 7 with $\delta_n = 1/\log n$.

Lemma 14 (Asymptotically Linear Representation of Series Estimator). *Suppose conditions (i)-(vii) hold. Then we have the following asymptotically linear representation:*

$$\Omega_n^{-1/2} \sqrt{n}(\hat{\beta}_n - \beta_n) = \Omega_n^{-1/2} Q_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n p_n(Z_i) \epsilon_i + o_{P_n}(1/\log n).$$

Proof of Lemma 14. As in Newey (1997), we have the following representation: with probability approaching one,

$$\hat{\beta}_n - \beta_n = n^{-1} \hat{Q}_n^{-1} \sum_{i=1}^n p_n(V_i) \epsilon_i + \nu_n, \quad (\text{I.1})$$

where $\hat{Q}_n \equiv \mathbb{E}_n[p_n(V_i)p_n(V_i)']$, $\epsilon_i \equiv Y_i - E_{P_n}[Y|V = V_i]$, $\nu_n \equiv n^{-1} \hat{Q}_n^{-1} \sum_{i=1}^n p_n(V_i) A_n(v)$, where $A_n(v) := \theta_n(v) - p_n(v)' \beta_n$. As shown in the proof of Theorem 1 of Newey (1997), we

³³This condition, which is based on Newey (1997) can be relaxed to $(\log n)^c K_n^{-s+1} \rightarrow 0$ and $(\log n)^c \sqrt{n} K_n^{-s}/\zeta'_n \rightarrow 0$, using the recent results of Belloni, Chen, and Chernozhukov (2010) for least squares series estimators.

have $\|\nu_n\| = O_P(K_n^{-\alpha})$. In addition, write

$$\bar{R}_n := \left[\hat{Q}_n^{-1} - Q_n^{-1} \right] n^{-1} \sum_{i=1}^n p_n(V_i) \epsilon_i = Q_n^{-1} \left[Q_n - \hat{Q}_n \right] n^{-1} \hat{Q}_n^{-1} \sum_{i=1}^n p_n(V_i) \epsilon_i.$$

Then it follows from the proof of Theorem 1 of Newey (1997) that

$$\|\bar{R}_n\| = O(\xi(K_n)K_n/n),$$

where $\xi(K_n) \equiv \sup_v \|p_n(v)\| = \sqrt{K_n}$ by condition (i). Combining the results above gives

$$\hat{\beta}_n - \beta_n = n^{-1} Q_n^{-1} \sum_{i=1}^n p_n(V_i) \epsilon_i + R_n, \quad (\text{I.2})$$

where the remainder term R_n satisfies

$$\|R_n\| = O\left(\frac{K_n^{3/2}}{n} + K_n^{-\alpha}\right).$$

Note that by condition (iv), eigenvalues of S_n^{-1} are bounded above. In other words, using the notation used in Corollary 1 in the main text, we have that $\tau_n \lesssim 1$. Then

$$\Omega_n^{-1/2} \sqrt{n} (\hat{\beta}_n - \beta_n) = n^{-1/2} \sum_{i=1}^n u_{i,n} + r_n, \quad (\text{I.3})$$

where

$$u_{i,n} := \Omega_n^{-1/2} Q_n^{-1} p_n(V_i) \epsilon_i, \quad (\text{I.4})$$

and the new remainder term r_n satisfies

$$\|r_n\| = O\left[n^{1/2} \left(K_n^{3/2}/n + K_n^{-\alpha}\right)\right].$$

Therefore, $r_n = o_{P_n}(1/\log n)$ if

$$(\log n) n^{1/2} \left(K_n^{3/2}/n + K_n^{-\alpha}\right) \rightarrow 0, \quad (\text{I.5})$$

which is satisfied under conditions (iii) and (vii), and we have proved the lemma. \blacksquare

APPENDIX J. ASYMPTOTIC LINEAR REPRESENTATION FOR LOCAL POLYNOMIAL ESTIMATOR OF A CONDITIONAL MEAN

In this section we provide details of Example 7 that are omitted in the main text. Recall that we assume that $P_n = P$ is fixed in this example, and impose the following conditions:

- (i) for each $j \in \mathcal{J}$, $\theta(z, j)$ is $(p+1)$ times continuously differentiable with respect to $z \in \mathcal{Z}$, where \mathcal{Z} is convex.
- (ii) the probability density function

f of Z_i is bounded above and bounded below from zero with continuous derivatives on \mathcal{Z} ; (iii) for $Y_i(j) := m(X_i, \mu, j)$, $Y_i := (Y_i(j), j \in \mathcal{J})'$, and $U_i := Y_i - E_P[Y_i|Z_i]$; and U_i is a bounded random vector; (iv) for each j , the conditional on Z_i density of U_i exists and is uniformly bounded from above and below, or, more generally, condition R stated in Appendix G holds; (v) $K(\cdot)$ has support on $[-1, 1]^d$, is twice continuously differentiable, $\int uK(u)du = 0$, and $\int K(u)du = 1$; (vi) $h_n \rightarrow 0$, $nh_n^{d+|\mathcal{J}|+1} \rightarrow \infty$, and $nh_n^{d+2(p+1)} \rightarrow 0$ at polynomial rates in n .

Let $\mathbf{K}(z/h) \equiv \mathbf{e}'_1 S_p^{-1} K_h(z) \mathbf{u}_p(z/h)$, and let

$$g_v(U, Z) := \frac{e'_j U}{(h_n^d)^{1/2} f(z)} \mathbf{K} \left(\frac{Z - z}{h_n} \right).$$

Then results obtained in Kong, Linton, and Xia (2010) give the following lemma, which verifies the linear expansion required in condition (b) of Theorem 8 with $\delta_n = 1/\log n$.

Lemma 15 (Asymptotically Linear Representation of Local Polynomial Estimator). *Suppose conditions (i)-(vi) hold. Then we have the following asymptotically linear representation: uniformly in $v = (z, j) \in \mathcal{V} \subseteq \mathcal{Z} \times \mathcal{J}$,*

$$(nh_n^d)^{1/2}(\hat{\theta}_n(v) - \theta_n(v)) = \mathbb{G}_n(g_v) + o_P(1/\log n).$$

Proof of Lemma 15. We first verify Assumptions A1-A7 in Kong, Linton, and Xia (2010) (KLX hereafter). In our example, $\rho(y; \theta) = \frac{1}{2}(y - \theta)^2$ using the notation in KLX. Then $\varphi(y; \theta)$ in Assumptions A1 and A2 in KLX is $\varphi(y; \theta) = \varphi(y - \theta) = -(y - \theta)$. Then Assumption A1 is satisfied since the pdf of U_i is bounded and U_i is a bounded random vector. Assumption A2 is trivially satisfied since $\varphi(u) = -u$. Assumption A3 follows since $K(\cdot)$ has compact support and is twice continuous differentiable. Assumption A4 holds by condition (ii) since X_i and X_j are independent in our example ($i \neq j$). Assumption A5 is implied directly by Condition (i). Since we have i.i.d. data, mixing coefficients ($\gamma[k]$ using the notation of KLX) are identically zeros for any $k \geq 1$. The regression error U_i is assumed to be bounded, so that ν_1 in KLX can be arbitrary large. Hence, to verify Assumption A6 of KLX, it suffices to check that for some $\nu_2 > 2$, $h_n \rightarrow 0$, $nh_n^d/\log n \rightarrow \infty$, $h_n^{d+2(p+1)}/\log n < \infty$, and $n^{-1}(nh_n^d/\log n)^{\nu_2/8} d_n \log n / M_n^{(2)} \rightarrow \infty$, where $d_n = (nh_n^d/\log n)^{-1/2}$ and $M_n^{(2)} = M^{1/4}(nh_n^d/\log n)^{-1/2}$ for some $M > 2$, by choosing $\lambda_2 = 1/2$ and $\lambda_1 = 3/4$ on page 1540 in KLX. By choosing a sufficiently large ν_2 (at least greater than 8), the following holds: $n^{-1}(nh_n^d)^{\nu_2/8} \rightarrow \infty$. Then condition (vi) implies

Assumption A6. Finally, condition (iv) implies Assumption A7 since we have i.i.d. data. Thus, we have verified all the conditions in KLX.

Let $\delta_n = 1/\log n$. Then it follows from Corollary 1 and Lemmas 8 and 10 of KLX that

$$\hat{\theta}_n(z, j) - \theta_n(z, j) = \frac{1}{nh_n^d f(z)} \mathbf{e}_1' S_p^{-1} \sum_{i=1}^n (e_j' U_i) K_h(Z_i - z) \mathbf{u}_p \left(\frac{Z_i - z}{h_n} \right) + B_n(z, j) + R_n(z, j), \quad (\text{J.1})$$

where \mathbf{e}_1 is a $|A_p| \times 1$ vector whose first element is one and all others are zeros, S_p is a $|A_p| \times |A_p|$ matrix such that $S_p = \{\int z^u (z^v)' du : u \in A_p, v \in A_p\}$, $\mathbf{u}_p(z)$ is a $|A_p| \times 1$ vector such that $\mathbf{u}_p(z) = \{z^u : u \in A_p\}$,

$$B_n(z, j) = O(h_n^{p+1}) \text{ and } R_n(z, j) = o_P \left(\frac{\delta_n}{(nh_n^d)^{1/2}} \right),$$

uniformly in $(z, j) \in \mathcal{V}$. The exact form of $B_n(z, j)$ is given in equation (12) of KLX. The result that $B_n(z, j) = O(h_n^{p+1})$ uniformly in (z, j) follows from the standard argument based on Taylor expansion given in Fan and Gijbels (1996), KLX, or Masry (1996). The condition that $nh_n^{d+2(p+1)} \rightarrow 0$ at a polynomial rate in n corresponds to the undersmoothing condition. Now the lemma follows from (J.1) immediately since $\mathbf{K}(z/h) \equiv \mathbf{e}_1' S_p^{-1} K_h(z) \mathbf{u}_p(z/h)$ is a kernel of order $(p+1)$ (See Section 3.2.2 of Fan and Gijbels (1996)). ■

APPENDIX K. LOCAL ASYMPTOTIC POWER COMPARISONS

We have shown in the main text that the test of $H_0 : \theta_{na} \leq \theta_{n0}$ of the form

$$\text{Reject } H_0 \text{ if } \theta_{na} > \hat{\theta}_{n0}(p),$$

can reject all local alternatives θ_{na} that are more distant than $\bar{\sigma}_n \bar{a}_n$. We now provide a couple of examples of local alternatives against which our test has non-trivial power, but for which the CvM statistic of Andrews and Shi (forthcoming), henceforth AS, does not. See also Armstrong (2011) for a comprehensive analysis of power properties of KS statistic of Andrews and Shi (forthcoming). It is evident from the results of AS on local asymptotic power that there are also models for which their CvM statistic will have power against some $n^{-1/2}$ alternatives, where our approach will not.³⁴ We conclude that neither approach dominates.

We consider two examples in which

$$Y_i = \theta_n(V_i) + U_i,$$

³⁴For the formal results, see AS Section 7, Theorem 4. In the examples that follow their Assumption LA3' is violated, as is also the case in the example covered in their Section 13.5.

where U_i are iid with $E[U_i|V_i] = 0$ and V_i are iid random variables uniformly distributed on $[-1, 1]$. Suppose that for all $v \in [-1, 1]$ we have

$$\theta^* \leq E[Y_i|V_i = v],$$

equivalently

$$\theta^* \leq \theta_0 = \min_{v \in [-1, 1]} \theta_n(v).$$

In the examples below we consider two specifications of the bounding function $\theta_n(v)$, each with

$$\min_{v \in [-1, 1]} \theta_n(v) = 0,$$

and we analyze asymptotic power against a local alternative $\theta_{na} > \theta_0$.

Following AS, consider the CvM test statistic

$$T_n(\theta) := \int \left[n^{1/2} \frac{\overline{m}_n(g; \theta)}{\widehat{\sigma}_n(g; \theta) \vee \varepsilon} \right]_-^2 dQ(g), \quad (\text{K.1})$$

for some $\varepsilon > 0$, where $[u]_- := -u1(u < 0)$ and θ is the parameter value being tested. In the present context we have

$$\overline{m}_n(g; \theta) := \frac{1}{n} \sum_{i=1}^n (Y_i - \theta) g(V_i),$$

where $g \in \mathcal{G}$ are instrument functions used to transform the conditional moment inequality $E[Y - \theta|V = v]$ a.e. $v \in \mathcal{V}$ to unconditional inequalities, and $Q(\cdot)$ is a measure on the space \mathcal{G} of instrument functions as described in AS Section 3.4. $\widehat{\sigma}_n(g; \theta)$ is a uniformly consistent estimator for $\sigma_n(g; \theta)$, the standard deviation of $n^{1/2}\overline{m}_n(g; \theta)$.

We can show that $T_n(\theta) = \tilde{T}_n(\theta) + o_p(1)$, where

$$\tilde{T}_n(\theta) := \int [\beta_n(\theta, g) / (\sigma_n(g; \theta) \vee \varepsilon) + w(\theta, g)]_-^2 dQ(g),$$

where $w(\theta, g)$ is a mean zero Gaussian process, and $\beta_n(\theta, g)$ is a deterministic function of the form

$$\beta_n(\theta, g) \equiv \sqrt{n} E\{[\theta_n(V_i) - \theta] g(V_i)\}.$$

For any θ , the testing procedure based on the CvM statistic rejects $H_0 : \theta \leq \theta_{n0}$ if

$$T_n(\theta) > c(\theta, 1 - \alpha),$$

where $c(\theta, 1 - \alpha)$ is a generalized moment selection (GMS) critical value that satisfies

$$c(\theta, 1 - \alpha) = (1 - \alpha)\text{-quantile of } \left(\int [\varphi_n(\theta, g) / (\sigma_n(g; \theta) \vee \varepsilon) + w(\theta, g)]_-^2 dQ(g) \right) + o_p(1).$$

$\varphi_n(\theta, g)$ is a GMS function that satisfies $0 \leq \varphi_n(\theta, g) \leq \beta_n(\theta, g)$ with probability approaching 1 whenever $\beta_n(\theta, g) \geq 0$, see AS Section 4.4 for further details. Relative to $\tilde{T}_n(\theta)$, in the integrand of the expression above $\varphi_n(\theta, g)$ is replaced with $\beta_n(\theta, g)$. Hence if

$$\sup_{g \in \mathcal{G}} [\beta_n(\theta_{na}, g)]_- \rightarrow 0,$$

for the sequence of local alternatives θ_{na} , then

$$\liminf_{n \rightarrow \infty} \mathbb{P}(T_n(\theta_{na}) > c(\theta_{na}, 1 - \alpha)) \leq \alpha,$$

since asymptotically $c(\theta_{na}, 1 - \alpha)$ exceeds the $1 - \alpha$ quantile of $\tilde{T}_n(\theta)$. It follows that the CvM test has only trivial power against such a sequence of alternatives. The same conclusion holds using plug-in asymptotic critical values, since these are no smaller than GMS critical values.

In the following two examples we now verify that $\sup_{g \in \mathcal{G}} [\beta_n(\theta_{na}, g)]_- \rightarrow 0$. We assume that instrument functions g are either boxes or cubes, defined in AS Section 3.3, and hence bounded between zero and one.

K.1. Example J.1 (Unique, well-defined optimum). Let the function $\theta(\cdot)$ be specified as

$$\theta_n(v) = |v|^a,$$

for some $a \geq 1$.

Let us now proceed to bound

$$\begin{aligned} [\beta_n(\theta_{na}, g)]_- &= \sqrt{n} [E \{[\theta_n(V_i) - \theta_{na}] g(V_i)\}]_- \\ &\leq \sqrt{n} [E [\theta_n(V_i) - \theta_{na}]]_- \\ &\leq \sqrt{n} E \{[\theta_n(V_i) - \theta_{na}]_-\} \\ &= \sqrt{n} \int_{-1}^1 (\theta_{na} - |v|^a) 1\{|v|^a \leq \theta_{na}\} dv \\ &= 2\sqrt{n} \int_0^1 (\theta_{na} - v^a) 1\{v \leq \theta_{na}^{1/a}\} dv \\ &= \frac{2a}{a+1} \sqrt{n} \theta_{na}^{(a+1)/a} \\ &\equiv \bar{\beta}_n. \end{aligned}$$

Note that

$$\theta_{na} = o\left(n^{-a/[2(a+1)]}\right) \Rightarrow \bar{\beta}_n \rightarrow 0.$$

Thus, in this case the asymptotic rejection probability of the CvM test for the local alternative θ_{na} is bounded above by α . On the other hand, by Theorems 1 and 2 of the main text, our test rejects all local alternatives θ_{na} that are more distant than $\bar{\sigma}_n \bar{a}_n$ with probability at least α asymptotically. It suffices to find a sequence of local alternatives θ_{na} such that $\theta_{na} = o\left(n^{-a/[2(a+1)]}\right)$ but $\theta_{na} \gg \bar{\sigma}_n \bar{a}_n$.

For instance, consider the case where $a = 2$. Then

$$\sqrt{n}\theta_{na}^{3/2} \rightarrow 0 \Rightarrow \bar{\beta}_n \rightarrow 0,$$

i.e. $\theta_{na} = o\left(n^{-1/3}\right) \Rightarrow \bar{\beta}_n \rightarrow 0$, so the CvM test has trivial asymptotic power against θ_{na} . In contrast, since this is a very smooth case, our approach can achieve $\bar{\sigma}_n \bar{a}_n = O(n^{-\delta})$ for some δ that can be close to $1/2$, for instance by using a series estimator with a slowly growing number of terms, or a higher-order kernel or local polynomial estimator. Our test would then be able to reject any θ_{na} that converges to zero faster than $n^{-1/3}$ but more slowly than $n^{-\delta}$.

K.2. Example J.2 (Deviation with Small Support). Now suppose that the form of the conditional mean function, $\theta_n(v) \equiv E[Y_i|V_i = v]$, is given by

$$\theta_n(v) := \bar{\theta}(v) - \tau_n^a (\phi(v/\tau_n) - \phi(0)),$$

where τ_n is a sequence of positive constants converging to zero and $\phi(\cdot)$ is the standard normal density function. Let $\bar{\theta}(v)$ be minimized at zero so that

$$\theta_0 = \min_{v \in [-1,1]} \theta_n(v) = \min_{v \in [-1,1]} \bar{\theta}(v) = 0.$$

Let the alternative be $\tilde{\theta}_{na} \equiv \tau_n^a \phi(0)$. Again, the behavior of the AS statistic is driven by $\left[\beta_n(\tilde{\theta}_{na}, g)\right]_-$, which we bound from above as follows.

$$\begin{aligned}
\left[\beta_n(\tilde{\theta}_{na}, g)\right]_- &= \sqrt{n} \left[E \left\{ \left[\theta_n(V_i) - \tilde{\theta}_{na} \right] g(V_i) \right\} \right]_- \\
&\leq \sqrt{n} \left[E \left\{ \theta_n(V_i) - \tilde{\theta}_{na} \right\} \right]_- \\
&= \sqrt{n} \left[E \left\{ \bar{\theta}(V_i) - \tau_n^a \phi(V_i/\tau_n) \right\} \right]_- \\
&\leq \sqrt{n} E \left\{ \tau_n^a \phi(V_i/\tau_n) \right\} \\
&= \frac{\sqrt{n}}{2} \int_{-1}^1 \tau_n^a \phi(v/\tau_n) dv \\
&\leq \frac{\sqrt{n}}{2} \tau_n^{a+1} \equiv \bar{\beta}_n.
\end{aligned}$$

Consider the case $a = 2$. If $\tau_n = o(n^{-1/6})$ then $\bar{\beta}_n \rightarrow 0$, so that again the CvM test has only trivial asymptotic power. If $\tau_n = n^{-1/6-c/2}$ for some small positive constant c , then $\tilde{\theta}_{na} \equiv n^{-1/3-c} \phi(0)$. Note that

$$f(v) := \tau_n^2 \phi(v/\tau_n) \Rightarrow f''(v) = \phi''(v/\tau_n) \leq \bar{\phi}'' < \infty,$$

for some constant $\bar{\phi}''$. Hence, if $\bar{\theta}(v)$ is twice continuously differentiable, we can use a series or kernel estimator to estimate $\theta_n(v)$ uniformly at the rate of $(\log n)^d n^{-2/5}$ for some $d > 0$, leading to non-trivial power against alternatives $\tilde{\theta}_{na}$ for sufficiently small c .

APPENDIX L. RESULTS OF ADDITIONAL MONTE CARLO EXPERIMENTS

In this section we present the results of some additional Monte Carlo experiments to further illustrate the finite-sample performance of our method. We consider two types of additional data-generating processes (DGPs). The first set of DGPs, DGPs 5-8, are motivated by Manski and Pepper (2009), discussed briefly in Example B of the main text. The second set, DGPs 9-12, are from Section 10.3 of AS.

L.1. Monte Carlo Designs. In DGPs 5-8 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

$$\max_{u \leq v} E[Y_i \cdot 1\{Z_i = t\} + y_0 \cdot 1\{Z_i \neq t\} | V_i = u], \quad (\text{L.1})$$

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

In DGP5, $V_0 = \mathcal{V}$ and the MIV assumption has no identifying power. In other words, the bound-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 DGP6, the MIV assumption has identifying power, and V_0 is a strict subset of \mathcal{V} . In DGPs 7-8, we set V_0 to be a singleton set.

Specifically, for DGPs 5-8 we generated 1000 independent samples as follows:

$$V_i \sim \text{Unif}[-2, 2], Z_i = 1\{\varphi_0(V_i) + \varepsilon_i > 0\}, \text{ and } Y_i = \min\{\max\{-0.5, \sigma_0(V_i)U_i\}, 0.5\},$$

where $\varepsilon_i \sim N(0, 1)$, $U_i \sim N(0, 1)$, $\sigma_0(V_i) = 0.1 \times |V_i|$, and (V_i, U_i) are statistically independent ($i = 1, \dots, n$). The bounding function has the form

$$\begin{aligned} \theta(v) &:= E[Y_i \cdot 1\{Z_i = 1\} + y_0 \cdot 1\{Z_i \neq 1\} | V_i = v] \\ &= -0.5\Phi[-\varphi_0(v)], \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. For DGP5, we set $\varphi_0(v) \equiv 0$. In this case, the bounding function is completely flat ($\theta_l(v) = -0.25$ for each $v \in \mathcal{V} = [-2, 1.5]$). For DGP6, an alternative specification is considered:

$$\varphi_0(v) = v1(v \leq 1) + 1(v > 1).$$

In this case, $v \mapsto \theta(v)$ is strictly increasing on $[-2, 1]$ and is flat on $[1, 2]$, and $V_0 = [1, 1.5]$ is a strict subset of $\mathcal{V} = [-2, 1.5]$. For DGP7, we consider

$$\varphi_0(v) = -2v^2.$$

In this case, $v \mapsto \theta_l(v)$ has a unique maximum at $v = 0$, and thus, $V_0 = \{0\}$ is singleton. For DGP8, we consider

$$\varphi_0(v) = -10v^2.$$

In this case, $v \mapsto \theta(v)$ has a unique maximum at $v = 0$ and is more peaked than that of DGP7. Figures 3 and 4 show data realizations and bounding functions for these DGPs.

DGPs 9-12 use the bounding functions in Section 10.3 of AS. DGP9 and DGP10 feature a roughly plateau-shaped bounding function given by

$$\theta(v) := L\phi(v^{10}), \tag{L.2}$$

instead of $\theta(v) := L\phi(v)$ as in Section 7 of the main text. DGP11 and DGP12 use the roughly double-plateau-shaped bounding function

$$\theta(v) := L \cdot \max \{ \phi((v - 1.5)^{10}), \phi((v + 1.5)^{10}) \}. \quad (\text{L.3})$$

Specifically, we generated 1000 independent samples from the model:

$$V_i \sim \text{Unif}[-2, 2], U_i = \min\{\max\{-3, \sigma\tilde{U}_i\}, 3\}, \text{ and } Y_i = \theta(V_i) + U_i,$$

where $\tilde{U}_i \sim N(0, 1)$, with L and σ as follows:

$$\text{DGP9 and DGP11: } L = 1 \text{ and } \sigma = 0.1; \quad \text{DGP10 and DGP12: } L = 5 \text{ and } \sigma = 0.1.$$

Figure 5 illustrates the bounding function and data realizations for DGPs 9 and 10; figure 6 provides the same for DGPs 11 and 12. Interest again lies in inference on $\theta_0 = \sup_{v \in \mathcal{V}} \theta(v)$, which in these DGPs is $\theta_0 = L\phi(0)$.

L.2. Simulation Results. To evaluate the relative performance of our inference method in DGPs 5-8, we used our method with cubic B-splines with knots equally spaced over the sample quantiles of V_i , and we also implemented one of the inference methods proposed by AS, specifically their Cramér-von Mises-type (CvM) statistic with PA/Asy and GMS/Asy critical values. Implementation details for our method with B-splines are the same as in Section 7.2 of the main text. Turning parameters for CvM were chosen exactly as in AS (see Section 9).³⁵ We considered sample sizes $n = 250$, $n = 500$, and $n = 1000$.

The coverage probability is evaluated at the true lower bound θ_0 (with the nominal level of 95%), and the false coverage probability (FCP) is evaluated at a θ value outside the identified set. For DGP5, we set $\theta = \theta_0 - 0.03$; for DGP6-DGP7, $\theta = \theta_0 - 0.05$; and for DGP8, $\theta = \theta_0 - 0.07$. These points are chosen differently across different DGPs to ensure that the FCPs have similar values. This type of FCP was reported in AS, along with a so-called “CP-correction” (similar to size correction in testing). We did not do CP-correction in our reported results. There were 1,000 replications for each experiment. Table 1 summarizes the results of Monte Carlo experiments. CLR and AS refer to our inference method and that of AS, respectively.

First, we consider Monte Carlo results for DGP5. The discrepancies between nominal and actual coverage probabilities are not large across all methods, implying that all of them perform well in finite samples. For DGP5, since the true argmax set V_0 is equal to \mathcal{V} , an

³⁵In DGPs 5-8 our Monte Carlo design differs from that of AS, and alternative choices of tuning parameters could perform more or less favorably in our design. We did not examine sensitivity to the choice of tuning parameters for their method.

estimated V_0 should be the entire set \mathcal{V} . Thus the simulation results are the same whether or not estimating V_0 since for most of simulation draws, $\widehat{V}_n = \mathcal{V}$. Similar conclusions hold for AS with CvM between PA/Asy and GMS/Asy critical values. In terms of false coverage probability, CvM with either critical value performs better than our method.

We now move to DGPs 6-8. In DGP6, the true argmax set V_0 is $[1, 1.5]$ and in DGP7 and DGP8, V_0 is a singleton set. In these cases the true argmax set V_0 is a strict subset of \mathcal{V} . Hence, we expect that it is important to estimate V_0 . On average, for DGP6, the estimated sets were $[-0.793, 1.5]$ when $n = 250$, $[-0.359, 1.5]$ when $n = 500$, and $[-0.074, 1.5]$ when $n = 1,000$; for DGP7, the estimated sets were $[-0.951, 0.943]$ when $n = 250$, $[-0.797, 0.798]$ when $n = 500$, and $[-0.684, 0.680]$ when $n = 1,000$; for DGP8, the estimated sets were $[-1.197, 0.871]$ when $n = 250$, $[-0.662, 0.645]$ when $n = 500$, and $[-0.403, 0.402]$ when $n = 1,000$.

Hence, an average estimated set is larger than V_0 ; however, it is still a strict subset of \mathcal{V} and gets smaller as n gets large. For all the methods, the Monte Carlo results are consistent with asymptotic theory. Unlike in DGP5, the CLR method performs better than the AS method in terms of false coverage probability.³⁶ As can be seen from the table, the CLR method performs better when V_0 is estimated in terms of making the coverage probability less conservative and also of making the false coverage probability smaller. Similar gains are obtained for the CvM with GMS/Asy critical values, relative to that with PA/Asy critical values.

We now turn to DGPs 9-12. AS, Section 10.3, report results for their approach using their CvM and KS statistics, and we refer the reader to their paper for results using their method. They also include results for our approach using B-splines and local-linear estimation of the bounding function. Here we provide further investigation of the performance of our method in additional Monte Carlo simulations.

From Figures 5 and 6 we see that the bounding function is very nearly flat in some regions, including areas close to V_0 , and also has very large derivatives a bit further away from V_0 . The functions are smooth, but the steep derivatives mimic discontinuity points, and are challenging for nonparametric estimation methods. The AS approach does not rely on smoothness of the bounding function, and performs better in most - though not all - of the comparisons of AS. The performance of our approach improves with the sample size, as expected.

³⁶As in Section 7, this conclusion will remain valid even with CP-correction as in AS, since our method performs better in DGPs 6-8 where we have over-coverage.

Our Monte Carlo experiments for DGPs 9-12 further examine the performance of our method in such a setup. In all experiments we report coverage probabilities (CPs) at $\theta_0 = L\phi(0)$ and FCPs at $\theta_0 - 0.02$ as in AS. We provide results for sample sizes $n = 500$ and $n = 1000$, both with and without estimation of the contact set. In Table 2 we report the results of series estimation via B-splines. We used cubic B-splines and our implementation was identical to that described in Section 7 of the main text. Compared to the results in Table 1 of the main text for DGPs 1-4, we see that the average number of series terms is much higher. This is due to a higher number of terms selected during cross-validation, presumably because of the regions with very large derivatives. Our coverage probabilities are below the nominal level, but they improve with the sample size, as in AS, ranging from .885 to .937 across DGPs at $n = 1000$. Moreover, we see that our method using $\mathbf{V} = \mathcal{V}$ rather than $\mathbf{V} = \widehat{V}_n$ actually performs better in this setup.

To further investigate the challenge of nonparametrically estimating a bounding function with steep derivatives, we implemented our method with a locally-constant Nadaraya-Watson kernel estimator. The functions are in fact nearly locally constant at most points, with the exception of the relatively narrow regions with steep derivatives. The top half of Table 3 presents the results with a bandwidth selected the same way as for the local-linear estimator in Section 7.3, equation (7.3). When V_n is estimated coverage probabilities in these DGPs range from .903 to .923 when $n = 500$ and .926 to .945 when $n = 1000$, becoming closer to the nominal level. The procedure exhibits good power in all cases, with FCPs decreasing with the sample size. These results are qualitatively similar to those reported in AS for the local linear estimator. We also include results when \mathcal{V} is used instead of estimating the contact set. This results in higher coverage probabilities for θ_0 , in most cases closer to the nominal level, but also somewhat higher FCPs. Overall performance remains reasonably good.

The bottom half of Table 3 gives results for locally-constant kernel estimation using an identical rule of thumb for bandwidth selection, $h = \widehat{h}_{ROT} \times \widehat{s}_v$, but without applying an undersmoothing factor. The proofs of our asymptotic results use undersmoothing, but with a locally-constant kernel estimator this does not appear essential. Our exploratory Monte Carlo results are very good, offering support to that view. Compared to the results with undersmoothing, all coverage probabilities increase, and all FCPs decrease. This suggests that future research on the possibility to abandon undersmoothing may be warranted.

The overall results of this section support the conclusions reached in Section 7 of the main text regarding comparisons to AS. In completely flat cases, the AS method outperforms our method, whereas in non-flat cases, our method outperforms the AS method. In this

section we also considered some intermediate cases. In DGP7, where the bounding function is partly-flat, our method performed favorably. More generally there is a wide range of intermediate cases that could be considered, and we would expect the approach of AS to perform favorably in some cases too. Indeed, in DGPs 9-12 from AS, the bounding function exhibits areas with extremely steep derivatives. Their results indicate that in these DGPs their approach performs better at smaller sample sizes ($n = 100, 250$) than does our approach, which is based on nonparametric estimation methods. However, at larger sample sizes ($n = 500, 1000$) even with the very steep derivatives of DGPs 9-12 our approach performs well, and in a handful of cases, e.g. DGP10 with kernel estimation, can even perform favorably. The main conclusions we draw from the full range of Monte Carlo experiments are that our inference method generally performs well both in coverage probabilities and false coverage probabilities and that in terms of a comparison between our approach and that of AS, each has their relative advantages and neither approach dominates.

TABLE 1. Results for Monte Carlo Experiments

DGP	Sample Size	Critical Value	Cov. Prob.	False Cov. Prob.
CLR with Series Estimation using B-splines				
		Estimating V_n ?		
5	250	No	0.914	0.709
5	250	Yes	0.914	0.709
5	500	No	0.935	0.622
5	500	Yes	0.935	0.622
5	1000	No	0.947	0.418
5	1000	Yes	0.947	0.418
6	250	No	0.953	0.681
6	250	Yes	0.942	0.633
6	500	No	0.967	0.548
6	500	Yes	0.941	0.470
6	1000	No	0.973	0.298
6	1000	Yes	0.957	0.210
7	250	No	0.991	0.899
7	250	Yes	0.980	0.841
7	500	No	0.996	0.821
7	500	Yes	0.994	0.697
7	1000	No	0.987	0.490
7	1000	Yes	0.965	0.369
8	250	No	0.999	0.981
8	250	Yes	0.996	0.966
8	500	No	1.000	0.984
8	500	Yes	0.999	0.951
8	1000	No	0.998	0.909
8	1000	Yes	0.995	0.787
AS with CvM (Cramér-von Mises-type statistic)				
5	250	PA/Asy	0.951	0.544
5	250	GMS/Asy	0.945	0.537
5	500	PA/Asy	0.949	0.306
5	500	GMS/Asy	0.945	0.305
5	1000	PA/Asy	0.962	0.068
5	1000	GMS/Asy	0.956	0.068
6	250	PA/Asy	1.000	0.941
6	250	GMS/Asy	0.990	0.802
6	500	PA/Asy	1.000	0.908
6	500	GMS/Asy	0.980	0.674
6	1000	PA/Asy	1.000	0.744
6	1000	GMS/Asy	0.980	0.341
7	250	PA/Asy	1.000	1.000
7	250	GMS/Asy	0.997	0.948
7	500	PA/Asy	1.000	0.997
7	500	GMS/Asy	0.997	0.916
7	1000	PA/Asy	1.000	0.993
7	1000	GMS/Asy	0.997	0.823
8	250	PA/Asy	1.000	1.000
8	250	GMS/Asy	1.000	0.988
8	500	PA/Asy	1.000	1.000
8	500	GMS/Asy ₃₈	0.999	0.972
8	1000	PA/Asy	1.000	1.000
8	1000	GMS/Asy	1.000	0.942

Notes: CLR and AS refer to our inference methods and those of Andrews and Shi (forthcoming), respectively. There were 1000 replications per experiment.

TABLE 2. Results for Monte Carlo Experiments (Series Estimation Using B-splines)

DGP	Sample Size	Critical Value	Ave. Smoothing Parameter	Cov. Prob.	False Cov. Prob.	Ave. Argmax Set Min.	Max.
CLR with Series Estimation Using B-splines							
Estimating V_n ?							
9	500	No	35.680	0.920	0.562	-1.799	1.792
9	500	Yes	35.680	0.870	0.475	-1.001	1.001
9	1000	No	39.662	0.937	0.487	-1.801	1.797
9	1000	Yes	39.662	0.913	0.380	-0.977	0.977
10	500	No	39.090	0.887	0.534	-1.799	1.792
10	500	Yes	39.090	0.825	0.428	-0.912	0.912
10	1000	No	41.228	0.920	0.477	-1.801	1.797
10	1000	Yes	41.228	0.891	0.351	-0.902	0.903
11	500	No	35.810	0.880	0.462	-1.799	1.792
11	500	Yes	35.810	0.853	0.399	-1.799	1.792
11	1000	No	40.793	0.937	0.374	-1.801	1.797
11	1000	Yes	40.793	0.912	0.299	-1.801	1.797
12	500	No	39.474	0.836	0.459	-1.799	1.792
12	500	Yes	39.474	0.811	0.386	-1.799	1.792
12	1000	No	42.224	0.917	0.367	-1.801	1.797
12	1000	Yes	42.224	0.885	0.294	-1.801	1.797

Notes: DGPs 9-12 correspond to DGPs 1-4 in Andrews and Shi (forthcoming) Section 10.3. The last two columns report the average values of the minimum and maximum of the argmax set. The estimated set is allowed to be disconnected and so the interval between the minimum and maximum of the argmax set is just an outer set for the estimated argmax set.

TABLE 3. Results for Monte Carlo Experiments (Nadaraya-Watson Kernel Estimation)

DGP	Sample Size	Critical Value	Ave. Smoothing Parameter	Cov. Prob.	False Cov. Prob.	Ave. Argmax Set Min.	Max.
Using Bandwidth $h = \hat{h}_{ROT} \times \hat{s}_v \times n^{1/5} \times n^{-2/7}$							
Estimating V_n ?							
9	500	No	0.206	0.944	0.496	-1.799	1.792
9	500	Yes	0.206	0.916	0.391	-0.955	0.954
9	1000	No	0.169	0.966	0.326	-1.801	1.796
9	1000	Yes	0.169	0.945	0.219	-0.941	0.940
10	500	No	0.166	0.945	0.523	-1.799	1.792
10	500	Yes	0.166	0.903	0.411	-0.879	0.880
10	1000	No	0.136	0.963	0.387	-1.801	1.796
10	1000	Yes	0.136	0.926	0.266	-0.868	0.868
11	500	No	0.195	0.938	0.403	-1.799	1.792
11	500	Yes	0.195	0.923	0.345	-1.799	1.792
11	1000	No	0.160	0.951	0.201	-1.801	1.796
11	1000	Yes	0.160	0.926	0.152	-1.801	1.796
12	500	No	0.169	0.937	0.439	-1.799	1.792
12	500	Yes	0.169	0.917	0.365	-1.799	1.792
12	1000	No	0.138	0.947	0.235	-1.801	1.796
12	1000	Yes	0.138	0.933	0.176	-1.801	1.796
Using Bandwidth $h = \hat{h}_{ROT} \times \hat{s}_v$ (no undersmoothing factor)							
Estimating V_n ?							
9	500	No	0.351	0.968	0.470	-1.799	1.792
9	500	Yes	0.351	0.944	0.360	-0.908	0.907
9	1000	No	0.306	0.977	0.211	-1.801	1.796
9	1000	Yes	0.306	0.956	0.138	-0.885	0.884
10	500	No	0.283	0.959	0.520	-1.799	1.792
10	500	Yes	0.283	0.931	0.409	-0.839	0.839
10	1000	No	0.247	0.977	0.285	-1.801	1.796
10	1000	Yes	0.247	0.959	0.188	-0.828	0.828
11	500	No	0.333	0.955	0.316	-1.799	1.792
11	500	Yes	0.333	0.939	0.261	-1.799	1.792
11	1000	No	0.290	0.960	0.118	-1.801	1.796
11	1000	Yes	0.290	0.943	0.079	-1.801	1.796
12	500	No	0.287	0.956	0.376	-1.799	1.792
12	500	Yes	0.287	0.944	0.295	-1.799	1.792
12	1000	No	0.250	0.960	0.154	-1.801	1.796
12	1000	Yes	0.250	0.948	0.111	-1.801	1.796

Notes: DGPs 9-12 correspond to DGPs 1-4 in Andrews and Shi (forthcoming) Section 10.3. The last two columns report the average values of the minimum and maximum of the argmax set. The estimated set is allowed to be disconnected and so the interval between the minimum and maximum of the argmax set is just an outer set for the estimated argmax set.

FIGURE 1. Simulated Data and Bounding Functions: DGP1 and DGP2

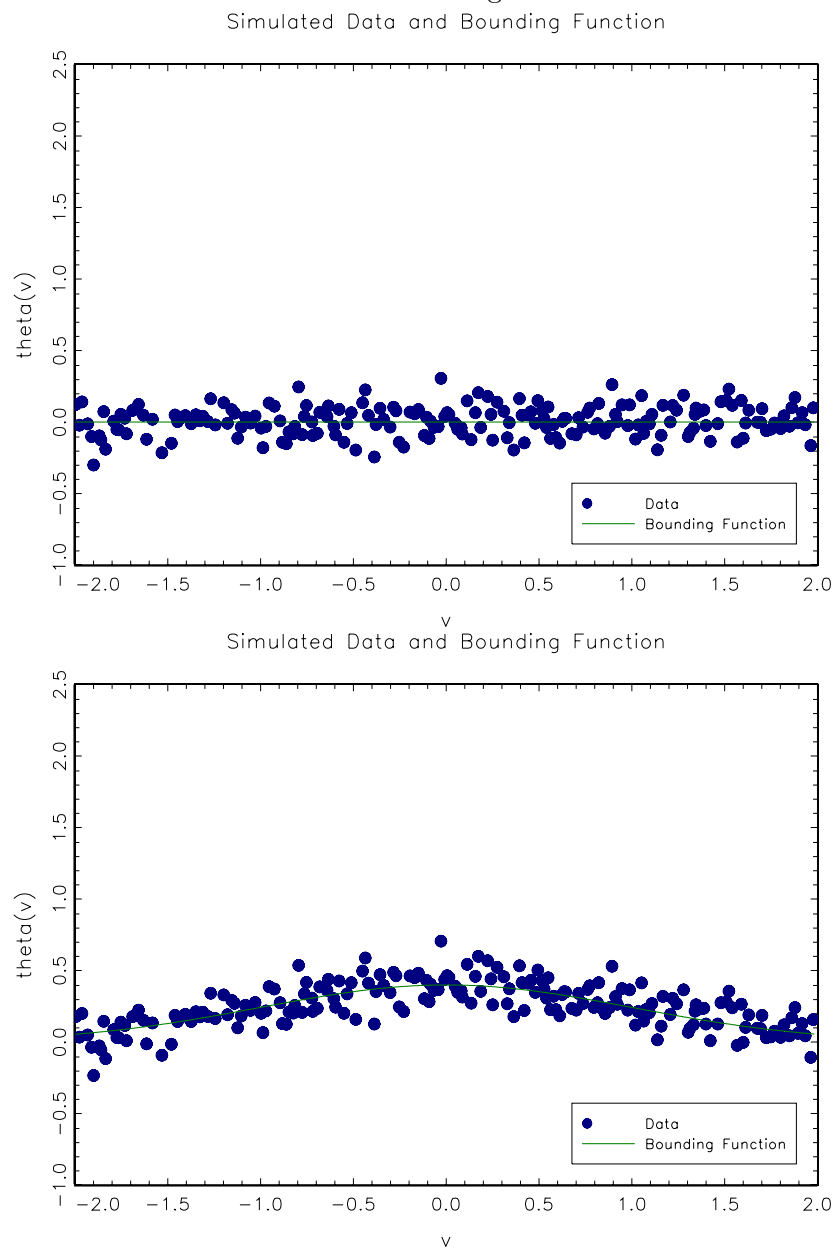


FIGURE 2. Simulated Data and Bounding Functions: DGP3 and DGP4

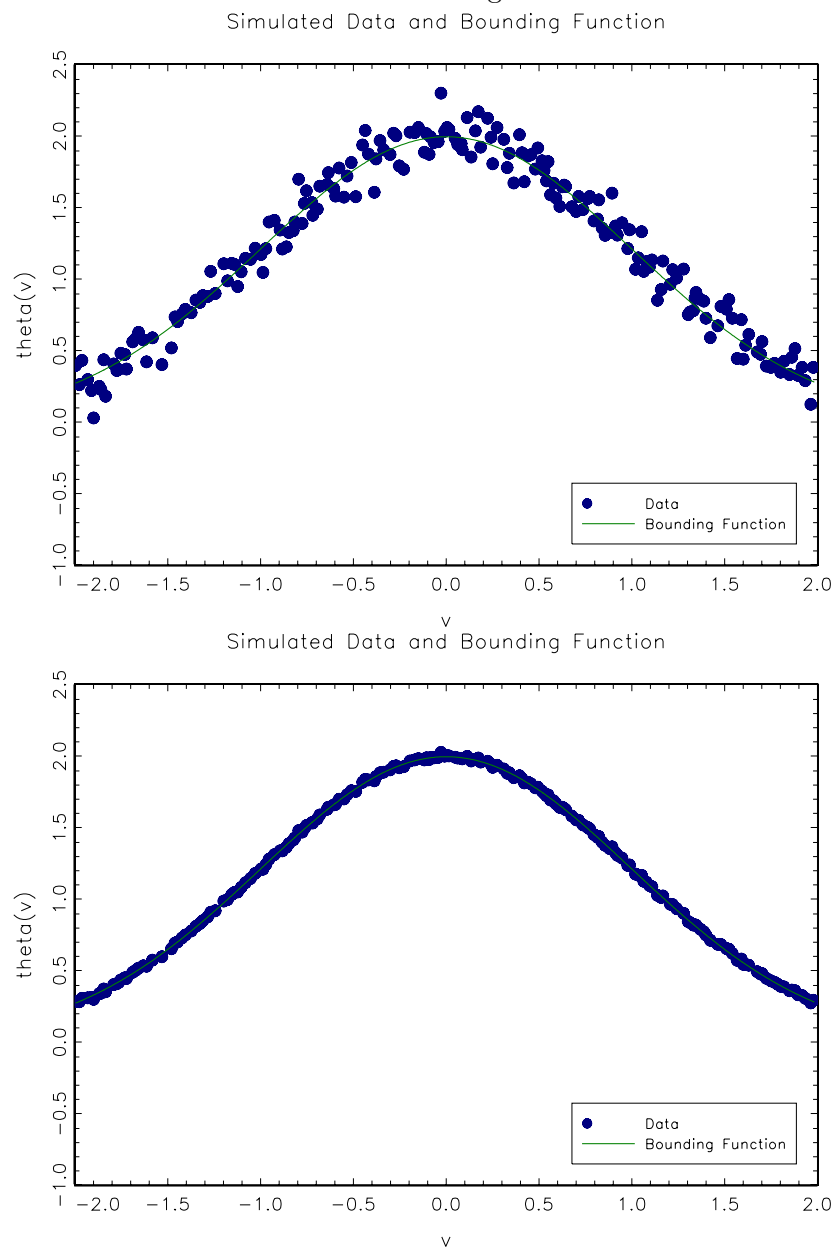


FIGURE 3. Simulated Data and Bounding Functions: DGP5 and DGP6

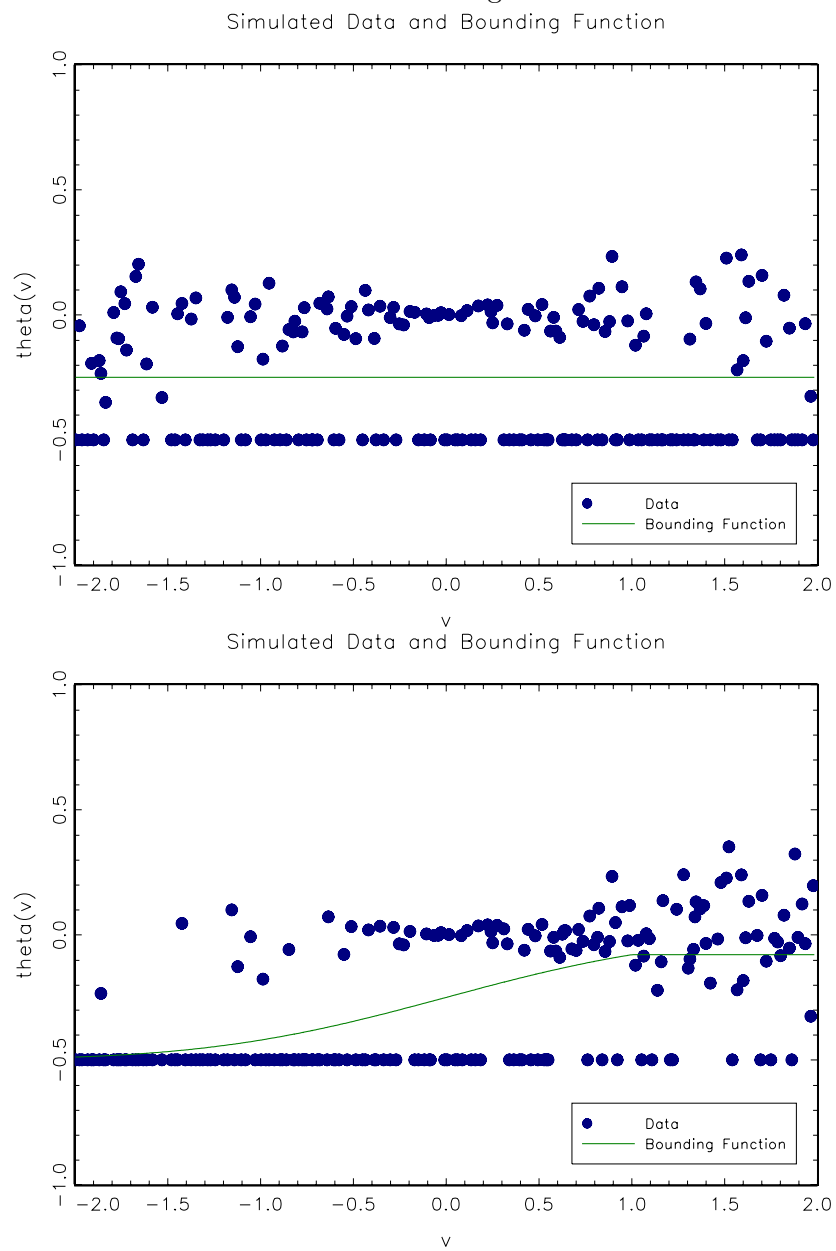


FIGURE 4. Simulated Data and Bounding Functions: DGP7 and DGP8
 Simulated Data and Bounding Function

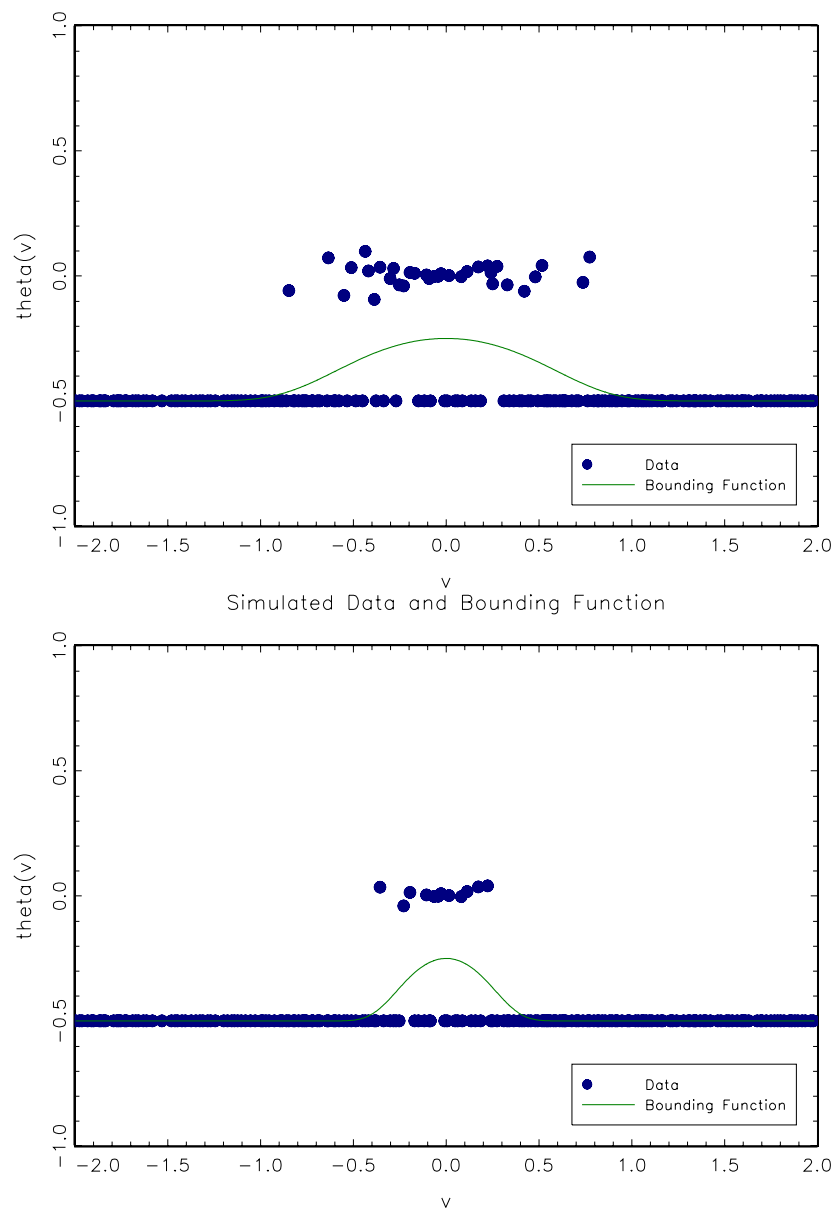


FIGURE 5. Simulated Data and Bounding Functions: DGP9 and DGP10
 Simulated Data and Bounding Function

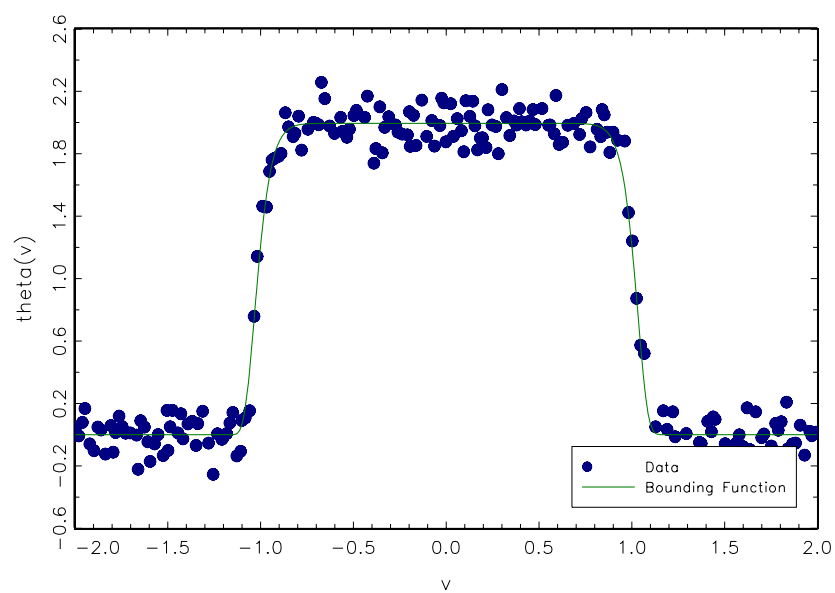
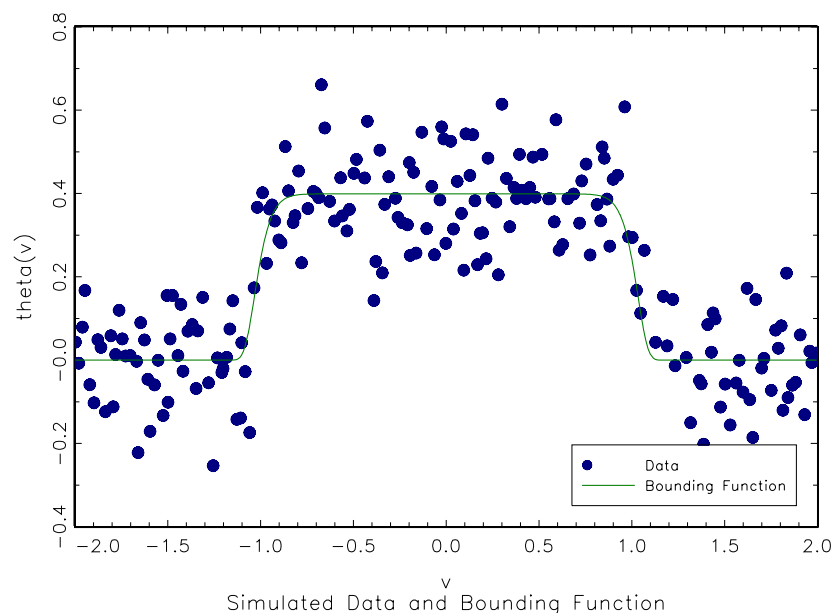
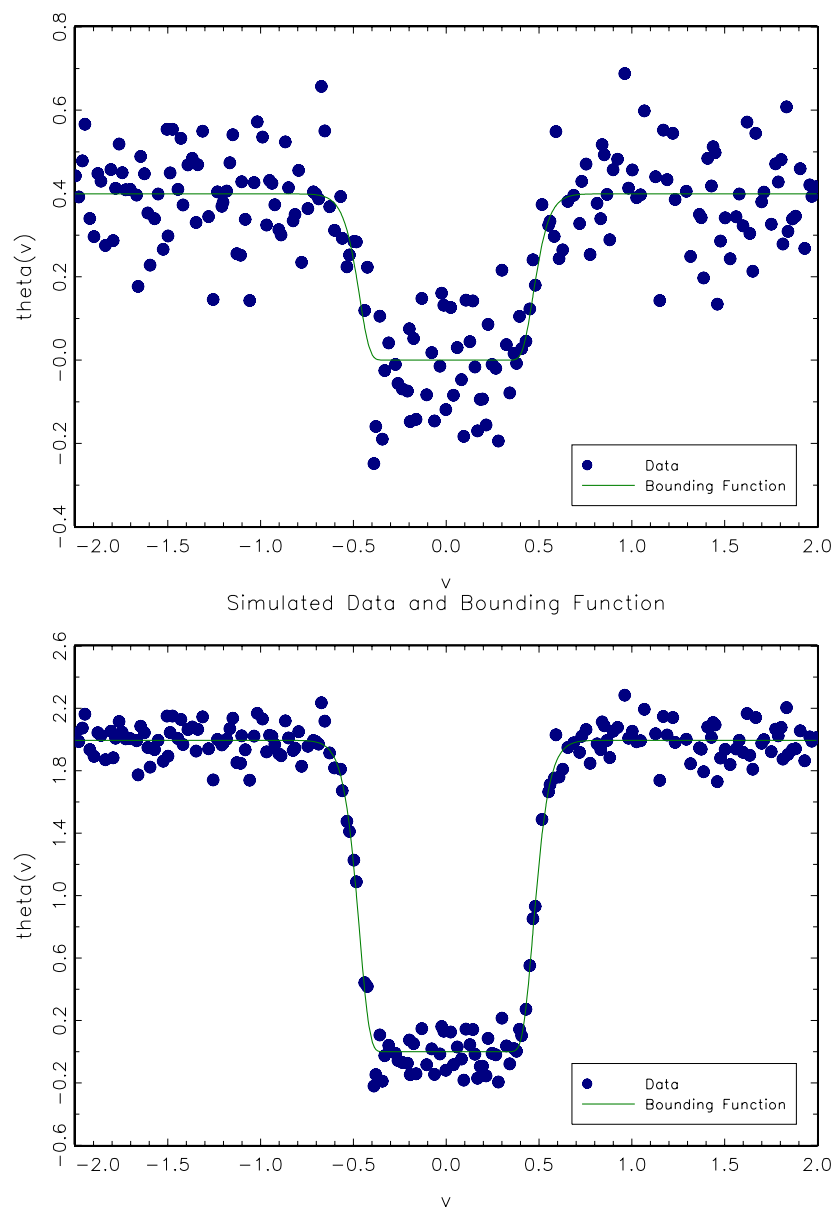


FIGURE 6. Simulated Data and Bounding Functions: DGP11 and DGP12
 Simulated Data and Bounding Function



REFERENCES

- ANDREWS, D. W. K., AND X. SHI (forthcoming): “Inference for Parameters Defined by Conditional Moment Inequalities,” *Econometrica*.
- BELLONI, A., X. CHEN, AND V. CHERNOZHUKOV (2010): “(Some) New Asymptotic Theory for Linear Series Estimators,” working paper, MIT.
- BELLONI, A., V. CHERNOZHUKOV, AND I. FERNANDEZ-VAL (2011): “Conditional Quantile Processes under Increasing Dimensions,” CeMMAP working paper CWP19/11.
- CHEN, X. (2007): “Large Sample Sieve Estimation of Semi-Nonparametric Models (Chapter 76),” vol. 6, Part 2 of *Handbook of Econometrics*, pp. 5549 – 5632. Elsevier.
- CHERNOZHUKOV, V., S. LEE, AND A. ROSEN (2012): “Intersection Bounds: Estimation and Inference,” CeMMAP working paper.
- CHETVERIKOV, D. (2011): “Adaptive rate optimal test of conditional moment inequalities,” working paper, MIT.
- FAN, J., AND I. GIJBELS (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London, UK.
- GHOSAL, S., A. SEN, AND A. W. VAN DER VAART (2000): “Testing Monotonicity of Regression,” *Annals of Statistics*, 28, 1054–1082.
- HOROWITZ, J. L., AND E. MAMMEN (2004): “Nonparametric estimation of an additive model with a link function,” *Ann. Statist.*, 32(6), 2412–2443.
- KONG, E., O. LINTON, AND Y. XIA (2010): “Uniform Bahadur representation for local polynomial estimates of m-regression and its application to the additive model,” *Econometric Theory*, 26, 1529–1564.
- MANSKI, C. F., AND J. V. PEPPER (2009): “More on Monotone Instrumental Variables,” *Econometrics Journal*, 12(s1), S200–S216.
- MASRY, E. (1996): “Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates,” *Journal of Time Series Analysis*, 17, 571–599.
- MASSART, P. (1989): “Strong Approximation for Multivariate Empirical and Related Processes, via KMT Constructions,” *Annals of Probability*, 17(1), 266–291.
- NEWBY, W. K. (1997): “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, 79, 147–168.
- POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer, New York, NY.
- RIO, E. (1994): “Local Invariance Principles and Their Application to Density Estimation,” *Probability Theory and Related Fields*, 98, 21–45.
- RUDELSON, M. (1999): “Random vectors in the isotropic position,” *Journal of Functional Analysis*, 164(1), 60–72.
- VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY.