

Uniform inference for value functions

Sergio Firpo*

Antonio F. Galvao[†]

Thomas Parker[‡]

October 25, 2022

Abstract

We propose a method to conduct uniform inference for the (optimal) value function, that is, the function that results from optimizing an objective function marginally over one of its arguments. Marginal optimization is not Hadamard differentiable (that is, compactly differentiable) as a map between the spaces of objective and value functions, which is problematic because standard inference methods for nonlinear maps usually rely on Hadamard differentiability. However, we show that the map from objective function to an L_p functional of a value function, for $1 \leq p \leq \infty$, are Hadamard directionally differentiable. As a result, we establish consistency and weak convergence of nonparametric plug-in estimates of Cramér-von Mises and Kolmogorov-Smirnov test statistics applied to value functions. For practical inference, we develop detailed resampling techniques that combine a bootstrap procedure with estimates of the directional derivatives. In addition, we establish local size control of tests which use the resampling procedure. Monte Carlo simulations assess the finite-sample properties of the proposed methods and show accurate empirical size and nontrivial power of the procedures. Finally, we apply our methods to the evaluation of a job training program using bounds for the distribution function of treatment effects.

Keywords: uniform inference, value function, bootstrap, delta method, directional differentiability

JEL codes: C12, C14, C21

*Insper, Sao Paulo, Brazil. firmo@insper.edu.br

[†]Department of Economics, Michigan State University, East Lansing, USA. agalvao@msu.edu

[‡]Corresponding author. Department of Economics, University of Waterloo, Waterloo, Canada.
tmparker@uwaterloo.ca

1 Introduction

The (optimal) value function, that is, a function defined by optimizing an objective function over one of multiple arguments, is used widely in economics, finance, statistics and other fields. The statistical behavior of estimated value functions appears, for instance, in the study of distribution and quantile functions of treatment effects under partial identification (see, e.g., Firpo and Ridder (2008, 2019) and Fan and Park (2010, 2012)). More recently, investigating counterfactual sensitivity, Christensen and Connault (2019) construct interval-identified counterfactual outcomes as the distribution of unobservables varies over a nonparametric neighborhood of an assumed model specification. Statistical inference for value functions can be analyzed at a point as in Römisch (2006) or Cárcamo, Cuevas, and Rodríguez (2020) and inference can be carried out using the techniques of Hong and Li (2018) or Fang and Santos (2019).

This paper extends these pointwise inference methods to uniform inference for value functions that are estimated with nonparametric plug-in estimators. Uniform inference has an important role for the analysis of statistical models.¹ However, uniform inference for the value function can be quite challenging. In order to use standard statistical inference procedures, one must show that the map from objective function to value function is Hadamard (i.e., compactly) differentiable as a map between two functional spaces, which may require restrictive regularity conditions. As a result, the delta method, as the conventional tool used to analyze the distribution of nonlinear transformations of sample data, may not apply to marginal optimization as a map between function spaces due to a lack of differentiability. We propose a solution to this problem that allows for the use of the delta method to conduct uniform inference for value functions.

Because in general the map of objective function to value function is not differentiable, our solution is to directly analyze statistics that are applied to the value function. This is in place of a more conventional analysis that would first establish that the value function is well-behaved and as a second step use the continuous mapping theorem to determine the distribution of test statistics. In settings involving a chain of maps, this can be seen as choosing how to define the “links” in the chain to which the chain rule applies. The results are presented for L_p functionals for $1 \leq p \leq \infty$, that are applied to a value function, which should cover many cases of interest. In particular, this family of functionals includes Kolmogorov-Smirnov and Cramér-von Mises tests.

By considering L_p functionals of the value function, we may bypass the most serious impediment to uniform inference. However, these L_p functionals are only Hadamard directionally differentiable. A directional generalization of Hadamard (or compact) differentiability was developed for statistical

¹For example, Angrist, Chernozhukov, and Fernández-Val (2006) discuss the importance of uniform inference *vis-à-vis* pointwise inference in a quantile regression framework. Uniform testing methods allow for multiple comparisons across the entire distribution without compromising confidence levels, and uniform confidence bands differ from corresponding pointwise confidence intervals because they guarantee coverage over the family of confidence intervals simultaneously.

applications in Shapiro (1990) and Dümbgen (1993). There is a more recent growing literature of applications using Hadamard directional differentiability. See, for example, Sommerfeld and Munk (2017), Chetverikov, Santos, and Shaikh (2018), Cho and White (2018), Hong and Li (2018), Fang and Santos (2019), Christensen and Connault (2019), Masten and Poirier (2020), and Cattaneo, Jansson, and Nagasawa (2020). We separately analyze L_p functionals for $1 \leq p < \infty$ and the supremum norm, as well as one-sided variants used for tests that have a sense of direction. As an intermediate step towards showing the differentiability of supremum-norm statistics, we show directional differentiability of a minimax map applied to functions that are not necessarily convex-concave.

Directional differentiability of L_p functionals provides the minimal amount of regularity needed for feasible statistical inference. We use directional differentiability to find the asymptotic distributions of test statistics estimated from sample data. The distributions generally depend on features of the objective function, and we show how to combine estimation with resampling to estimate the limiting distributions, as was proposed by Fang and Santos (2019). This technique can be tailored to impose a null hypothesis and is simple to implement in practice. We provide three examples throughout the paper to illustrate the proposed methods — dependency bounds, stochastic dominance, and cost and profit functions. Moreover, we establish local size control of the resampling procedure in general and uniform size control for some functionals.

As a practical illustration of the developed framework, we consider bounds for the distribution function of a treatment effect. We use this model in both the Monte Carlo simulations and empirical application. Treatment effects models have provided a valuable analytic framework for the evaluation of causal effects in program evaluation studies. When the treatment effect is heterogeneous, its distribution function and related features, for example its quantile function, are often of interest. Nevertheless, it is well known (e.g. Heckman, Smith, and Clements (1997); Abbring and Heckman (2007)) that features of the distribution of treatment effects beyond the average are not point identified unless one imposes significant theoretical restrictions on the joint distribution of potential outcomes. Therefore it has become common to use partial identification of the distribution of such features, and to derive bounds on their distribution without making any unwarranted assumptions about the copula that connects the marginal outcome distributions.²

Monte Carlo simulations (reported in a supplementary appendix) assess the finite-sample properties of the proposed methods. The simulations suggest that Kolmogorov-Smirnov statistics used to construct uniform confidence bands have accurate empirical size and power against local alternatives. A Cramér-von Mises type statistic is used to test a stochastic dominance hypothesis using bound functions, similar to the breakdown frontier tests of Masten and Poirier (2020). This

²Bounds for the CDF and quantile functions at one point in the support of the treatment effects distribution were developed in Makarov (1982); Rüschendorf (1982); Frank, Nelsen, and Schweizer (1987); Williamson and Downs (1990). These bounds are minimal in the sense that they rely only on the marginal distribution functions of the observed outcomes, not the joint distribution of (potential) outcomes.

test has accurate size at the breakdown frontier and power against alternatives that violate the dominance relationship. All results are improved when the sample size increases and are powerful using a modest number of bootstrap repetitions.

We illustrate the methods empirically using job training program data set first analyzed by LaLonde (1986) and subsequently by many others, including Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2001, 2005), Imbens (2003), and Firpo (2007), without making any assumptions on dependence between potential outcomes in the experiment. This experimental data set has information from the National Supported Work Program used by Dehejia and Wahba (1999). We document strong heterogeneity of the job training treatment across the distribution of earnings. The uniform confidence bands for the treatment effect distribution function are imprecisely estimated at some parts of the earnings distribution. These large uniform confidence bands may in part be attributed to the large number of zeros contained in the data set, but are also inherent to the fact that the distribution function is everywhere only partially identified.

The remainder of the paper is organized as follows. Section 2 defines the statistical model of interest for the value function. Section 3 establishes directional differentiability for the functionals of interest. Inference procedures are established in Section 4. An empirical application to job training is discussed in Section 5. Finally, Section 6 concludes. Monte Carlo simulation descriptions and results and all proofs are relegated to the supplementary appendix.

Notation

For any set $T \subseteq \mathbb{R}^d$ let $\ell^\infty(T)$ denote the space of bounded functions $f : T \rightarrow \mathbb{R}$ and let $\mathcal{C}(T)$ denote the space of continuous functions $f : T \rightarrow \mathbb{R}$, both equipped with the uniform norm $\|f\|_\infty = \sup_{t \in T} |f(t)|$. Given a sequence $\{f_n\}_n \subset \ell^\infty(T)$ and limiting random element f we write $f_n \rightsquigarrow f$ to denote weak convergence in $(\ell^\infty(T), \|\cdot\|_\infty)$ in the sense of Hoffmann-Jørgensen (van der Vaart and Wellner, 1996). Let $[x]_+ = \max\{x, 0\}$ and $[x]_- = \min\{x, 0\}$. A set-valued map (or correspondence) S that maps elements of X to the collection of subsets of Y is denoted $S : X \rightrightarrows Y$. For set-valued map $S : X \rightrightarrows Y$, let $\text{gr}S$ denote the graph of S as a set in $X \times Y$.

2 The model

Consider two sets $U \subseteq \mathbb{R}^{d_U}$ and $X \subseteq \mathbb{R}^{d_X}$, a set-valued map $A : X \rightrightarrows U$ that serves as a choice set, and an objective function $f \in \ell^\infty(\text{gr}A)$. The objective function and the value function are linked by a marginal optimization step that maps one functional space to another. Let $\psi : \ell^\infty(\text{gr}A) \rightarrow \ell^\infty(X)$ map the objective function to the value function obtained by marginally optimizing the objective f with respect to $u \in A(x)$ for each $x \in X$. Without loss of generality, we consider only marginal

maximization with respect to u for each x :

$$\psi(f)(x) = \sup_{u \in A(x)} f(u, x). \quad (1)$$

The value function $\psi(f)$ is the object of interest and we would like to conduct uniform inference using a plug-in estimator $\psi(f_n)$. First we introduce examples of the value function model represented in equation (1).

Example A (Dependency bounds). Dependency bounds describe bounds on the distribution function of the sum of two random variables X and Y , F_{X+Y} , using only the marginal distribution functions F_X and F_Y and no information from the joint distribution F_{XY} . Frank, Nelsen, and Schweizer (1987) showed that bounds derived for F_{X+Y} by Makarov (1982) are related to copula bounds and can be extended to other binary operations. Upper and lower dependency bounds for the CDFs F_{X+Y} , F_{X-Y} , $F_{X \cdot Y}$ and $F_{X/Y}$ and their associated quantile functions are provided in Williamson and Downs (1990).

Williamson and Downs (1990) show that for any x , $L_{X+Y}(x) \leq F_{X+Y}(x) \leq U_{X+Y}(x)$, where

$$L_{X+Y}(x) = \left[\sup_{u \in \mathbb{R}} (F_X(u) + F_Y(x - u) - 1) \right]_+ \quad (2)$$

$$U_{X+Y}(x) = 1 - \left[\inf_{u \in \mathbb{R}} (F_X(u) + F_Y(x - u)) \right]_- . \quad (3)$$

Similarly, letting f^{-1} denote the generalized inverse of an increasing function, bounds for the quantile function $F_{X+Y}^{-1}(\tau)$ may be considered for $0 < \tau < 1$. The functions above may be inverted to find that for each quantile level τ , $U_{X+Y}^{-1}(\tau) \leq F_{X+Y}^{-1}(\tau) \leq L_{X+Y}^{-1}(\tau)$, and the quantile function bounds are

$$U_{X+Y}^{-1}(\tau) = \sup_{u \in (0, \tau)} \{F_X^{-1}(u) + F_Y^{-1}(\tau - u)\} \quad (4)$$

$$L_{X+Y}^{-1}(\tau) = \inf_{u \in (\tau, 1)} \{F_X^{-1}(u) + F_Y^{-1}(\tau - u + 1)\} . \quad (5)$$

The functions U_{X+Y}^{-1} and L_{X+Y}^{-1} are value functions where marginal optimization takes place over a set-valued map that varies with τ .

All these related functions may be written as maps that depend on a value function in some part of a chain of maps. For example, take L_{X+Y} . For a pair of functions $f = (f_x, f_y) \in (\ell^\infty(\mathbb{R}))^2$, let $\Pi^+ : (\ell^\infty(\mathbb{R}))^2 \rightarrow \ell^\infty(\mathbb{R}^2)$

$$\Pi^+(f)(s, t) = f_x(s) + f_y(t - s) - 1 \quad (6)$$

and let ψ be as in (1) with $A(x) = \mathbb{R}$ for all x . Then $L_{X+Y} = \psi(\Pi^+(F) \vee 0)$. Similarly, consider

U_{X+Y}^{-1} . For a pair of CDFs $f = (f_x, f_y)$ where $f_x, f_y : \mathbb{R} \rightarrow [0, 1]$, define $\tilde{\Pi}^+ : (\ell^\infty(\mathbb{R}))^2 \rightarrow \ell^\infty([0, 1]^2)$ by³

$$\tilde{\Pi}^+(f)(s, t) = f_x^{-1}(s) + f_y^{-1}(t - s) \quad (7)$$

and let ψ use $A(\tau) = (0, \tau)$ for each $\tau \in (0, 1)$. Then $U_{X+Y}^{-1} = \psi(\tilde{\Pi}^+(F))$. We will focus on constructing uniform confidence bands for L_{X+Y} and U_{X+Y}^{-1} , since the calculations for the other functions are analogous. \square

Example B (Stochastic dominance). Inference for features of F_{X-Y} at a given x , when $X - Y$ may be considered a treatment effect distribution, has been considered in Fan, Guerre, and Zhu (2017) and the references cited therein. Consider testing for first order stochastic dominance without assuming point identification. Suppose that (X_A, X_B, X_0) have marginal CDFs $(G_A, G_B, G_0) \in (\ell^\infty(\mathbb{R}))^3$, and let F_{Δ_A} and F_{Δ_B} be the distribution functions of the treatment effects $\Delta_A = X_A - X_0$ and $\Delta_B = X_B - X_0$. Without point identification we cannot test the hypothesis $H_0 : F_{\Delta_A} \succeq_{FOSD} F_{\Delta_B}$, which is equivalent to the condition that $F_{\Delta_A}(x) - F_{\Delta_B}(x) \leq 0$ for all x . However, the distributions of Δ_A and Δ_B may be bounded. For example, using references from the previous example it can be shown that for each x ,

$$\begin{aligned} F_{\Delta_A}(x) &\geq L_A(x) = \left[\sup_{u \in \mathbb{R}} (G_A(u) - G_0(u - x)) \right]_+, \\ F_{\Delta_B}(x) &\leq U_B(x) = 1 + \left[\inf_{u \in \mathbb{R}} (G_B(u) - G_0(u - x)) \right]_-. \end{aligned}$$

Under the null hypothesis, $L_A \leq F_{\Delta_A} \leq F_{\Delta_B} \leq U_B$, so that $L_A(x) - U_B(x) \leq F_{\Delta_A} - F_{\Delta_B} \leq 0$ for all x , and the null hypothesis may be rejected whenever there exists an x' such that $L_A(x') > U_B(x')$. This holds regardless of the correlations between treatment outcomes. Let $\Pi^- : (\ell^\infty(\mathbb{R}))^2 \rightarrow \ell^\infty(\mathbb{R}^2)$ be, for $f = (f_0, f_1)$

$$\Pi^-(f)(u, x) = f_1(u) - f_0(u - x)$$

and let ψ be defined by (1) with $A(x) = \mathbb{R}$ for all x . The function used to indicate a violation of dominance maps three distribution functions into a function in $\ell^\infty(\mathbb{R})$: for each $x \in \mathbb{R}$,

$$L_A(x) - U_B(x) = \sup_{u \in \mathbb{R}} (G_A(u) - G_0(u - x)) - 1 - \inf_{u \in \mathbb{R}} (G_B(u) - G_0(u - x)) \quad (8)$$

$$= \psi(\Pi^-(G_0, G_A))(x) - 1 + \psi(-\Pi^-(G_0, G_B))(x). \quad (9)$$

We would like to test the hypothesis that F_A dominates F_B , without knowledge of the dependence between X_0 , X_A and X_B , by looking for x where $L_A(x) - U_B(x) > 0$, a case in which a one-sided functional $f \mapsto \|[f]_+\|_p$ is preferred for testing. \square

Example C (Cost and profit functions). Suppose a firm produces outputs $x \in \mathbb{R}_+^n$ for $n \geq 1$. Given cost function $c : \mathbb{R}^n \rightarrow \mathbb{R}$, the firm's profit at given product prices $p \in \mathbb{R}^n$ is $\pi(p) =$

³We have simplified the problem by excluding $\{0, 1\}$ from the definition of $\tilde{\psi}$ for ease of exposition — see p.115-117 of Williamson and Downs (1990) for more complex cases which include these points.

$\sup_{x \in \mathbb{R}_+^n} (\langle p, x \rangle - c(x))$. Suppose that we have an accurate model of costs and would like to conduct inference for the profit curve, for example by putting a confidence band around an estimated profit curve.

The Legendre-Fenchel transform can be used to find the dual conjugate function of a convex function f or the convex hull of a non-convex function f (Rockafellar and Wets, 1998, Chapter 11). For convex $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the conjugate $f^* : \mathbb{R}^n \rightarrow \mathbb{R}$ is $f^*(x) = \sup_{u \in \mathbb{R}^n} (\langle u, x \rangle - f(u))$, that is, the smallest function such that $f^*(x) + f(u) \geq \langle u, x \rangle$. We can recast this as an optimization map by defining the Legendre transform as a map $\mathcal{L} : \ell^\infty(\mathbb{R}^n) \rightarrow \ell^\infty(\mathbb{R}^n)$ defined by

$$\mathcal{L}(f)(x) = \sup_{u \in \mathbb{R}^n} (\langle u, x \rangle - f(u)).$$

Then the profit function can be seen as the conjugate dual function of the cost function: $\pi(p) = \mathcal{L}(c)(p) = \sup_{x \in \mathbb{R}^n} (\langle p, x \rangle - f(x))$. As discussed in the text, it is possible (i.e., when it is not possible to directly calculate π from the cost function) that π is not Hadamard differentiable. Instead, consider the supremum of the absolute value of transformed functions $\check{\lambda}(f) = \|\mathcal{L}(f)\|_\infty$. Since $\pi = \mathcal{L}(c)$, we can find a confidence band for π using $\check{\lambda}(c)$ using a sample analog $\check{\lambda}(c_n)$. \square

In Example A, the bounds on the quantile function are an example where $A(x)$ varies with x . It would be notationally simpler to consider optimization problems over the rectangle $\mathcal{X} \times \mathcal{U}$. However, given an x -varying set of constraints we would need to use “indicator functions” as they are used in the optimization literature, e.g. we would change the optimization problem to $\tilde{f}(u, x) = f(u, x) + \delta_A(u, x)$, where $\delta_A(u, x) = 0$ for $u \in A(x)$ and (for maximization) $\delta_A(u, x) = -\infty$ for $u \notin A(x)$. This takes us out of the realm of bounded functions and towards the consideration of epi- or hypographs, which may be more complicated. For example, see Bücher, Segers, and Volgushev (2014). The literature on weak convergence in the space of bounded functions is voluminous and more familiar to researchers in economics.

To analyze the asymptotic properties of the above examples with plug-in estimates, one would typically rely on the delta method because ψ is a nonlinear map from objective function to value function. In the next section we discuss the difficulties encountered with the delta method for value functions and a solution for the purposes of uniform inference.

3 Differentiability properties

We use the delta method to establish uniform inference methods for value functions. This depends on the notion of Hadamard (or compact) differentiability to linearize the map near the population distribution (see, for example, van der Vaart and Wellner (1996, Section 3.9.3)), and has the advantage of dividing the analysis into a deterministic part and a statistical part. We first consider

Hadamard derivatives without regard to sample data, before explicitly considering the behavior of sample statistics in the next section.

3.1 Directional differentiability

The delta method between two metric spaces usually relies on (full) Hadamard differentiability (van der Vaart and Wellner, 1996, Section 3.9). Shapiro (1990), Dümbgen (1993) and Fang and Santos (2019) discuss Hadamard directional differentiability and show that this weaker notion also allows for the application of the delta method.⁴

Definition 3.1 (Hadamard directional differentiability). Let \mathbb{D} and \mathbb{E} be Banach spaces and consider a map $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \rightarrow \mathbb{E}$. ϕ is *Hadamard directionally differentiable* at $f \in \mathbb{D}_\phi$ tangentially to a set $\mathbb{D}_0 \subseteq \mathbb{D}$ if there is a continuous map $\phi'_f : \mathbb{D}_0 \rightarrow \mathbb{E}$ such that

$$\lim_{n \rightarrow \infty} \left\| \frac{\phi(f + t_n h_n) - \phi(f)}{t_n} - \phi'_f(h) \right\|_{\mathbb{E}} = 0$$

for all sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbb{R}_+$ such that $h_n \rightarrow h \in \mathbb{D}_0$ and $t_n \rightarrow 0^+$ as $n \rightarrow \infty$ and $f + t_n h_n \in \mathbb{D}_\phi$ for all n .

In case the derivative map ϕ'_f is linear and t_n may approach 0 from both negative and positive sides, the map is fully Hadamard differentiable. The delta method and chain rule can be applied to maps that are either Hadamard differentiable or Hadamard directionally differentiable.

To discuss the differentiability of ψ in equation (1), for any $\epsilon \geq 0$ let $U_f : X \rightrightarrows U$ define the set-valued map of ϵ -maximizers of $f(\cdot, x)$ in u for each $x \in X$:

$$U_f(x, \epsilon) = \{u \in A(x) : f(u, x) \geq \psi(f)(x) - \epsilon\}. \quad (10)$$

Results on the directional derivatives of the optimization map $f \mapsto \sup_X f$ date to Danskin (1967). For example, Cárcamo, Cuevas, and Rodríguez (2020, Theorem 2.1) show that, in our notation, for fixed $x \in X$, ψ is directionally differentiable in $\ell^\infty(A(x) \times \{x\})$, and for directions $h(\cdot, x)$

$$\psi'_f(h)(x) = \lim_{\epsilon \rightarrow 0^+} \sup_{u \in U_f(x, \epsilon)} h(u, x). \quad (11)$$

Assuming the stronger conditions that A is continuous and compact-valued and that f is continuous on the graph of A ($\text{gr}A$), the maximum theorem implies that U_f is non-empty, compact-valued and upper hemicontinuous (Aliprantis and Border, 2006, Theorem 17.31), and Cárcamo, Cuevas, and Rodríguez (2020, Corollary 2.2) show that tangentially to $\mathcal{C}(U \times \{x\})$, the derivative of $\psi(f)(x)$

⁴The derivative ϕ'_f in the definition is also called a *semiderivative* (Rockafellar and Wets, 1998, Definition 7.20).

simplifies to $\psi'_f(h)(x) = \sup_{u \in U_f(x,0)} h(u, x)$. Further, when (for fixed x) $U_f(x, 0)$ is a singleton set, $\psi(f)(x)$ is fully differentiable, not just directionally so.

The pointwise directional differentiability of $\psi(f)(x)$ at each x might lead one to suspect that ψ is differentiable more generally as a map from $\ell^\infty(\text{gr}A)$ to $\ell^\infty(X)$. However, this is not true. The following example illustrates a case where ψ is not Hadamard directionally differentiable as a map from $\ell^\infty(\text{gr}A)$ to $\ell^\infty(X)$, although f and h are both continuous functions.

Example D. Let $U \times X = [0, 1] \times [-1, 1]$ and define $f : U \times X \rightarrow \mathbb{R}$ by

$$f(u, x) = \begin{cases} 0 & x \in [-1, 0] \\ x(u^2 - u) & x \in (0, 1] \end{cases}.$$

Let $h(u, x) = -u^2 + u$. Then $\psi(f)(x) = 0$ for all x and for any $t > 0$,

$$\psi(f + th)(x) = \begin{cases} t/4 & x \in [-1, 0] \\ (t - x)/4 & x \in (0, t] \\ 0 & x \in (t, 1] \end{cases}.$$

Therefore for each x ,

$$\lim_{t \rightarrow 0^+} \frac{\psi(f + th)(x) - \psi(f)(x)}{t} = \frac{1}{4}I(x \leq 0).$$

However, for any $t > 0$,

$$\sup_{x \in [-1, 1]} |(\psi(f + th)(x) - \psi(f)(x))/t - (1/4)I(x \leq 0)| = 1/4.$$

This implies that no derivative exists as an element of $\ell^\infty([-1, 1])$. The functions f and h are well behaved, but because the candidate derivative $I(x \leq 0)/4$ is not continuous in x , uniform convergence of the quotients to the candidate fails. \square

3.2 Directional differentiability of norms of value functions

The lack of uniformity of the convergence to $\psi'_f(h)(\cdot)$ appears to preclude the delta method for uniform inference, because the typical path of analysis would use a well-behaved limit of $r_n(\psi(f_n) - \psi(f))$ and the continuous mapping theorem to find the distribution of statistics applied to the limit. However, real-valued statistics are used to find uniform inference results, so we examine maps that include not only the marginal optimization step but also a functional applied to the resulting value function.

Consider L_p norms (for $1 \leq p \leq \infty$) applied to $\psi(f)$ or $[\psi(f)]_+$. Letting m denote the domi-

nating measure, define $\lambda_j : \ell^\infty(\text{gr}A) \rightarrow \mathbb{R}$ for $j = 1, \dots, 4$ by

$$\begin{aligned} \lambda_1(f) &= \sup_{x \in X} \left| \sup_{u \in A(x)} f(u, x) \right|, & \lambda_2(f) &= \sup_{x \in X} \left[\sup_{u \in A(x)} f(u, x) \right]_+, \\ \lambda_3(f) &= \left(\int_X \left| \sup_{u \in A(x)} f(u, x) \right|^p dm(x) \right)^{1/p}, & \lambda_4(f) &= \left(\int_X \left[\sup_{u \in A(x)} f(u, x) \right]_+^p dm(x) \right)^{1/p}. \end{aligned} \quad (12)$$

The maps λ_2 and λ_4 are included because one-sided comparisons may also be of interest, and the map $f \mapsto [f]_+$ is pointwise directionally differentiable but not differentiable as a map from $\ell^\infty(X)$ to $\ell^\infty(X)$. We illustrate its use in a stochastic dominance example below.

We make the following definitions in order to construct well-defined derivatives. First, let $\mu(f)$ be the supremum of f over $\text{gr}A$ and let $\sigma(f)$ be the maximin value of f over $\text{gr}A$:

$$\mu(f) = \sup_{(u,x) \in \text{gr}A} f(u, x) \quad (13)$$

$$\sigma(f) = \sup_{x \in X} \inf_{u \in A(x)} f(u, x). \quad (14)$$

Define $A_f^\mu \subseteq \text{gr}A$ for any $\epsilon > 0$ to be the set of ϵ -maximizers of f in (u, x) over $\text{gr}A$. Next, let $X_f^\sigma \subseteq X$ collect the ϵ -maximinimizers of f in x over X . That is, define

$$A_f^\mu(\epsilon) = \{(u, x) \in \text{gr}A : f(u, x) \geq \mu(f) - \epsilon\} \quad (15)$$

$$X_f^\sigma(\epsilon) = \left\{ x \in X : \inf_{u \in A(x)} f(u, x) \geq \sigma(f) - \epsilon \right\}. \quad (16)$$

Also, note that $U_{(-f)}(x, \epsilon)$ is the set of ϵ -minimizers of $f(u, x)$ in u for given x .

Cárcamo, Cuevas, and Rodríguez (2020) show that the map (13) has directional derivative

$$\mu'_f(h) = \lim_{\epsilon \rightarrow 0^+} \sup_{x \in A_f^\mu(\epsilon)} h(u, x) \quad (17)$$

and Lemma B.1 in the Appendix shows that (14) has directional derivative

$$\sigma'_f(h) = \lim_{\delta \rightarrow 0^+} \sup_{x \in X_f^\sigma(\delta)} \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x). \quad (18)$$

To analyze derivatives of the one-sided $\lambda_2(f)$ and $\lambda_4(f)$, define the contact set

$$X_0 = \{x \in X : \psi(f)(x) = 0\}. \quad (19)$$

The following two theorems describe the form of the four directional derivatives generally and

under the condition that the value function is zero everywhere or on some subset of X .

Theorem 3.2. *Suppose $f, h \in \ell^\infty(\text{gr}A)$, where $A : X \rightrightarrows U$ is non-empty-valued for all $x \in X$. Define μ by (13) and σ by (14), and their derivatives μ'_f and σ'_f as in (17) and (18). Then:*

1. $\lambda_1(f)$ is Hadamard directionally differentiable and

$$\lambda'_{1f}(h) = \begin{cases} \mu'_f(h), & \mu(f) > \sigma(-f) \\ \max \left\{ \mu'_f(h), \sigma'_{(-f)}(-h) \right\}, & \mu(f) = \sigma(-f) \\ \sigma'_{(-f)}(-h), & \mu(f) < \sigma(-f) \end{cases}.$$

2. If $\|\psi(f)\|_\infty = 0$, $\lambda_1(f)$ is Hadamard directionally differentiable and

$$\begin{aligned} \lambda'_{1f}(h) &= \max \left\{ \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X,\epsilon)} h(u,x), \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X} \inf_{u \in U_f(x,\epsilon)} (-h(u,x)) \right\} \\ &= \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X} \left| \sup_{u \in U_f(x,\epsilon)} h(u,x) \right|. \end{aligned}$$

3. $\lambda_2(f)$ is Hadamard directionally differentiable and

$$\lambda'_{2f}(h) = \begin{cases} \mu'_f(h), & \mu(f) > 0 \\ \left[\mu'_f(h) \right]_+, & \mu(f) = 0 \\ 0, & \mu(f) < 0 \end{cases}.$$

4. If $\|[\psi(f)]_+\|_\infty = 0$, $\lambda_2(f)$ is Hadamard directionally differentiable and

$$\lambda'_{2f}(h) = \begin{cases} \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X_0,\epsilon)} [h(u,x)]_+, & \mu(f) = 0 \\ 0, & \mu(f) < 0 \end{cases}.$$

The first part of Theorem 3.2 essentially takes advantage of the identity $|x| = \max\{x, -x\}$, and breaks the evaluation of the supremum of the absolute value of the value function into two parts that are well-defined, a maximization and a minimization problem. The second part of the theorem is simpler because the $x \mapsto [x]_+$ map simplifies the evaluation of the supremum.

The next result shows that the L_p functionals for $1 \leq p < \infty$ defined in (12), are Hadamard directionally differentiable. The form of the derivatives for these maps also change depending on whether $\lambda_j(f) = 0$ or $\lambda_j(f) \neq 0$.

Theorem 3.3. *Suppose that $f, h \in \ell^\infty(\text{gr}A)$ where $A : X \rightrightarrows U$ is non-empty-valued for all $x \in X$ and $m(X) < \infty$. Recall ψ'_f defined in (11). Then:*

1. If $\|\psi(f)\|_p \neq 0$, then λ_3 is Hadamard directionally differentiable and

$$\lambda'_{3f}(h) = \|\psi(f)\|_p^{1-p} \int_X \text{sgn}(\psi(f)(x)) |\psi(f)(x)|^{p-1} \psi'_f(h)(x) dm(x).$$

2. If $\|\psi(f)\|_p = 0$, then λ_3 is Hadamard directionally differentiable and

$$\lambda'_{3f}(h) = \left(\int_X |\psi'_f(h)(x)|^p dm(x) \right)^{1/p}.$$

3. If $\|[\psi(f)]_+\|_p \neq 0$, then λ_4 is Hadamard directionally differentiable and

$$\lambda'_{4f}(h) = \|[\psi(f)]_+\|_p^{1-p} \int_X [\psi(f)(x)]_+^{p-1} \psi'_f(h)(x) dm(x).$$

4. If $\|[\psi(f)]_+\|_p = 0$, then λ_4 is Hadamard directionally differentiable and

$$\lambda'_{4f}(h) = \left(\int_{X_0} [\psi'_f(h)(x)]_+^p dm(x) \right)^{1/p}.$$

Theorem 3.3 is related to the results of Chen and Fang (2019), who dealt with a squared L_2 statistic. It is interesting to note here that using an L_2 statistic instead of its square results in first-order (directional) differentiability of the map, unlike the squared L_2 norm, which has a degenerate first-order derivative but nondegenerate second-order derivative.

We conclude this section by revisiting Examples A, B, and C and describing the computation of the directional derivatives in the corresponding cases.

Example A (Dependency bounds continued). To construct a uniform confidence band for L_{X+Y} we can invert a test for the hypothesis that $L_{X+Y} = L_0$. Recalling the map Π^+ defined in (6), consider the statistic $\lambda : (\ell^\infty(\mathbb{R}))^2 \rightarrow \mathbb{R}$ defined by

$$\lambda(F) = \sup_{x \in \mathbb{R}} |\max\{\sup_{u \in \mathbb{R}} \Pi^+(F)(u, x), 0\} - L_0(x)|. \quad (20)$$

This is challenging because of the pointwise maximum inside the absolute value. Lemma B.3 in the appendix, which uses calculations similar to those in the proof of Theorem 3.2, shows that, letting

$$\begin{aligned} X_0 &= \{x \in \mathbb{R} : \sup_{u \in \mathbb{R}} (F_X(u) + F_Y(x - u)) = 1\} \\ X_+ &= \{x \in \mathbb{R} : \sup_{u \in \mathbb{R}} (F_X(u) + F_Y(x - u)) > 1\} \\ U_F(x, \epsilon) &= \{u \in \mathbb{R} : F_X(u) + F_Y(x - u) \geq \sup_{u \in \mathbb{R}} (F_X(u) + F_Y(x - u)) - \epsilon\}, \end{aligned}$$

λ is Hadamard directionally differentiable and its derivative is, for direction $h = (h_X, h_Y)$,

$$\lambda'_F(h) = \max \left\{ \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X_0} \sup_{u \in U_F(x, \epsilon)} [h_X(u) + h_Y(x - u)]_+, \right. \\ \left. \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X_+} \left| \sup_{u \in U_F(x, \epsilon)} (h_X(u) - h_Y(x - u)) \right| \right\}. \quad (21)$$

Similarly, we may work with the bound U_{X+Y}^{-1} on the quantile function F_{X+Y}^{-1} . When a CDF F is continuously differentiable on a set $\mathcal{T} = [F^{-1}(p) - \epsilon, F^{-1}(q) + \epsilon]$ for $0 < p < q < 1$ and $\epsilon > 0$, it is fully Hadamard differentiable in $\ell^\infty(\mathcal{T})$ with derivative in direction h at any $\tau \in \mathcal{T}$ given by $-h(F^{-1}(\tau))/f(F^{-1}(\tau))$, where f is the density associated with F (van der Vaart and Wellner, 1996, Lemma 3.9.23). Recall the definition of $\tilde{\Pi}^+$ in (7) and consider, for $F = (F_X, F_Y)$,

$$\tilde{\lambda}(F) = \sup_{\tau \in \mathcal{T}} \left| \sup_{u \in (0, \tau)} \tilde{\Pi}^+(F)(u, x) - U_0^{-1}(x) \right|. \quad (22)$$

Under differentiability of the CDFs, we have

$$(\tilde{\Pi}^+)'_F(h)(s, t) = -\frac{h_X(F_X^{-1}(s))}{f_X(F_X^{-1}(s))} - \frac{h_Y(F_Y^{-1}(t - s))}{f_Y(F_Y^{-1}(t - s))}.$$

Then this fact, Theorem 3.2 and the chain rule imply that, letting $\tilde{U}(\tau, \epsilon) = \{u \in (0, \tau) : F_X^{-1}(u) + F_Y^{-1}(\tau - u) \geq \sup_{u \in (0, \tau)} (F_X^{-1}(u) - F_Y^{-1}(\tau - u)) - \epsilon\}$,

$$\tilde{\lambda}'_F(h) = \lim_{\epsilon \rightarrow 0} \sup_{\tau \in \mathcal{T}} \left| \sup_{u \in \tilde{U}(\tau, \epsilon)} \left(-\frac{h_X(F_X^{-1}(u))}{f_X(F_X^{-1}(u))} - \frac{h_Y(F_Y^{-1}(\tau - u))}{f_Y(F_Y^{-1}(\tau - u))} \right) \right|.$$

□

Example B (Stochastic dominance continued). Similarly to the previous example, we may test the necessary condition for stochastic dominance using the test statistic

$$\Lambda = \left(\int [(L_A(x) - U_B(x))]_+^2 dm(x) \right)^{1/2}. \quad (23)$$

This test statistic is zero when $L_A \leq U_B$ and greater than zero otherwise. The map Π^- is fully differentiable with derivative $(\Pi)'_f(h) = h_1(u) - h_0(u - x)$. Then Λ is directionally differentiable

and, extending Theorem 3.3 to two functions inside the integral,

$$\Lambda'_G(h) = \left(\int_{L_A=U_B} \left[\lim_{\epsilon \rightarrow 0^+} \sup_{u \in U_{\Pi^-(G_0, G_A)}(\epsilon, x)} \Pi^-(h_0, h_A)(u, x) - \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{-\Pi^-(G_0, G_B)}(\epsilon, x)} \Pi^-(h_0, h_B)(u, x) \right]_+^2 dm(x) \right)^{1/2}. \quad (24)$$

□

Example C (Cost and profit functions continued). This example shows how some of the technical difficulties encountered in the previous two examples may be alleviated with assumptions on the economic model. Recall that a uniform confidence band for $\pi = \mathcal{L}(c)$ depends on the derivative of $\check{\lambda}(f) = \sup_x |\mathcal{L}(f)(x)|$. Suppose that $(u, x) \in \mathcal{U} \times \mathcal{X}$, a compact set. More generally, the assumptions that firms will not operate with negative profits and that the cost function is 1-coercive (that is, that $\lim_{\|x\| \rightarrow \infty} c(x)/\|x\| = +\infty$) would ensure that π is continuous. Given the existence of optimizers and a maximum, write

$$\begin{aligned} \check{\lambda}(c) &= \max_p \left| \max_x (p \cdot x - c(x)) \right| \\ &= \max \left\{ \max_{p, x} (p \cdot x - c(x)), \max_p \min_x (c(x) - p \cdot x) \right\}. \end{aligned}$$

Since c is estimated and c only depends on x , it is sufficient to perturb the above statistic in a direction that does not depend on p , i.e., $h(x)$, to calculate derivatives.

Define the derivative of $\check{\mu}(c) = \max_{p, x} (p \cdot x - c(x))$ by

$$\check{\mu}'_c(h) = \max_{\operatorname{argmax}_{(p, x)} (p \cdot x - c(x))} (-h(x)),$$

and the derivative of $\check{\sigma}(c) = \max_p \min_x (c(x) - p \cdot x)$ by

$$\check{\sigma}'_c(h) = \max_{\operatorname{argmax}_p (\operatorname{argmin}_x (c(x) - p \cdot x))} \min_{\operatorname{argmin}_x (c(x) - p \cdot x)} h(x).$$

With strict convexity of c the argmin expressions in $\check{\mu}'$ and $\check{\sigma}'$ would be single-valued, and these would be simpler. The directional derivative of $\lambda(c)$ is

$$\lambda'_c(h) = \begin{cases} \check{\mu}'_c(h), & \check{\mu}(c) > \check{\sigma}(c), \\ \check{\mu}'_c(h) \vee \check{\sigma}'_c(h), & \check{\mu}(c) = \check{\sigma}(c), \\ \check{\sigma}'_c(h) & \check{\mu}(c) < \check{\sigma}(c) \end{cases}. \quad (25)$$

□

4 Inference

In this section we derive asymptotic distributions for uniform statistics applied to value functions, and propose bootstrap estimators of their distributions that can be used for practical inference.

4.1 Asymptotic distributions

The derivatives developed in the previous subsections can be used in a functional delta method. We make a high-level assumption on the way that observations are related to estimated functions, but conditions under which such convergence holds are well-known, and examples will be shown below.

A1. Assume that for each n there is a random sample $\{Z_i\}_{i=1}^n$ with P denoting their joint probability distribution. Assume there is a map $\{Z_i\}_{i=1}^n \mapsto f_n$ where $f_n \in \ell^\infty(U \times X)$. In case λ_3 or λ_4 is used, also assume $m(X) < \infty$. Furthermore, for some sequence $r_n \rightarrow \infty$, $r_n(f_n - f) \rightsquigarrow \mathcal{G}_P$, where \mathcal{G}_P is a tight random element of $\ell^\infty(\text{gr}A)$.

In case an L_p statistic is used (with $p < \infty$), we restrict the space of allowable functions. In order to ensure that L_p statistics are well-defined we make the assumption that the measure of X is finite. While stronger than the assumption used for the supremum statistic, it is sufficient to ensure that the L_p statistics of the value function are finite. This assumption also has the advantage of providing an easy-to-verify condition based on the objective function f , rather than the more direct, but less obvious condition $\psi(f) \in L_p(m)$. Lifting this restriction would require some other restriction on the objective function f that is sufficient to ensure the value function is p -integrable.⁵

Given assumption **A1** and the derivatives of the last section, the asymptotic distribution of test statistics applied to the value function is straightforward.

Theorem 4.1. *Under Assumption **A1**,*

$$r_n(\lambda_j(f_n) - \lambda_j(f)) \rightsquigarrow \lambda'_{jf}(\mathcal{G}_P),$$

where for $j \in \{1, \dots, 4\}$, λ_j are defined in (12) and λ'_{jf} are defined in Theorems 3.2 and 3.3.

This theorem is abstract and hides the fact that the limiting distributions may depend on features of λ_j and f . Therefore it is only indirectly useful for inference. A resampling scheme is the subject of the next section.

⁵We attempted to show p -integrability by assuming that f is bounded and integrable (note that this would imply f is p -integrable in $\text{gr}A$), but were unable to show that this implies that the value function is integrable (in X). Note that if one were able to show that f is bounded and integrable, then the elegant results in Kaji (2019, Section 3) would apply for the purposes of verifying Assumption **A1**.

The delta method as used in Theorem 4.1 is tailored to arguments that have weak limits in the space $\ell^\infty(\text{gr}A)$, which is an intrinsic feature of the delta method. That implies some limitations on the type of functions that may be considered. For example, functions estimated using kernel methods do not converge to a weak limit as assumed in Assumption **A1**. One can in general only show the stochastic order of the sequence $r_n(f_n - f)$, and methods for uniform inference must shown using something other than the delta method. We leave uniform inference for value functions based on less regular estimators to future research.

Our bootstrap technique ahead is designed for resampling under the null hypothesis that $\lambda_j(f) = 0$. We assume that when $j = 1$ or $j = 3$ (that is, conventional two-sided statistics are used), the statistics are used to test the null and alternative hypotheses

$$H_0 : \psi(f)(x) = 0 \quad \text{for all } x \in X \quad (26)$$

$$H_1 : \psi(f)(x) \neq 0 \quad \text{for some } x \in X. \quad (27)$$

Meanwhile, when $j = 2$ or $j = 4$ (the one-sided cases) we assume the hypotheses are

$$H_0 : \psi(f)(x) \leq 0 \quad \text{for all } x \in X \quad (28)$$

$$H_1 : \psi(f)(x) > 0 \quad \text{for some } x \in X. \quad (29)$$

The probability measure P is assumed to belong to a collection \mathcal{P} which describes the set of probability measures allowed by the model. A few subcollections of \mathcal{P} serve to organize the asymptotic results below. When P is such that $\psi(f) \equiv 0$, we label $P \in \mathcal{P}_{00}^E$. For functional inequalities, the behavior of test statistics under the null is more complicated. If P is such that $\psi(f) \leq 0$ everywhere, then label $P \in \mathcal{P}_0^I$. When $P \in \mathcal{P}_0^I$ makes $\psi(f)(x) = 0$ for at least one $x \in X$, then we label $P \in \mathcal{P}_{00}^I$.

The following corollary combines the derivatives from Theorem 3.2 and Theorem 3.3 with the result of Theorem 4.1 for these distribution classes. Recall that the set X_0 , defined in (19), represents the subset of X where $\psi(f)$ is zero.

Corollary 4.2. *Under Assumption **A1**, for $j \in \{1, \dots, 4\}$,*

1. For each $P \in \mathcal{P}_{00}^E$, $r_n(\lambda_1(f_n) - \lambda_1(f)) \rightsquigarrow \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X} \left| \sup_{u \in U_f(x, \epsilon)} \mathcal{G}_P(u, x) \right|$.
2. (a) For each $P \in \mathcal{P}_{00}^I$, $r_n(\lambda_2(f_n) - \lambda_2(f)) \rightsquigarrow \lim_{\epsilon \rightarrow 0^+} \sup_{(u, x) \in U_f(X_0, \epsilon)} [\mathcal{G}_P(u, x)]_+ \cdot$
(b) For each $P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$, $r_n(\lambda_2(f_n) - \lambda_2(f)) \xrightarrow{P} 0$.
3. For each $P \in \mathcal{P}_{00}^E$, $r_n(\lambda_3(f_n) - \lambda_3(f)) \rightsquigarrow \left(\int_X \left| \psi'_f(\mathcal{G}_P)(x) \right|^p dm(x) \right)^{1/p}$.
4. (a) For each $P \in \mathcal{P}_{00}^I$, $r_n(\lambda_4(f_n) - \lambda_4(f)) \rightsquigarrow \left(\int_{X_0} \left[\psi'_f(\mathcal{G}_P)(x) \right]_+^p dm(x) \right)^{1/p}$.

(b) For each $P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$, $r_n(\lambda_4(f_n) - \lambda_4(f)) \xrightarrow{P} 0$.

The distributions described in the above corollary under the assumption that $P \in \mathcal{P}_{00}^E$ or \mathcal{P}_{00}^I are those that we emulate using resampling methods in the next section.

4.2 Resampling

In the previous section we established asymptotic distributions for the test statistics of interest. However, for practical inference we turn to resampling to estimate their distributions. This section suggests the use of a resampling strategy that was proposed by Fang and Santos (2019), and it combines a standard bootstrap procedure with estimates of the directional derivatives λ'_{jf} . Hadamard directional differentiability also implies that the numerical approximation methods of Hong and Li (2018) may be used to estimate these directional derivatives, an approach that we address in simulations (in the supplementary appendix).

The resampling scheme described below is designed to reproduce the null distribution under the assumption that the statistic is equal to zero, or in other terms, that $P \in \mathcal{P}_{00}^E$ or $P \in \mathcal{P}_{00}^I$. This is achieved by restricting the form of the estimates $\hat{\lambda}'_{jn}$. The bootstrap routine is consistent under more general conditions, as in Theorem 3.1 of Fang and Santos (2019). However, we discuss behavior of bootstrap-based tests under the assumption that one of the null conditions described in the previous section holds.

All the derivative formulas in Corollary 4.2 require some form of an estimate of the near-maximizers of f in u , that is, of the set $U_f(x, \epsilon)$ defined in (10) for various x . The set estimators we use are similar to those used in Linton, Song, and Whang (2010), Chernozhukov, Lee, and Rosen (2013) and Lee, Song, and Whang (2018) and depend on slowly decreasing sequences of constants to estimate the relevant sets in the derivative formulas. For a sequence $a_n \rightarrow 0^+$ we estimate $U_f(x, \epsilon)$ with the plug-in estimator $U_{f_n}(x, a_n)$. The sequence a_n should decrease more slowly than the rate at which f_n converges uniformly to f , an assumption which will be formalized below.

The one-sided estimates $\hat{\lambda}'_{2n}$ and $\hat{\lambda}'_{4n}$ use a second estimate. For another sequence $b_n \rightarrow 0^+$, estimate the contact set X_0 defined in (19) with

$$\hat{X}_0 = \begin{cases} \{x \in X : |\psi(f_n)(x)| \leq b_n\} & \text{if } \{x \in X : |\psi(f_n)(x)| \leq b_n\} \neq \emptyset \\ X & \text{if } \{x \in X : |\psi(f_n)(x)| \leq b_n\} = \emptyset \end{cases}.$$

This is a method used by Linton, Song, and Whang (2010) of achieving uniformity in the convergence of estimators that depend on contact set estimation under the null hypothesis $P \in \mathcal{P}_{00}^I$.

Define estimators of λ'_{1f} and λ'_{2f} by

$$\begin{aligned}\hat{\lambda}'_{1n}(h) &= \max \left\{ \sup_{(u,x) \in U_{f_n}(X, a_n)} h(u, x), \sup_{x \in X} \inf_{u \in U_{f_n}(x, a_n)} (-h(u, x)) \right\} \\ &= \sup_{x \in X} \left| \sup_{u \in U_{f_n}(x, a_n)} h(u, x) \right|\end{aligned}\tag{30}$$

and

$$\hat{\lambda}'_{2n}(h) = \sup_{(u,x) \in U_{f_n}(\hat{X}_0, a_n)} [h(u, x)]_+.\tag{31}$$

These formulas impose the condition that $P \in \mathcal{P}_{00}^E$ or \mathcal{P}_{00}^I on the derivative estimates, following the forms shown in Corollary 4.2. For more general estimation we would need to devise more complex estimators to deal with the variety of limits discussed in Theorems 3.2 and 3.3, and also the technique of Hong and Li (2018) could be used in that case. However, inference is usually of primary interest, and generating a bootstrap distribution that respects the null hypothesis can improve the performance of bootstrap inference procedures. When $P \in \mathcal{P}_{00}^E$ the imposition amounts to the assumption that $\psi(f)(x) = 0$ for each x , so that the set of population ϵ -maximizers in $\text{gr}A$ is $U_f(X, \epsilon)$, and the estimator $\hat{\lambda}'_{1n}$ emulates that. The condition $P \in \mathcal{P}_{00}^I$ implies that X_0 is not empty, and \hat{X}_0 is meant to estimate this contact set, while using all of X when the distribution is not as is hypothesized.

The estimators $\hat{\lambda}'_{3n}$ and $\hat{\lambda}'_{4n}$ are defined similarly to the supremum-norm estimators: let

$$\hat{\lambda}'_{3n}(h) = \left(\int_X \left| \sup_{u \in U_{f_n}(x, a_n)} h(u, x) \right|^p dm(x) \right)^{1/p}\tag{32}$$

and

$$\hat{\lambda}'_{4n}(h) = \left(\int_{\hat{X}_0} \sup_{u \in U_{f_n}(x, a_n)} [h(u, x)]_+^p dm(x) \right)^{1/p}.\tag{33}$$

We use an exchangeable bootstrap, which depends on a set of weights $\{W_i\}_{i=1}^n$ that put probability mass W_i at each observation Z_i . This type of bootstrap describes many well-known bootstrap techniques including sampling with replacement from the observations (van der Vaart and Wellner, 1996, Section 3.6.2).

We make the following assumptions to ensure the bootstrap is well-behaved.

A2. Assume $a_n, b_n \rightarrow 0^+$, $a_n r_n \rightarrow \infty$ and $b_n r_n \rightarrow \infty$.

A3. Suppose that for each n , W is independent of the data Z and there is a map $\{Z_i, W_i\}_{i=1}^n \mapsto f_n^*$ where $f_n^* \in \ell^\infty(\text{gr}A)$. Suppose $r_n(f_n^* - f_n)$ is asymptotically measurable; for all continuous

and bounded g , $g(r_n(f_n^* - f_n))$ is a measurable function of $\{W_i\}$ outer almost surely in $\{Z_i\}$; and $r_n(f_n^* - f_n) \rightsquigarrow \mathcal{G}_P$, where \mathcal{G}_P was defined in Assumption **A1**.

We remark that it is difficult to propose a general method for choosing how to set a_n and b_n in practice. We numerically investigated their effect in simulations. We apply similar rules to those used in Linton, Song, and Whang (2010) and Lee, Song, and Whang (2018). However, in general situations, where the variance of the objective function changes significantly over its domain, one would need to estimate the pointwise standard deviation of the function. For example, Fan and Park (2010) provide pointwise asymptotic distributions for estimated dependency bound functions for F_{X-Y} . Nevertheless, we leave this extension for future research.

Assumption **A2** is used to ensure consistency of the ϵ -maximizer set estimators used in the bootstrap algorithm. Assumption **A3** is a condensed version of Assumption 3 of Fang and Santos (2019) to ensure that $r_n(f_n^* - f_n)$ behaves asymptotically like $r_n(f_n - f)$.

Resampling routine to estimate the distribution of $r_n(\lambda_j(f_n) - \lambda_j(f))$

1. Estimate $\hat{\lambda}'_{jn}$ using sample data and formulas (30)-(33) above.

Then repeat steps 2-3 for $r = 1, \dots, R$:

2. Use an exchangeable bootstrap to construct $r_n(f_n^* - f_n)$.
3. Calculate the resampled test statistic $\lambda_r^* = \hat{\lambda}'_{jn}(r_n(f_n^* - f_n))$ using the estimate from step 1.

Finally,

4. Let $\hat{q}_{\lambda^*}(1 - \alpha)$ be the $(1 - \alpha)$ -th sample quantile from the bootstrap distribution of $\{\lambda_r^*\}_{r=1}^R$, where $\alpha \in (0, 1)$ is the nominal size of the test. Reject the null hypothesis if $r_n \lambda_j(f_n)$ is larger than $\hat{q}_{\lambda^*}(1 - \alpha)$.

The consistency of this resampling procedure under the null hypothesis is summarized in the following theorem. To discuss weak convergence it is easiest to use the space of bounded Lipschitz functions, which indicates weak convergence. That is, defining $BL_1(\mathbb{R}) = \{g \in C_b(\mathbb{R}) : \sup_x |g(x)| \leq 1, \sup_{x \neq y} |g(x) - g(y)| \leq |x - y|\}$, X_n converges weakly to X if and only if $\sup_{g \in BL_1(\mathbb{R})} |\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]| \rightarrow 0$ (van der Vaart and Wellner, 1996, §1.12).

Theorem 4.3. *Under Assumptions **A1-A3**, if either $j \in \{1, 3\}$ and $P \in \mathcal{P}_{00}^E$, or $j \in \{2, 4\}$ and $P \in \mathcal{P}_{00}^I$, then for any $\epsilon > 0$,*

$$\limsup_{n \rightarrow \infty} P \left\{ \sup_{g \in BL_1(\mathbb{R})} \left| \mathbb{E} \left[g \left(\hat{\lambda}'_{jn}(r_n(f_n^* - f_n)) \right) \mid \{Z_i\}_{i=1}^n \right] - \mathbb{E} \left[g \left(\lambda'_{jf}(\mathcal{G}_P) \right) \right] \right| > \epsilon \right\} = 0.$$

We conclude this section by revisiting Examples A, B, and C and describing the resampling procedures in the corresponding cases.

Example A (Dependency bounds continued). Recall that tests using the statistic $\lambda(F)$ defined in (20) may be inverted to find a uniform confidence band for L_{X+Y} (here we focus on the lower dependency bound for the CDF F_{X+Y} only). Suppose that we observe two independent random samples, $\{X_i\}_{i=1}^{n_X}$ and $\{Y_i\}_{i=1}^{n_Y}$ and $\mathbb{F}_n = (\mathbb{F}_{Xn}, \mathbb{F}_{Yn})$ be their empirical CDFs. Then

$$\lambda(\mathbb{F}_n) = \sup_{x \in \mathbb{R}} |\mathbb{L}_n(x) - L_0(x)|, \quad (34)$$

where $\mathbb{L}_n(x) = \mathbb{L}_n(x, \mathbb{F}_n) = \max \{ \sup_{u \in \mathbb{R}} (\mathbb{F}_{Xn}(u) + \mathbb{F}_{Yn}(x - u)) - 1, 0 \}$.

Standard conditions ensure that, with $n = n_X + n_Y$, $\sqrt{n}(\mathbb{F}_n - F) \rightsquigarrow \mathcal{G}_F$, where \mathcal{G}_F is a Gaussian process in $(\ell^\infty(\mathbb{R}))^2$. Under the hypothesis that L_0 is the true lower bound function, $\sqrt{n}\lambda(\mathbb{F}_n) \rightsquigarrow \lambda'_F(\mathcal{G}_F)$, where λ'_F is defined in (21). This derivative needs to be estimated as in (30), estimating the set-valued map $U_F(x, \epsilon)$ with the plug-in estimate

$$U_{\mathbb{F}_n}(x, a_n) = \{u \in \mathbb{R} : \max\{\mathbb{F}_{Xn}(u) + \mathbb{F}_{Yn}(x - u), 0\} \geq \mathbb{L}_n(x) - a_n\}, \quad (35)$$

where a_n is a sequence that converges slowly to zero, and with an additional slowly-decreasing sequence b_n that is used to determine the regions where $\Pi^+(\mathbb{F}_n)$ falls in the interval $(-b_n, b_n)$ or is above b_n to estimate \hat{X}_0 and \hat{X}_+ . Labeling the estimated derivative $\hat{\lambda}'_F$, slight modifications of Corollary 4.2 and Theorem 4.3 imply that for all $q \geq 0$,

$$P \left\{ \hat{\lambda}'_F(\sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n)) \leq q \mid \{X_i\}_{i=1}^{n_X}, \{Y_i\}_{i=1}^{n_Y} \right\} \xrightarrow{P} P \left\{ \lambda'_F(\mathcal{G}_F) \leq q \right\}.$$

We can find a critical value of the asymptotic distribution using the bootstrap by estimating

$$\hat{q}_{\lambda^*}(1 - \alpha) = \min \left\{ q : P \left\{ \hat{\lambda}'_F(\sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n)) \leq c \mid \{X_i\}_{i=1}^{n_X}, \{Y_i\}_{i=1}^{n_Y} \right\} \geq 1 - \alpha \right\}$$

by simulation. For $r = 1, \dots, R$, let $\lambda_r^* = \hat{\lambda}'_F(\sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n))$, and let $\hat{q}_{\lambda^*}(1 - \alpha)$ be the $(1 - \alpha)$ -th quantile of the bootstrap sample $\{\lambda_r^*\}_{r=1}^R$. Simulation evidence presented in the supplementary appendix verifies that the coverage probability of these intervals is accurate in moderate sample sizes by examining the empirical rejection probabilities of the associated tests. \square

Example B (Stochastic dominance continued). Suppose that three independent random samples $\{X_{0i}\}_{i=1}^n$, $\{X_{Ai}\}_{i=1}^n$ and $\{X_{Bi}\}_{i=1}^n$ are observed. Let $\mathbb{G}_n = (\mathbb{G}_{0n}, \mathbb{G}_{An}, \mathbb{G}_{Bn})$ be their empirical distribution functions, and define the sample statistic

$$\hat{\Lambda}(\mathbb{G}_n) = \left(\int_{\mathbb{R}} [\mathbb{L}_{An}(x) - \mathbb{U}_{Bn}(x)]_+^2 dm(x) \right)^{1/2}, \quad (36)$$

where \mathbb{L}_{An} is the plug-in estimate of L_A using \mathbb{G}_{0n} and \mathbb{G}_{An} and \mathbb{U}_{Bn} is the plug-in estimate of U_B

using \mathbb{G}_{0n} and \mathbb{G}_{Bn} . Under standard conditions, with $n = n_0 + n_A + n_B$, $\sqrt{n}(\mathbb{G}_n - G) \rightsquigarrow \mathcal{G}_G$, where \mathcal{G}_G is a Gaussian process in $(\ell^\infty(\mathbb{R}))^3$. The theory above can be extended in a straightforward way to show that $\hat{\Lambda}(\sqrt{n}\mathbb{G}_n) \rightsquigarrow \Lambda'_G(\mathcal{G}_G)$, where $\Lambda'_G(h)$ was defined in (24). To estimate the distribution of $\Lambda'_G(\mathcal{G}_G)$, some estimates of the derivative are required. Given sequence $\{b_n\}$, let

$$\hat{X}_0 = \{x \in \mathbb{R} : |\psi(\Pi^-(\mathbb{G}_0, \mathbb{G}_A))(x) - \psi(-\Pi^-(\mathbb{G}_0, \mathbb{G}_B))(x)| \leq b_n\},$$

and estimate the near-maximizers in u for each dependency bound and all x as in the previous example. Estimate the distribution by calculating resampled statistics, for $r = 1, \dots, R$,

$$\begin{aligned} \Lambda_r^* = \sqrt{n} \left(\int_{\hat{X}_0} \left[\sup_{u \in U_{\Pi^-(\mathbb{G}_{0n}, \mathbb{G}_{An})}(x, a_n)} \Pi(\mathbb{G}_{0n}^* - \mathbb{G}_{0n}, \mathbb{G}_{An}^* - \mathbb{G}_{An})(u, x) \right. \right. \\ \left. \left. + \sup_{u \in U_{-\Pi^-(\mathbb{G}_{0n}, \mathbb{G}_{Bn})}(x, a_n)} (-\Pi)(\mathbb{G}_{0n}^* - \mathbb{G}_{0n}, \mathbb{G}_{Bn}^* - \mathbb{G}_{Bn})(u, x) \right]_+^2 dm(x) \right)^{1/2}. \end{aligned}$$

A test can be conducted by comparing $\sqrt{n}\hat{\Lambda}(\mathbb{G}_n)$ to the $(1 - \alpha)$ -th quantile of the bootstrap distribution $\{\Lambda_r^*\}_{r=1}^R$. A simulation in the supplementary appendix illustrates the accurate size and power of this testing strategy. \square

Example C (Cost and profit functions continued). Now consider estimating a confidence band for a conjugate function. Suppose that we observe $\{X_i\}_{i=1}^n$ and have a parametric model $c(x, \theta)$, estimated using $c(x, \hat{\theta})$. Supposing that $\theta \in \mathbb{R}^p$, this is a VC-subgraph class of functions and $\sqrt{n}(c(\cdot, \hat{\theta}) - c(\cdot, \theta)) \rightsquigarrow \mathcal{G}_c$ where \mathcal{G}_c is a Gaussian process (van der Vaart, 1998, Example 19.7). To estimate the directional derivative $\lambda'(\mathcal{G}_c)$ where λ' is defined in (25), we need to estimate several quantities. Let $\hat{\mu} = \max_{p,x}(px - c(x, \hat{\theta}))$ and $\hat{\sigma} = \max_p \min_x(c(x) - px)$. For a slowly decreasing sequence a_n , decide whether $\hat{\mu} + a_n < \hat{\sigma}$, $|\hat{\mu} - \hat{\sigma}| \leq a_n$ or $\hat{\mu} > \hat{\sigma} + a_n$. Given this determination, estimate the derivative $\lambda'(\mathcal{G}_c)$ by $\hat{\sigma}'(\mathcal{G}_c^*)$, $\max\{\hat{\mu}'(\mathcal{G}_c^*), \hat{\sigma}'(\mathcal{G}_c^*)\}$ or $\hat{\mu}'(\mathcal{G}_c^*)$ respectively, where $\mathcal{G}_c^* = \sqrt{n}(c(\cdot, \hat{\theta}^*) - c(\cdot, \hat{\theta}))$. Given $\check{q}_{\lambda^*}(1 - \alpha/2)$, a quantile from the bootstrap sample $\{\check{\lambda}_r^*\}_{r=1}^R$ calculated as just described, we could estimate a symmetric level $(1 - \alpha)$ uniform confidence band for $\mathcal{L}(c)$ using $\mathcal{L}(c(\cdot, \hat{\theta})) \pm \check{q}_{\lambda^*}(1 - \alpha/2)$. \square

4.3 Local size control

It is also of interest to examine how these tests behave under sequences of distributions local to distributions that satisfy the null hypothesis. We consider sequences of local alternative distributions $\{P_n\}$ such that for each n , $\{Z_i\}_{i=1}^n$ are distributed according to P_n , and P_n converges towards a limit P that satisfies the null hypothesis. To describe this process, for $t \geq 0$ define a path $t \mapsto P_t$,

where P_t is an element of the space of distribution functions \mathcal{P} , such that

$$\lim_{t \rightarrow 0} \int \left(((dP_t)^{1/2} - (dP_0)^{1/2})/t - \frac{1}{2}h(dP_0)^{1/2} \right)^2 = 0, \quad (37)$$

where the score function $h \in L_2(F)$ satisfies $E[h] = 0$ and P_0 is a distribution that satisfies the null hypothesis. The direction that the sequence approaches the null is described asymptotically by the score h . We assume that by letting $t = c/r_n$ for $c \in \mathbb{R}$, we can parameterize distributions that are local to P_0 and for $t \geq 0$ denote $f(P_t)$ as the function f under distribution $f(P_t)$ so that the unmarked f described above can be rewritten $f = f(P_0)$. See, e.g., van der Vaart and Wellner (1996, Section 3.10.1) for more details.

The following assumption ensures that f remains suitably regular under such local perturbations to the null distribution.

A4. For all $c \in \mathbb{R}$,

- (a) For given c , there exists some $f'(c) \in \ell^\infty(\text{gr}A)$ such that $\|r_n(f(P_{c/r_n}) - f(P_0)) - f'(c)\|_\infty \rightarrow 0$, where P_{c/r_n} satisfy (37).
- (b) $r_n(f_n - f(P_{c/r_n})) \rightsquigarrow \mathcal{G}_{P_0}$ in $\ell^\infty(\text{gr}A)$, where for each n , $\{Z_i\}_{i=1}^n \sim P_{c/r_n}$.

Both parts of Assumption **A4** ensure that f_n behave regularly as distributions drift towards $P_0 \in \mathcal{P}_{00}^E$ or $P_0 \in \mathcal{P}_{00}^I$. These additional regularity conditions allow us to describe the size and local power properties of the test statistics.

Theorem 4.4. *Under Assumptions **A1-A4**:*

1. Suppose $j = 1$ or 3 , $P_0 \in \mathcal{P}_{00}^E$ and for each n , the observations $\{Z_i\}_{i=1}^n$ are distributed according to $P_n = P_{c/r_n} \in \mathcal{P}_{00}^E$. Suppose that the CDF of $\lambda'_{jf}(\mathcal{G}_{P_0})$ is continuous and increasing at $q_{1-\alpha}$, its $(1-\alpha)$ -th quantile. Then

$$\limsup_{n \rightarrow \infty} P_n \{r_n \lambda_j(f_n) > \hat{q}_{\lambda^*}(1-\alpha)\} \geq P_0 \{\lambda'_{jf}(\mathcal{G}_{P_0} + f'(c)) > q_{1-\alpha}\}.$$

This holds with equality if $q_{1-\alpha}$ is a continuity point of the CDF of $\lambda'_{jf}(\mathcal{G}_{P_0} + f'(c))$. If $c = 0$ then the limiting rejection probability is equal to α .

2. Suppose $j = 2$ or 4 , $P_0 \in \mathcal{P}_{00}^I$ and for each n , the observations $\{Z_i\}_{i=1}^n$ are distributed according to $P_n = P_{c/r_n} \in \mathcal{P}_{00}^I$. Suppose that the CDF of $\lambda'_{jf}(\mathcal{G}_{P_0})$ is continuous and increasing at its $(1-\alpha)$ -th quantile. Then

$$\limsup_{n \rightarrow \infty} P_n \{r_n \lambda_j(f_n) > \hat{q}_{\lambda^*}(1-\alpha)\} \leq \alpha.$$

Theorem 4.4 shows that the size of tests can be controlled locally to the null region (and the nominal rejection probability matches the intended probability) in only some cases. In particular, the tests for the null that $P_0 \in \mathcal{P}_{00}^E$ cannot be shown to control local size without more information about the direction of the local alternatives. This is because we must make assumptions about the underlying objective functions without being able to make similar assumptions about the corresponding value functions — note that Assumption **A4** is made with regard to $f(P)$, not $\psi(f(P))$. To see where the problem lies we may take λ_1 as an example. The result of Theorem 4.4 shows that when $P_n \in \mathcal{P}_{00}^E$ for each n ,

$$\sqrt{r_n}\lambda_1(f_n) \rightsquigarrow \lambda'_{1f}(\mathcal{G}_{P_0} + f'(c)).$$

Although $P \in \mathcal{P}_{00}^E$ implies $\lambda'_f(f'(c)) = 0$, as shown in the proof of Theorem 4.4, it does not imply that $f'(c) \equiv 0$. For example, if f are CDFs corresponding to a location-shift family of distributions, i.e., for all $c \in \mathbb{R}$, $G_c(x) = G_0(x - c)$ with differentiable densities that have square-integrable derivatives. Then since the h in (37) is $cg'_0(x)/g_0(x)$, the $f'(c)$ of assumption **A4** would be $-cg_0(x)$. This non-zero derivative can cause problems when optimizing it interacts with the absolute value function. Generally $|\sup_u (f(u) + g(u))| \not\leq |\sup_u f(u)| + |\sup_u g(u)|$ (take f and g strictly less than zero), and we do not achieve a simplification that would imply the probability on the right-hand side is less than or equal to α . We cannot assume regularity of $\psi(f)$ because this would require the existence of a derivative $\psi'_f(\cdot) \in \ell^\infty(X)$, but this derivative cannot generally exist. It may be that size control could be shown for special cases (for example when theory implies a shape for the value function), but not in general. On the other hand, for the same reason, the size of one-sided test statistics is as intended. That is, when the value function is found through maximization, the positive-part map in the definitions of λ_2 and λ_4 interacts with it in a way that maintains size control. In contrast, the negative-part map would not guarantee size control. A stronger result in this vein will be shown in the next section.

The inequality in the second part of this theorem results from the one-sidedness of the test statistics λ_2 and λ_4 . This is related to a literature in econometrics on moment inequality testing. Tests may exhibit size that is lower than nominal for local alternatives that are from the interior of the null region. A few possible solutions to this problem have been proposed. For example, one might evaluate the region where the moments appear to hold with equality, which leads to contact set estimates like in the bootstrap routine described above (Linton, Song, and Whang, 2010). Alternatively, we may alter the reference distribution by shifting it in the regions where equality does not seem to hold (Andrews and Shi, 2017).

4.4 Uniform size control

Uniformity of inference procedures in the data distribution was introduced in Giné and Zinn (1991) and Sheehy and Wellner (1992), and uniformity has become a topic of great interest in the econometrics literature (see, for example, Andrews and Guggenberger (2010), Linton, Song, and Whang (2010), Romano and Shaikh (2012) or Hong and Li (2018) for an application like ours). Under stronger assumptions that ensure the underlying f_n converge weakly to a limiting process uniformly over the set of possible data distributions, some of the above results can be extended from size control under local alternative distributions to size control that holds uniformly over the class of distributions that satisfy the null. To ensure uniformity we assume the following regularity conditions are satisfied.

A5. Suppose that $r_n(f_n - f) \rightsquigarrow \mathcal{G}_P$ uniformly over $P \in \mathcal{P}$:

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{g \in BL_1(\ell^\infty(\text{gr}A))} |\mathbb{E}[g(r_n(f_n - f(P)))] - \mathbb{E}[g(\mathcal{G}_P)]| = 0.$$

A6. Suppose that the limiting process \mathcal{G}_P is tight uniformly over $P \in \mathcal{P}$:

$$\lim_{M \rightarrow \infty} \sup_{P \in \mathcal{P}} P\{\|\mathcal{G}_P\|_\infty \geq M\} = 0.$$

A7. Suppose that on \mathcal{P}_{00}^I , the processes \mathcal{G}_P are regular for λ'_{2f} or λ'_{4f} in the following sense: for any $0 < \alpha < 1/2$, there exists a $\underline{q} > 0$ depending only on α such that

$$\sup_{P \in \mathcal{P}_{00}^I} P\{\lambda'_{jf}(\mathcal{G}_P) < \underline{q}\} < 1 - \alpha$$

and for any $s > 0$,

$$\limsup_{\eta \rightarrow 0^+} \sup_{P \in \mathcal{P}_{00}^I} P\{|\lambda'_{jf}(\mathcal{G}_P) - s| \leq \eta\} = 0.$$

In addition, suppose that over $\mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$, the processes \mathcal{G}_P are regular for statistics related to the entire domain X instead of X_0 — letting $\tilde{\lambda}'_{2f}(h) = \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X,\epsilon)} [h(u,x)]_+$ and $\tilde{\lambda}'_{4f}(h) = (\int_X \psi'_f(h)(x) dm(x))^{1/p}$, assume that the above two displays are satisfied with $\tilde{\lambda}'_{jf}(\mathcal{G}_P)$ in the place of $\lambda'_{jf}(\mathcal{G}_P)$.

A8. Suppose that conditional on the observations $\{Z_i\}_{i=1}^n$, $r_n(f_n^* - f_n) \rightsquigarrow \mathcal{G}_P$ uniformly over $P \in \mathcal{P}$: for all $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left\{ \sup_{g \in BL_1(\ell^\infty(\text{gr}A))} |\mathbb{E}[g(r_n(f_n^* - f_n)) | \{Z_i\}_{i=1}^n] - \mathbb{E}[g(\mathcal{G}_P)]| > \epsilon \right\} = 0.$$

The above assumptions help define the collection \mathcal{P} on which we may assert that inference is

valid uniformly for $P \in \mathcal{P}$. Assumptions **A5** and **A6** require that the sample objective functions converge to well-behaved limits uniformly on \mathcal{P} . Assumption **A7** requires that also the limiting distributions of the sample statistics $r_n(\lambda_j(f_n) - \lambda_j(f))$ are regular enough over \mathcal{P}_0^I that their CDFs do not have extremely large atoms (i.e., extending above the $(1 - \alpha)$ -th quantile of the distribution) and are strictly increasing at the relevant critical value. This assumption may be simplified when the \mathcal{G}_P are Gaussian since the functionals λ_2 and λ_4 are convex (Davydov, Lifshits, and Smorodina, 1998, Theorem 11.1). Assumption **A8** requires that the bootstrap objective functions also converge uniformly, as the original sample functions do. This last assumption may be implied by Assumption **A5**, see Lemma A.2 of Linton, Song, and Whang (2010), for example.

Theorem 4.5. *Under Assumptions **A1-A3** and **A5-A8**, for $j = 2$ or 4 ,*

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0^I} P \{r_n \lambda_j(f_n) > \hat{q}_{\lambda^*}(1 - \alpha)\} \leq \alpha.$$

Uniformity for $P \in \mathcal{P}$ is maintained by the one-sided statistics when optimization matches the “direction” of the test. That is, when the value function matches the positive-part map in λ_2 and λ_4 , we have uniformity over $P \in \mathcal{P}$, and analogously uniform inference would be possible for minimization and the negative-part map. Uniformity can be lost when optimization and the direction of the test do not match, which may be the case with two-sided tests.

In the next section we illustrate the usefulness of our results by providing details on the construction of uniform confidence bands around bound functions for the CDF of a treatment effect distribution. See the first section of the supplemental appendix for numerical simulations evaluating the finite-sample performance of the tests developed in Examples A and B.

5 The treatment effect distribution of job training on wages

This section illustrates the inference methods with an evaluation of a job training program. We construct upper and lower bounds for both the distribution and quantile function of the treatment effects, confidence bands for these bound function estimates, and describe a few inference results. This application uses an experimental job training program data set from the National Supported Work (NSW) Program, which was first analyzed by LaLonde (1986) and later by many others, including Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2001, 2005), Imbens (2003), and Firpo (2007).

Recent studies in statistical inference for features of the treatment effects distribution in the presence of partial identification include, among others, Firpo and Ridder (2008, 2019); Fan and Park (2009, 2010); Fan and Wu (2010); Fan and Park (2012); Fan, Sherman, and Shum (2014); Fan, Guerre, and Zhu (2017). These studies have concentrated on distributions of finite-dimensional

functionals of the distribution and quantile functions, including these functions themselves evaluated at a point. Additional work includes, among others, Gautier and Hoderlein (2011), Chernozhukov, Lee, and Rosen (2013), Kim (2014) and Chesher and Rosen (2015). Each of these papers provides pointwise inference methods for bounds on the distribution or quantile functions, and often for more complex objects. We contribute to this literature by applying the general results in this paper to provide uniform inference methods for the bounds developed by Makarov and others, and hope that it may indicate the direction that pointwise inference for bounds in more involved models may be extended to be uniformly valid.

The data set we use is described in detail in LaLonde (1986). We use the publicly available subset of the NSW study used by Dehejia and Wahba (1999). The program was designed as an experiment where applicants were randomly assigned into treatment. The treatment was work experience in a wide range of possible activities, such as learning to operating a restaurant or a child care center, for a period not exceeding 12 months. Eligible participants were targeted from recipients of Aid to Families With Dependent Children, former addicts, former offenders, and young school dropouts. The NSW data set consists of information on earnings and employment (outcome variables), whether treated or not.⁶ We consider male workers only and focus on earnings in 1978 as the outcome variable of interest. There are a total of 445 observations, where 260 are control observations and 185 are treatment observations. Summary statistics for the two parts of the data are presented in Table 1.

	Treatment Group				Control Group			
	Mean	Median	Min.	Max.	Mean	Median	Min.	Max.
Earnings (1978)	6,349.1	4,232.3	0.0	60,307.9	4,554.8	3,138.8	0.0	39,483.5
	(7,867.4)				(5,483.8)			

Table 1: Summary statistics for the experimental National Supported Work (NSW) program data.

To provide a more complete overview of the data, we also compute the empirical CDFs of the treatment and control groups in Figure 1. From this figure we note that the empirical treatment CDF stochastically dominates the empirical control CDF, and that there are a large number of zeros in each sample. In particular, $\mathbb{F}_{tr,n}(0) \approx 0.24$ and $\mathbb{F}_{co,n}(0) \approx 0.35$.

Suppose a binary treatment is independent of two potential outcomes (X_{co}, X_{tr}) , where X_{co} denotes outcomes under a control regime and X_{tr} denotes outcomes under a treatment, and X_{co} and X_{tr} have marginal distribution functions F_{co} and F_{tr} respectively. Suppose that interest is in the distribution of the treatment effect $\Delta = X_{tr} - X_{co}$ but we are unwilling to make any assumptions regarding the dependence between X_{co} and X_{tr} . In this section we study the relationship between

⁶The data set also contains background characteristics, such as education, ethnicity, age, and employment variables before treatment. Nevertheless, since we only use the experimental part of the data we refrain from using this portion of the data.

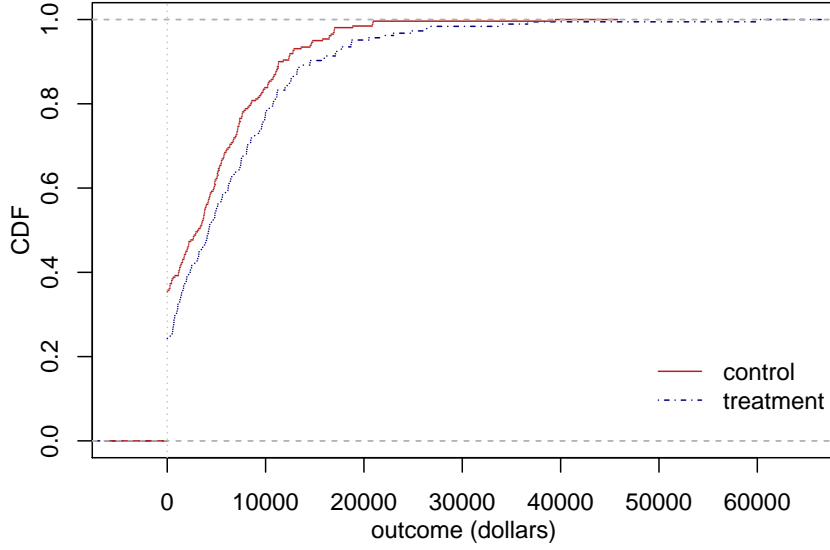


Figure 1: Empirical CDFs of treatment and control observations in the experimental NSW data. There are 260 control group observations and 185 treatment group observations, many outcomes equal to zero and the treatment group outcomes stochastically dominate the control group outcomes at first order.

the identifiable functions F_{co} and F_{tr} and functions that bound the distribution function of the unobservable random variable Δ .

$F_{\Delta}(\cdot)$ is not point-identified because the full bivariate distribution of (X_{co}, X_{tr}) is unidentified and the analyst has no knowledge of the dependence between potential outcomes. However, F_{Δ} can be bounded. Suppose we observe samples $\{X_{ki}\}_{i=1}^{n_k}$ for $k \in \{co, tr\}$. Define the empirical lower and upper bound functions

$$\mathbb{L}_{\Delta,n}(x) = \sup_{u \in \mathbb{R}} \{\mathbb{F}_{tr,n}(u) - \mathbb{F}_{co,n}(u - x)\} \quad (38)$$

$$\mathbb{U}_{\Delta,n}(x) = \inf_{u \in \mathbb{R}} \{1 + \mathbb{F}_{tr,n}(u) - \mathbb{F}_{co,n}(u - x)\}. \quad (39)$$

These are plug-in estimates of analogous population bounds L_{Δ} and U_{Δ} , and are similar to Example A because $F_{\Delta} = F_{(X_{tr}-X_{co})}$. We make the following assumptions on the observed samples of treatment and control observations.

B1. The observations $\{X_{co,i}\}_{i=1}^{n_{co}}$ and $\{X_{tr,i}\}_{i=1}^{n_{tr}}$ are iid samples and independent of each other and are distributed with marginal distribution functions F_{co} and F_{tr} respectively. Refer to the pair of distribution functions and their empirical distribution function estimates as $F = (F_{co}, F_{tr})$ and $\mathbb{F}_n = (\mathbb{F}_{co,n}, \mathbb{F}_{tr,n})$.

B2. The sample sizes n_{co} and n_{tr} increase in such a way that $n_k/(n_{co} + n_{tr}) \rightarrow \nu_k$ as $n_{co}, n_{tr} \rightarrow \infty$,

where $0 < \nu_k < 1$ for $k \in \{co, tr\}$. Define $n = n_{co} + n_{tr}$.

Under Assumptions **B1** and **B2**, it is a standard result (van der Vaart, 1998, Example 19.6) that for $k \in \{co, tr\}$, $\sqrt{n_k}(\mathbb{F}_{k,n} - F_k) \rightsquigarrow \mathcal{G}_k$, where \mathcal{G}_{co} and \mathcal{G}_{tr} are independent F_{co} - and F_{tr} -Brownian bridges, that is, mean-zero Gaussian processes with covariance functions $\rho_k(x, y) = F_k(x \wedge y) - F_k(x)F_k(y)$. This implies in turn that $\sqrt{n}(\mathbb{F}_n - F) \rightsquigarrow \mathcal{G}_F = (\mathcal{G}_{co}/\sqrt{\nu_{co}}, \mathcal{G}_{tr}/\sqrt{\nu_{tr}})$, where \mathcal{G}_F is a mean-zero Gaussian process with covariance process $\rho_F(x, y) = \text{diag}\{\rho_k(x, y)/\nu_k\}$.

Now we focus on the calculation of a uniform confidence band for L only. Because the bootstrap algorithm was described previously we only verify that the regularity conditions for this plan hold. Assumptions **B1** and **B2**, along with the above discussion of the weak convergence of $\sqrt{n}(\mathbb{F}_n - F)$ imply that Assumption **A1** is satisfied. The class of functions $\{(I(X \leq x), I(Y \leq y), x, y \in \mathbb{R})\}$ is uniform Donsker and the sequences a_n and b_n can be chosen to satisfy Assumption **A2** by making them converge to zero more slowly than $n^{-1/2}$. Assuming the weights are independent of the observations, assumption **A3** is satisfied by Lemma A.2 of Linton, Song, and Whang (2010), which implies that the bootstrap algorithm described above is consistent, as described in Theorem 4.3. It is also straightforward to verify that, under the high-level assumption (37), both parts of Assumption **A4** are satisfied — in the language of the assumption, $f(P_{c/\sqrt{n}})$ are the pair (F_{co}^n, F_{tr}^n) under the local probability distribution $P_{c/\sqrt{n}}$, and $f' = (\int_{-\infty}^{\cdot} h_{co}^c, \int_{-\infty}^{\cdot} h_{tr}^c)$ for direction h^c indexed by c and $\sqrt{n}(\mathbb{F}_n - F^n) \rightsquigarrow \mathcal{G}_F$ (van der Vaart and Wellner, 1996, Theorem 3.10.12).

The main objective is to provide uniform confidence bands for the CDF of the treatment effects distribution. We calculate the lower and upper bounds for the distribution function using the control and treatment samples as in equations (38)–(39). The bounds are computed on a grid with increments of \$100 dollars along the range of the common support of the bounds, which is roughly from -\$40,000 to \$60,000. We focus on the region between -\$40,000 and \$40,000, which contains almost all the observations (there is a single \$60K observation in the treated sample). The results are presented in Figure 2 and given by the black solid lines in the picture. The most prominent feature is that, as expected, the upper bound for the CDF of treatment effects stochastically dominates the corresponding lower bound.

Next, we compute the uniform confidence bands as described in the text. They are shown in Figure 2 as the dashed lines around the corresponding solid lines. Due to the large number of zero outcomes in both samples, these bounds have some interesting features that we discuss further. First, we note that for any ϵ greater than zero but smaller than the next smallest outcome (about \$45), $\mathbb{F}_{tr,n}(0) - \mathbb{F}_{co,n}(-\epsilon) = 0.24$, which explains the jump in the lower bound estimate near zero (it is really for a point in the grid just above zero). Likewise, for the same ϵ , $\mathbb{F}_{tr,n}(-\epsilon) - \mathbb{F}_{co,n}(0) = -0.35$, which explains the jump in the upper bound just below zero. Without these point masses at zero, both bounds would more smoothly tend towards 0 or 1. Second, the point masses at zero imply another feature of the bounds that can be discerned in the picture. The upper bound to the left of 0

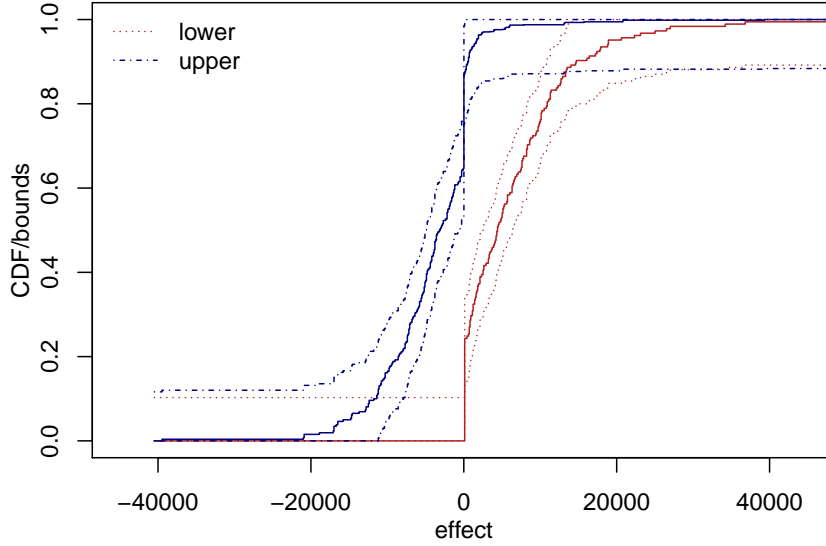


Figure 2: Bound functions and their uniform confidence bands. These confidence bands were constructed by inverting Kolmogorov-Smirnov-type test statistics as described in the text.

is the same as $1 - \mathbb{F}_{co,n}(-x)$ and the lower bound to the right of zero is the same as $\mathbb{F}_{tr,n}(x)$. Taking the lower bound as an example, for each $x > 0$ find the closest observation from the control sample y_{co,i^*} , and set $x^*(x) = X_{co,i^*} + \epsilon$, leading to the supremum $\mathbb{F}_{tr,n}(X_{co,i} + \epsilon)$ at every point where $X_{co,i} + \epsilon < X_{tr,j}$ for all j in the treated sample. It is identical to the empirical treatment CDF for the entire positive part of the support because the treatment first-order stochastically dominates the control. The situation would be different if there were a jump in the empirical control CDF at least as large as the jump in the empirical treatment CDF at zero. Because the opposite is the case for the upper bound, it does change slightly above the zero mark, tending from $1 - \mathbb{F}_{tr,n}(0) + \mathbb{F}_{co,n}(0)$ to 1 as x goes from 0 to the right.

We can also use the confidence bands for the bound functions to construct a confidence band for the true distribution function of treatment effects. This is shown in Figure 3. A $1 - \alpha$ confidence band can be constructed by using the upper $1 - \alpha/2$ limit of the upper bound confidence band and the lower $\alpha/2$ limit of the lower confidence band. This band is a uniform asymptotic confidence band for the true CDF, and uniform over correlation between the potential outcomes between samples. In other words, if \mathcal{P} is the collection of bivariate distributions that have marginal distributions F_{tr} and F_{co} , then

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} P \{F_{\Delta}(x) \in CB(x) \text{ for all } x\} \geq 1 - \alpha.$$

This confidence band is likely conservative, since

$$\begin{aligned}
& P\{\exists x : F_{\Delta}(x) \notin CB(x)\} \\
&= P\left\{\{\exists x : F_{\Delta}(x) < \mathbb{L}_{\Delta,n}(x) - q_{L,1-\alpha/2}^*/\sqrt{n}\} \cup \{\exists x : F_{\Delta}(x) > \mathbb{U}_{\Delta,n}(x) + q_{U,1-\alpha/2}^*/\sqrt{n}\}\right\} \\
&\leq P\left\{\{\exists x : L_{\Delta}(x) < \mathbb{L}_{\Delta,n}(x) - q_{L,1-\alpha/2}^*/\sqrt{n}\} \cup \{\exists x : U_{\Delta}(x) > \mathbb{U}_{\Delta,n}(x) + q_{U,1-\alpha/2}^*/\sqrt{n}\}\right\}.
\end{aligned}$$

See Kim (2014) and Firpo and Ridder (2019) for a more thorough discussion of the sense in which these bounds are not uniformly sharp for the treatment effect distribution function. We leave more sophisticated, potentially tighter confidence bands for future research. Note that the technique of Imbens and Manski (2004) cannot be used to tighten these bounds, because the parameter, a function, could violate the null hypothesis at both sides of the confidence band.

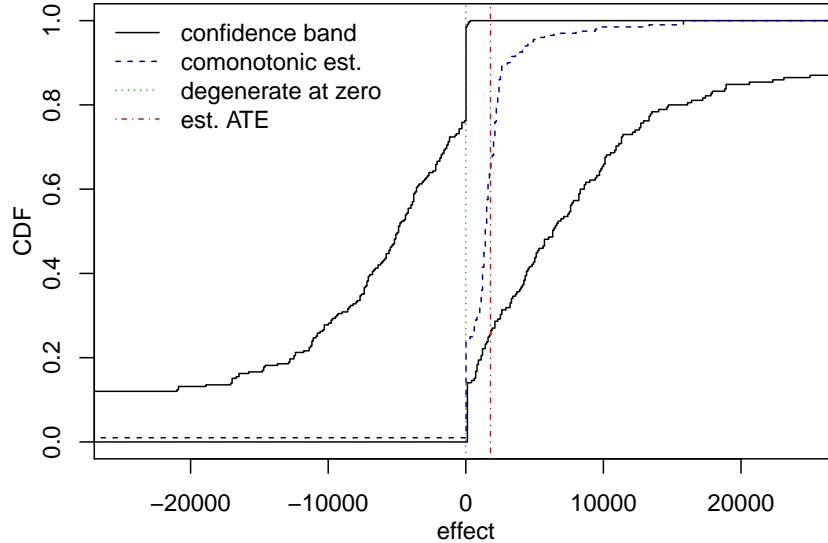


Figure 3: A uniform confidence band for the true treatment effect CDF. This bound is constructed by using the lower $\alpha/2$ limit of the lower bound and the upper $\alpha/2$ limit of the upper bound. A few other estimates of the treatment effect distribution are made: the vertical line at zero represents an informal hypothesis test that the effect is zero across the entire distribution, and is rejected (see the calculations at the beginning of this section about the nontrivial parts of the bounds to see why this is so). The vertical line to the right of zero is the average treatment effect, and it can be seen that this average ignores some variation in treatment effect outcomes. The dotted curve in between the bounds is the same as the (inverted) quantile treatment effect, which is equivalent to assuming rank invariance between potential treatment and control outcomes.

We plot some other features in this figure for context. First, the dotted vertical line is positioned at $y = 0$, and it can be seen that we (just) reject the null hypothesis $H_0 : P\{\Delta = 0\} = 1$. This hypothesis is closest to non-rejection, and it is clearer that one should reject the null that the treatment effect distribution is degenerate at any other point besides zero. This supports the notion that treatment effect heterogeneity is an important feature of these observations, especially

because this band is completely agnostic about the form of the joint distribution. On the other hand, by examining the bands at horizontal levels, it can be seen that for the median effect and a wide interval in the center of the distribution, the hypothesis of zero treatment effect cannot be rejected (although these are uniform bounds and not tests of individual quantile levels).

The final feature in the figure is the dashed curve that represents the estimate that one would make under the assumption of comonotonicity (or rank invariance) — the assumption that, had an individual been moved from the treatment to the control group, their rank in the control would be the same as their observed rank. Under this strong assumption the quantile treatment effects are the quantiles of the treatment distribution and they can be inverted to make an estimate. Clearly, the estimate under this assumption is just one point-identified treatment effect distribution function of many.

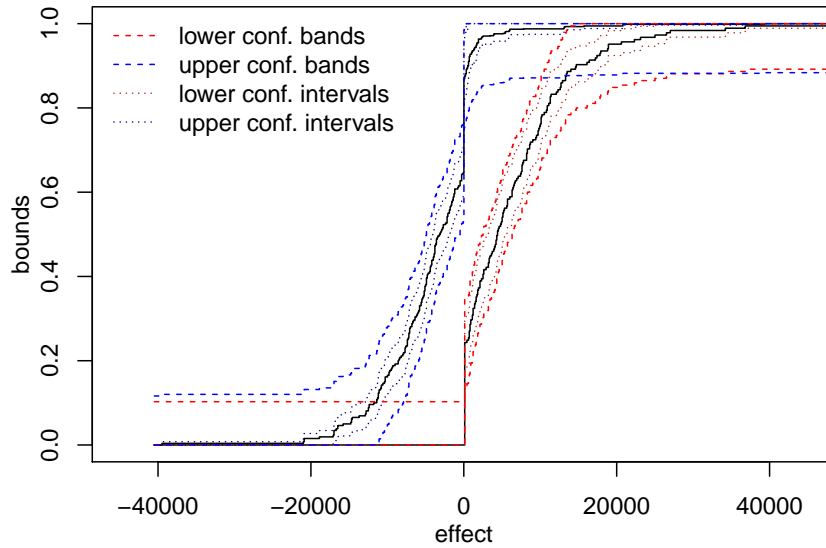


Figure 4: Uniform confidence bands for the CDF of treatment effects plotted with a collection of confidence intervals at treatment outcome levels. The confidence intervals are narrower because they represent confidence statements at each treatment effect individually, while the uniform bands give confidence statements about where the entire bound function lies. Pointwise confidence intervals were calculated using the method proposed in Fan and Park (2010).

Finally, to provide context for the uniform confidence band results within the literature on inference for bounds, we compare the proposed uniform bands to the pointwise confidence intervals suggested by Fan and Park (2010). We used Bickel and Sakov (2008)’s automatic procedure to choose subsample size and constructed confidence intervals for each individual point in the grid of the treatment effect support. This collection of pointwise confidence intervals are plotted along with uniform confidence bands in Figure 4. The results show that the uniform bands are farther from the bound estimates than the set of pointwise confidence intervals.

6 Conclusion

This paper develops uniform statistical inference methods for optimal value functions, that is, functions constructed by optimizing an objective function in one argument over all values of the remaining arguments. Value functions can be seen as a nonlinear transformation of the objective function. The map from objective function to value function is not Hadamard differentiable, but statistics used to conduct uniform inference are Hadamard directionally differentiable. We establish the asymptotic properties of nonparametric plug-in estimators of these uniform test statistics and develop a resampling technique to conduct practical inference. Examples involving dependency bounds for treatment effects distributions are used for illustration. Finally, we provide an application to the evaluation of a job training program, estimating a conservative uniform confidence band for the distribution function of the program's treatment effect without making any assumptions about the dependence between potential outcomes.

Acknowledgements

We would like to thank Timothy Christensen, Zheng Fang, Hidehiko Ichimura, Lynda Khalaf, Alexandre Poirier, Pedro H.C. Sant'Anna, Andres Santos, Rami Tabri, Brennan Thompson, Tiemen Woutersen, and seminar participants at the University of Kentucky and New York Camp Econometrics XIV for helpful comments and discussions. We would also like to thank the referees who helped improve the paper, especially the comments of one referee that improved subsection 4.4 considerably. All the remaining errors are ours. This research was enabled in part by support from the Shared Hierarchical Academic Research Computing Network (SHARCNET: <https://www.sharcnet.ca>) and Compute Canada (<http://www.computecanada.ca>). Code used to implement the methods described in this paper and to replicate the simulations can be found at https://github.com/tmparker/uniform_inference_for_value_functions.

Supplementary appendix to “Uniform inference for value functions”

This supplemental appendix contains the results of simulation experiments designed to test the size and power of tests described in the main text, and proofs of theorems and lemmas stated in the main text.

A Simulation study

We investigate the finite-sample behavior of the tests using examples A and B the main text. Specifically, we consider uniform inference for a lower bound function L_{X+Y} for the distribution function F_{X+Y} without knowledge of the dependence between X and Y by looking at empirical rejection probabilities for a two-sided test that $L_{X+Y} = L_0$ as in Example A. Then we investigate the performance of one-sided testing procedures through the stochastic dominance test as outlined in Example B. Code to reproduce these simulation experiments may be found through the third author’s personal website.⁷ In all cases, we test size and power in tests with nominal size 5% against local alternatives for samples of size 100, 500 and 1000 with respectively 499, 999 and 1999 bootstrap repetitions for each sample size, and use 1,000 simulation repetitions to calculate empirical rejection probabilities.

A.1 Tuning parameter selection

In this experiment, simulated observations are normally distributed and we shift the location of one distribution to examine size and power properties of tests using statistic (34) from Example A. The bound function L_{X+Y} was estimated using sample data and standard empirical distribution functions. The bounds must be equal to zero for small enough arguments and equal to one for large enough arguments, and are monotonically increasing in between. However, \mathbb{L}_n may change at all possible $X_i - Y_j$ combinations, and computing the function on all $n_X \times n_Y$ points could be computationally prohibitive. Therefore we compute the bound functions on a reasonably fine grid $\{x_k\}_{k=1}^K$ (we used increments of 0.05 in the experiment).

In order to make calculations as efficient as possible, it is helpful to know some features of the support of the bounds given sample data, where the support of a bound is the region where it is strictly inside the unit interval. The empirical lower bound at a given x may be written as the maximum of the function $\mathbb{F}_{Xn}(\cdot) - \mathbb{F}_{(x-Y)n}(\cdot)$, which takes steps of size $+1/n_X$ at sample observations $\{X_i\}_{i=1}^{n_X}$ and, given a value of x , steps of size $-1/n_Y$ at shifted observations $\{x -$

⁷https://github.com/tmperker/uniform_inference_for_value_functions.

$Y_j\}_{h,j=1}^{n_Y}$. This function is not right-continuous, but this is not a problem when Y is distributed continuously and calculations are done on a grid. The upper endpoint of the support of the lower bound is the smallest value for which it is equal to 1. If for some (large) x , $x - Y_j \geq X_i$ for all i and j , then the shape of $\mathbb{F}_{X_n}(u) - \mathbb{F}_{(x-Y)_n}(u)$ rises monotonically to one, then falls, as u increases. The x that satisfy this condition are $x \geq \max_{i,j}\{X_i + Y_j\} = \max_i X_i + \max_j Y_j$. To find the minimum of the support, the maximum value of x such that $\mathbb{L}_n(x) = 0$, note that the function $\mathbb{F}_{X_n}(u) - \mathbb{F}_{(x-Y)_n}(u)$ is always equal to zero for some u , but zero is the maximum only when \mathbb{F}_{X_n} first-order stochastically dominates $\mathbb{F}_{(x-Y)_n}$ and the functions are equal at at least one argument. Therefore the smallest x in the support of \mathbb{L}_n is the smallest x that shifts the $\mathbb{F}_{(x-Y)_n}$ distribution function enough so that it is dominated by \mathbb{F}_{X_n} . This can be estimated by computing the minimum that is, $\min_k\{\hat{Q}_{X_n}(\tau_k) - \hat{Q}_{(-Y)_n}(\tau_k)\}$ over a common set of quantiles $\{\tau_k\}_k$. Similar logic can also be applied to restrict the size of the grid to as small a search area as possible for the one-sided test experiment below.

We test the null hypothesis that the lower bound of the treatment effect distribution corresponds to the bound associated with two standard normal marginal distributions, which is $L_0(x) = 2\Phi(x/2) - 1$ for $x > 0$ and zero otherwise, where Φ is the CDF of the standard normal distribution.

We used this design to choose a sequence $\{a_n\}$ used to estimate the set of ϵ -maximizers in the estimation of the derivative ψ' and $\{b_n\}$ used to estimate contact sets in one-sided derivative estimates. We look for the tuning parameters a_n and b_n that provide size as close as possible to the nominal 5% rejection probability. We chose among $a_n = c_a \sqrt{\log(\log(n))/n}$ and $b_n = c_b \log(\log(n))/\sqrt{n}$ where $n = n_0 + n_1$ and $c_a, c_b \in \{0.5, 1.0, 1.5, \dots, 5\}$. Empirical rejection probabilities are reported in Tables 2 and 3.

We decided on $a_n = 0.5 \log(\log(n))/\sqrt{n}$, since in general smaller c_a resulted in more accurate size. For contact set estimation we chose $b_n = 3.5 \log(\log(n))/\sqrt{n}$, which is similar to the sequence used in Linton, Song, and Whang (2010), who concentrated on estimating the contact set in a stochastic dominance experiment.⁸

A.2 Two-sided tests

We verify that our proposed method provides accurate coverage probability for uniform confidence bands and power against local alternatives. We test the null hypothesis that the lower bound of the treatment effect distribution corresponds to the bound associated with two standard normal

⁸They used sequences of the form $c \log \log(\bar{n})/\sqrt{\bar{n}}$, where $\bar{n} = (n_1 + n_2)/2$, where they suggested c between 3 and 4. Using $c = 4$ but adjusting the formula to depend on n , the sum of the two samples, we have $4/\sqrt{2} \approx 2.8$, so there is some justification for setting the constant for b_n lower. In our simulations there was a slight improvement when using constant 3.5 instead of 3.

tab	$c_a = 0.5$	$c_a = 1$	$c_a = 1.5$	$c_a = 2$	$c_a = 2.5$	$c_a = 3$	$c_a = 3.5$	$c_a = 4$	$c_a = 4.5$	$c_a = 5$
s. size 100										
$c_b = 0.5$	0.8	-1.0	-1.0	-1.3	-0.8	-1.0	-1.4	-2.2	-1.6	-1.8
$c_b = 1$	0.8	0.5	-0.8	-1.8	-1.3	-1.9	-1.6	-2.0	-1.4	-1.5
$c_b = 1.5$	1.6	-1.3	-1.3	-1.7	-2.0	-2.2	-1.7	-2.6	-1.9	-2.1
$c_b = 2$	0.6	-0.5	0.2	-1.2	-2.1	-2.0	-2.0	-0.9	-1.2	-2.7
$c_b = 2.5$	0.7	-0.5	-1.7	-1.5	-1.3	-2.2	-1.8	-2.0	-1.3	-2.4
$c_b = 3$	2.0	1.4	-1.7	-0.1	-1.0	-2.2	-1.6	-2.3	-1.7	-2.0
$c_b = 3.5$	1.2	-0.7	0.1	-1.3	-0.8	-1.4	-1.5	-1.3	-2.0	-1.6
$c_b = 4$	0.6	-1.5	-0.3	-1.5	-2.0	-1.9	-1.6	-2.7	-2.0	-2.2
$c_b = 4.5$	1.0	-1.0	-1.5	-1.4	-2.6	-2.0	-1.8	-1.5	-2.2	-1.9
$c_b = 5$	0.4	-0.5	-0.8	-2.4	-1.7	-0.6	-1.7	-1.3	-1.4	-2.8
s. size 500										
$c_b = 0.5$	1.4	-0.4	-1.1	-2.1	-1.7	-2.0	-0.6	-2.4	-2.5	-1.7
$c_b = 1$	0.4	-1.2	-0.9	-1.1	-1.6	-1.9	-1.8	-2.3	-1.5	-2.4
$c_b = 1.5$	1.2	-1.8	-2.5	-1.4	-1.1	-2.5	-1.6	-1.4	-2.1	-1.2
$c_b = 2$	-0.4	-1.5	-1.8	-2.3	-2.4	-2.8	-2.3	-2.5	-1.8	-3.0
$c_b = 2.5$	0.9	-0.4	-1.9	-2.4	-2.2	-1.2	-2.4	-1.9	-2.1	-2.2
$c_b = 3$	0.5	-1.4	-1.6	-2.6	-1.5	-2.4	-1.8	-2.1	-2.5	-1.8
$c_b = 3.5$	-0.1	-0.3	-1.3	-1.7	-1.7	-1.5	-1.4	-1.2	-1.8	-2.1
$c_b = 4$	1.1	-0.8	-1.4	-1.7	-1.6	-2.6	-1.6	-2.2	-1.8	-1.5
$c_b = 4.5$	0.7	-2.1	-0.4	-1.7	-2.2	-1.7	-3.0	-2.4	-1.2	-2.7
$c_b = 5$	0.5	-1.2	-0.7	-1.4	-2.4	-2.6	-1.8	-2.5	-1.7	-2.4
s. size 1000										
$c_b = 0.5$	0.0	-1.5	-0.8	-1.6	-1.8	-2.3	-2.3	-2.4	-1.7	-2.5
$c_b = 1$	0.3	-0.4	-0.9	-2.3	-2.4	-1.5	-1.7	-1.1	-1.3	-1.9
$c_b = 1.5$	-0.9	-1.2	-1.3	-1.5	-2.1	-1.7	-1.8	-2.4	-1.6	-2.6
$c_b = 2$	-1.7	-0.8	-2.3	-0.8	-2.6	-1.7	-0.9	-2.4	-2.6	-2.4
$c_b = 2.5$	0.7	-1.8	-1.4	-1.9	-2.9	-2.1	-1.8	-1.1	-2.8	-2.5
$c_b = 3$	0.1	-1.2	-1.8	-2.2	-2.1	-1.9	-2.2	-2.1	-1.7	-1.6
$c_b = 3.5$	-0.1	-1.4	-1.2	-1.2	-0.8	-1.9	-2.5	-2.8	-1.2	-2.0
$c_b = 4$	-0.1	-1.5	-1.8	-2.2	-1.2	-2.2	-1.8	-1.7	-2.4	-2.0
$c_b = 4.5$	1.1	-1.2	-2.3	-2.1	-1.5	-0.7	-3.2	-2.0	-2.1	-1.6
$c_b = 5$	0.0	-2.0	-1.4	-3.1	-2.1	-1.6	-2.4	-2.4	-2.2	-1.9

Table 2: Empirical minus nominal 5% rejection frequency for a test that $L_{X+Y} \equiv L_0$ for known L_0 . Numbers in the table closer to zero are better. Rows describe tuning parameters for contact set estimation and columns for epsilon-maximizer estimation.

tab	$c_a = 0.5$	$c_a = 1$	$c_a = 1.5$	$c_a = 2$	$c_a = 2.5$	$c_a = 3$	$c_a = 3.5$	$c_a = 4$	$c_a = 4.5$	$c_a = 5$
s. size 100										
$c_b = 0.5$	1.1	-2.3	-1.7	-2.8	-2.9	-2.2	-2.3	-4.6	-2.3	-4.0
$c_b = 1$	1.8	-0.1	-1.6	-3.2	-2.6	-3.4	-4.0	-4.1	-3.2	-3.2
$c_b = 1.5$	1.3	-1.9	-1.7	-2.5	-1.8	-2.7	-2.8	-4.6	-3.1	-4.0
$c_b = 2$	0.9	-1.6	-1.2	-2.6	-3.2	-3.5	-2.3	-4.1	-3.0	-5.3
$c_b = 2.5$	0.8	-1.5	-1.8	-1.5	-2.4	-4.1	-3.3	-3.3	-1.9	-4.4
$c_b = 3$	3.0	0.9	-2.8	-2.9	-2.4	-3.5	-3.6	-4.5	-3.4	-4.2
$c_b = 3.5$	1.2	-1.5	-1.9	-3.4	-1.0	-2.4	-3.9	-3.1	-3.1	-3.2
$c_b = 4$	1.1	-1.0	-1.6	-2.5	-3.4	-4.6	-3.2	-5.0	-3.8	-3.3
$c_b = 4.5$	2.1	-0.4	-3.2	-3.0	-4.6	-3.3	-4.0	-2.4	-4.8	-3.3
$c_b = 5$	-0.8	-1.5	-2.1	-4.5	-3.2	-1.5	-3.0	-2.4	-2.8	-4.4
s. size 500										
$c_b = 0.5$	-0.1	-0.6	-2.3	-2.5	-4.3	-4.2	-2.5	-4.3	-4.2	-3.4
$c_b = 1$	0.3	-2.6	-2.2	-1.7	-2.7	-3.2	-4.6	-4.0	-3.3	-4.9
$c_b = 1.5$	1.4	-2.6	-2.8	-2.1	-2.7	-3.7	-2.9	-3.4	-4.5	-3.4
$c_b = 2$	0.5	-2.0	-2.7	-3.5	-3.6	-4.3	-3.2	-4.5	-3.4	-4.3
$c_b = 2.5$	1.3	-0.5	-3.4	-3.4	-4.2	-2.7	-4.3	-3.2	-4.2	-3.9
$c_b = 3$	-0.3	-3.6	-2.4	-5.1	-3.1	-4.0	-3.5	-2.5	-4.0	-3.8
$c_b = 3.5$	-0.1	-0.6	-2.6	-3.8	-2.6	-3.1	-2.7	-2.6	-4.0	-3.4
$c_b = 4$	1.3	-2.2	-3.2	-2.4	-3.2	-4.3	-3.6	-5.2	-3.8	-3.9
$c_b = 4.5$	1.9	-3.7	-2.0	-2.8	-4.5	-3.7	-4.5	-3.9	-2.5	-4.3
$c_b = 5$	0.8	-2.2	-1.6	-2.9	-4.3	-4.1	-3.2	-4.1	-2.9	-4.1
s. size 1000										
$c_b = 0.5$	-1.2	-3.2	-2.0	-3.9	-3.5	-4.3	-2.8	-4.1	-3.9	-4.8
$c_b = 1$	1.0	-1.7	-1.7	-4.9	-3.4	-2.8	-2.6	-3.6	-2.1	-2.9
$c_b = 1.5$	-1.0	-1.9	-3.5	-3.1	-2.9	-3.1	-3.0	-4.1	-3.5	-5.0
$c_b = 2$	-1.6	-2.0	-4.3	-2.3	-3.5	-3.6	-2.3	-5.0	-3.9	-4.6
$c_b = 2.5$	0.5	-2.5	-2.0	-3.0	-4.6	-4.7	-3.0	-3.5	-5.9	-4.6
$c_b = 3$	0.5	-2.6	-3.0	-3.5	-3.9	-3.6	-4.0	-4.1	-2.5	-3.7
$c_b = 3.5$	-0.2	-1.9	-2.0	-1.8	-1.6	-3.8	-4.7	-4.3	-3.0	-4.4
$c_b = 4$	-0.5	-1.9	-2.5	-3.6	-2.5	-3.3	-3.5	-3.4	-3.6	-3.6
$c_b = 4.5$	1.9	-2.2	-4.0	-3.3	-3.3	-2.1	-4.5	-3.8	-4.4	-2.5
$c_b = 5$	-0.1	-2.8	-1.8	-5.0	-3.2	-3.0	-3.6	-4.6	-4.5	-4.4

Table 3: Empirical minus nominal 10% rejection frequency for a test that $L_{X+Y} \equiv L_0$ for known L_0 . Numbers in the table closer to zero are better. Rows describe tuning parameters for contact set estimation and columns for epsilon-maximizer estimation.

marginal distributions, which is $L_0(x) = 2\Phi(x/2) - 1$ for $x > 0$ and zero otherwise, where Φ is the CDF of the standard normal distribution. The upper bound is symmetric so it is sufficient to examine only the lower bound test. Alternatives are local location-shift alternatives with the mean $\mu_X = 0$ while the mean μ_Y changes. In previous experiments we found good power against alternatives with $\mu_Y = k/\sqrt{n}$ for varying k , but thanks to referee suggestions we also consider a wider range of local alternative specifications: corresponding to those alternatives used in Hong and Li (2018), we let $\mu_Y \in \{0, n^{-1}, n^{-2/3}, n^{-1/2}, n^{-1/3}, n^{-1/6}, n^{-1/10}, 2\}$, and also consider these means multiplied by negative 1. These are tests that should reject for any alternative $\mu_Y \neq 0$, but nontrivial power against all local alternatives cannot be guaranteed for them. Results from this simulation experiment are collected in Tables 4, 5 and 6.

Hong and Li (2018) suggest estimating the directional derivative nonparametrically using

$$\hat{\lambda}'_{HL} = \frac{\lambda(\mathbb{F}_n + \epsilon_n \sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n)) - \lambda(\mathbb{F}_n)}{\epsilon_n},$$

where $\epsilon_n \rightarrow 0$ and $n^{1/2}\epsilon_n \rightarrow \infty$. However, it is difficult to use this estimate here, because it is difficult to impose a null hypothesis on the form of the derivative in this experiment. Specifically, the derivative (21) relies on L_0 . Although L_0 does not need to be known in the analytic construction of a derivative estimate, it is needed to estimate the derivative numerically using the above formula. We addressed this problem in two ways. First, we imposed the hypothesis that $L_{X+Y} = L_0$, that is, using statistic (34) exactly as written with the hypothesized L_0 in $\hat{\lambda}'_{HL}$. Second, we used a plug-in estimate of L_{X+Y} . Recalling (20), let

$$\begin{aligned} \tilde{\lambda}'_{HL} &= \epsilon_n^{-1} \left(\sup_{x \in \mathbb{R}} \left| \max_{u \in \mathbb{R}} \{ \sup \Pi^+(\mathbb{F}_n + \epsilon_n \sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n))(u, x), 0 \} - \mathbb{L}_n(x) \right| \right. \\ &\quad \left. - \sup_{x \in \mathbb{R}} \left| \max_{u \in \mathbb{R}} \{ \sup \Pi^+(\mathbb{F}_n)(u, x), 0 \} - \mathbb{L}_n(x) \right| \right) \\ &= \epsilon_n^{-1} \sup_{x \in \mathbb{R}} \left| \max_{u \in \mathbb{R}} \{ \sup \Pi^+(\mathbb{F}_n + \epsilon_n \sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n))(u, x), 0 \} - \mathbb{L}_n(x) \right|, \end{aligned}$$

where the second equality arises from the fact that $\mathbb{L}_n(x) = \max\{\sup_{u \in \mathbb{R}} \Pi^+(\mathbb{F}_n)(u, x), 0\}$ by definition. Neither of these options is very satisfactory, since L_0 is unlikely to be known and using \mathbb{L}_n as in $\tilde{\lambda}'_{HL}$ results in a statistic that appears to estimate the derivative inside the L_∞ norm. The method proposed in Hong and Li (2018) requires ϵ_n to be larger than $n^{-1/2}$, so size distortions for smaller choices of ϵ_n may be expected. However, those ϵ_n were also considered in their paper so we included them here.

Location	Analytic	Numeric (null known)				Numeric (null estimated)			
		$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$	$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$
-2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$-n^{-1/10}$	0.968	0.976	0.984	0.992	0.997	0.967	0.977	0.981	0.973
$-n^{-1/6}$	0.816	0.841	0.893	0.938	0.981	0.806	0.842	0.860	0.824
$-n^{-1/3}$	0.301	0.335	0.441	0.568	0.739	0.283	0.330	0.367	0.297
$-n^{-1/2}$	0.130	0.155	0.245	0.363	0.570	0.118	0.151	0.173	0.131
$-n^{-2/3}$	0.093	0.110	0.180	0.294	0.500	0.081	0.108	0.120	0.092
$-n^{-1}$	0.072	0.085	0.161	0.266	0.482	0.063	0.088	0.098	0.069
Null is true									
0	0.058	0.072	0.141	0.273	0.483	0.053	0.069	0.077	0.059
n^{-1}	0.048	0.061	0.120	0.246	0.433	0.043	0.057	0.066	0.049
$n^{-2/3}$	0.037	0.052	0.118	0.237	0.432	0.030	0.046	0.055	0.035
$n^{-1/2}$	0.030	0.050	0.132	0.266	0.427	0.025	0.036	0.045	0.030
$n^{-1/3}$	0.017	0.057	0.200	0.322	0.419	0.016	0.020	0.024	0.019
$n^{-1/6}$	0.242	0.497	0.750	0.850	0.836	0.222	0.283	0.320	0.254
$n^{-1/10}$	0.662	0.863	0.947	0.972	0.964	0.646	0.714	0.740	0.665
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4: Empirical rejection frequencies for a test that $L_{X+Y} \equiv L_0$ for known L_0 . Analytic derivative estimates are described in the text, while Numeric derivative estimates use the method suggested by Hong and Li (2018). Each row label gives the location parameter of one of the distributions, while the null assumes both parameters are zero. Samples of size 100, 499 bootstrap repetitions per test, 1000 simulation repetitions.

Location	Analytic	Numeric (null known)				Numeric (null estimated)			
		$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$	$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$
-2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$-n^{-1/10}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$-n^{-1/6}$	0.997	0.997	0.998	1.000	1.000	0.995	0.997	0.998	0.997
$-n^{-1/3}$	0.383	0.388	0.483	0.655	0.813	0.356	0.416	0.438	0.383
$-n^{-1/2}$	0.106	0.114	0.182	0.339	0.549	0.096	0.122	0.139	0.108
$-n^{-2/3}$	0.050	0.056	0.103	0.256	0.465	0.041	0.064	0.073	0.051
$-n^{-1}$	0.061	0.065	0.124	0.272	0.452	0.054	0.067	0.074	0.060
Null is true									
0	0.058	0.060	0.108	0.257	0.453	0.051	0.066	0.072	0.055
n^{-1}	0.050	0.056	0.111	0.240	0.430	0.042	0.057	0.069	0.050
$n^{-2/3}$	0.040	0.041	0.100	0.215	0.397	0.034	0.049	0.059	0.035
$n^{-1/2}$	0.035	0.042	0.113	0.241	0.362	0.027	0.039	0.044	0.034
$n^{-1/3}$	0.064	0.116	0.281	0.455	0.494	0.057	0.074	0.088	0.067
$n^{-1/6}$	0.904	0.953	0.989	0.996	0.991	0.896	0.924	0.933	0.909
$n^{-1/10}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 5: Empirical rejection frequencies for a test that $L_{X+Y} \equiv L_0$ for known L_0 . Analytic derivative estimates are described in the text, while Numeric derivative estimates use the method suggested by Hong and Li (2018). Each row label gives the location parameter of one of the distributions, while the null assumes both parameters are zero. Samples of size 500, 999 bootstrap repetitions per test, 1000 simulation repetitions.

Location	Analytic	Numeric (null known)				Numeric (null estimated)			
		$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$	$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$
-2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$-n^{-1/10}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$-n^{-1/6}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$-n^{-1/3}$	0.432	0.434	0.555	0.708	0.852	0.411	0.460	0.506	0.438
$-n^{-1/2}$	0.108	0.107	0.172	0.349	0.536	0.095	0.115	0.132	0.107
$-n^{-2/3}$	0.072	0.074	0.135	0.271	0.450	0.061	0.079	0.088	0.070
$-n^{-1}$	0.041	0.044	0.098	0.267	0.428	0.037	0.048	0.057	0.041
Null is true									
0	0.046	0.045	0.107	0.245	0.435	0.037	0.053	0.060	0.045
n^{-1}	0.051	0.053	0.116	0.259	0.419	0.046	0.058	0.068	0.050
$n^{-2/3}$	0.046	0.049	0.108	0.239	0.408	0.039	0.050	0.060	0.041
$n^{-1/2}$	0.027	0.030	0.103	0.264	0.387	0.024	0.032	0.035	0.026
$n^{-1/3}$	0.089	0.131	0.328	0.520	0.539	0.074	0.099	0.115	0.087
$n^{-1/6}$	0.994	0.998	0.999	1.000	0.999	0.993	0.995	0.996	0.995
$n^{-1/10}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 6: Empirical rejection frequencies for a test that $L_{X+Y} \equiv L_0$ for known L_0 . Analytic derivative estimates are described in the text, while Numeric derivative estimates use the method suggested by Hong and Li (2018). Each row label gives the location parameter of one of the distributions, while the null assumes both parameters are zero. Samples of size 1000, 1999 bootstrap repetitions per test, 1000 simulation repetitions.

The tables indicate power loss only for alternatives that are very small and positive (for locations of size $n^{-2/3}$ and $n^{1/2}$). The numerically estimated reference distribution with known L_0 was usually outperformed by the analytically estimated distribution and that using \mathbb{L}_n in the place of L_0 . When using a known L_0 , the choice of ϵ_n is crucial. On the other hand, the numerical derivatives do well when L is estimated and ϵ_n converges either slowly or very quickly to zero — surprisingly, the last column, with estimated null and $\epsilon_n = n^{-1}$ is closest to the behavior of the test with analytically estimated reference distribution. The analytic estimate does not require knowledge of L_0 and performs as well as the best-performing numeric estimates.

A.3 One-sided tests

In a second experiment we simulate uniformly distributed samples and test the condition that implies stochastic dominance of F_{Δ_A} over F_{Δ_B} using statistic (36). A control sample is made up of uniform observations on the unit interval, as is treatment B , while treatment A is distributed uniformly on $[\mu, \mu + 1]$. We test the hypothesis that $L_A - U_B \leq 0$. When $\mu = -1$, $L_A - U_B \equiv 0$ so that value of $\mu = -1$ represents a least-favorable null hypothesis.⁹ When $\mu > -1$, the null hypothesis is satisfied with a strict inequality and should not be rejected, while for $\mu < -1$ the null should be rejected. We examine local alternatives around the central value $\mu = -1$, which is normalized to zero in Tables 7-9. We test size and power against local alternatives for samples of size 100, 500 and 1000 with respectively 499, 999 and 1999 bootstrap repetitions for each sample size and a grid with increment size 0.02. 1,000 simulation repetitions were used for each sample size. In this experiment the method of Hong and Li (2018) can be implemented without issue. We compare the analytically-estimated derivative using the same a_n and b_n and numerically-estimated derivatives using the same choices for ϵ_n as in the previous experiment.

The reference distribution estimated analytically performs well in terms of size and power, although not as well as in the previous experiment — the empirical rejection probability is higher than expected at the boundary of the null region. The largest choice of ϵ_n appears to work best, although this leads to a conservative test for the largest sample size. The analytic test with sample size 1,000 show higher rejection rate for a very small location shift towards the interior of the null region (location n^{-1} in the table) than on the boundary of the null region. Theorem 4.5 predicts that this should disappear asymptotically.

⁹It can be verified that if $X \sim \text{Unif}[\mu, \mu + 1]$ and $Y \sim \text{Unif}[0, 1]$, then L_{X-Y} is the CDF of a $\text{Unif}[\mu, \mu + 1]$ random variable and U_{X-Y} is the CDF of a $\text{Unif}[\mu - 1, \mu]$ random variable.

Location	Analytic	Numeric			
		$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$
-2	1.000	1.000	1.000	1.000	1.000
$-n^{-1/10}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/6}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/3}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/2}$	0.645	0.479	0.819	0.984	1.000
$-n^{-2/3}$	0.266	0.179	0.410	0.721	0.961
$-n^{-1}$	0.114	0.070	0.199	0.438	0.812
Null bndry.					
0	0.100	0.059	0.193	0.407	0.778
n^{-1}	0.070	0.043	0.129	0.355	0.708
$n^{-2/3}$	0.032	0.018	0.059	0.177	0.432
$n^{-1/2}$	0.005	0.004	0.011	0.044	0.128
$n^{-1/3}$	0.000	0.000	0.000	0.000	0.003
$n^{-1/6}$	0.000	0.000	0.000	0.000	0.000
$n^{-1/10}$	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000

Table 7: Empirical rejection frequencies for a test that $L_A \leq U_B$. Analytic derivative estimates are described in the text, while Numeric derivative estimates use the method suggested by Hong and Li (2018). Each row label gives the location parameter of one of the distributions, while the null assumes both parameters are zero. Samples of size 100, 499 bootstrap repetitions per test, 1000 simulation repetitions.

Location	Analytic	Numeric			
		$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$
-2	1.000	1.000	1.000	1.000	1.000
$-n^{-1/10}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/6}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/3}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/2}$	0.668	0.526	0.781	0.983	1.000
$-n^{-2/3}$	0.172	0.099	0.245	0.570	0.943
$-n^{-1}$	0.090	0.057	0.140	0.363	0.802
Null bndry.					
0	0.085	0.049	0.146	0.358	0.777
n^{-1}	0.063	0.037	0.137	0.340	0.751
$n^{-2/3}$	0.029	0.013	0.053	0.162	0.482
$n^{-1/2}$	0.004	0.002	0.005	0.029	0.116
$n^{-1/3}$	0.000	0.000	0.000	0.000	0.000
$n^{-1/6}$	0.000	0.000	0.000	0.000	0.000
$n^{-1/10}$	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000

Table 8: Empirical rejection frequencies for a test that $L_A \leq U_B$. Analytic derivative estimates are described in the text, while Numeric derivative estimates use the method suggested by Hong and Li (2018). Each row label gives the location parameter of one of the distributions, while the null assumes both parameters are zero. Samples of size 500, 999 bootstrap repetitions per test, 1000 simulation repetitions.

Location	Analytic	Numeric			
		$\epsilon_n = n^{-1/6}$	$\epsilon_n = n^{-1/3}$	$\epsilon_n = n^{-1/2}$	$\epsilon_n = n^{-1}$
-2	1.000	1.000	1.000	1.000	1.000
$-n^{-1/10}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/6}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/3}$	1.000	1.000	1.000	1.000	1.000
$-n^{-1/2}$	0.653	0.528	0.776	0.978	1.000
$-n^{-2/3}$	0.169	0.104	0.261	0.559	0.932
$-n^{-1}$	0.084	0.054	0.135	0.355	0.793
Null bndry.					
0	0.067	0.031	0.107	0.334	0.730
n^{-1}	0.069	0.031	0.110	0.305	0.729
$n^{-2/3}$	0.033	0.017	0.054	0.167	0.518
$n^{-1/2}$	0.002	0.001	0.005	0.026	0.102
$n^{-1/3}$	0.000	0.000	0.000	0.000	0.000
$n^{-1/6}$	0.000	0.000	0.000	0.000	0.000
$n^{-1/10}$	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000

Table 9: Empirical rejection frequencies for a test that $L_A \leq U_B$. Analytic derivative estimates are described in the text, while Numeric derivative estimates use the method suggested by Hong and Li (2018). Each row label gives the location parameter of one of the distributions, while the null assumes both parameters are zero. Samples of size 1000, 1999 bootstrap repetitions per test, 1000 simulation repetitions.

B Proofs

Lemma B.1. *Suppose that $U \subseteq \mathbb{R}^{d_U}$ and $X \subseteq \mathbb{R}^{d_X}$ and $f, h \in \ell^\infty(\text{gr}A)$. Let $A : X \rightrightarrows U$ be non-empty-valued and define $\sigma : \ell^\infty(\text{gr}A) \rightarrow \mathbb{R}$ by $\sigma(f) = \sup_{x \in X} \inf_{u \in A(x)} f(u, x)$. Then σ is Hadamard directionally differentiable and*

$$\sigma'_f(h) = \lim_{\delta \rightarrow 0^+} \sup_{x \in X_f^\sigma(\delta)} \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x),$$

where U_f and X_f^σ are defined in (10) and (16).

Proof of Lemma B.1. It can be verified using that $|\sigma(f) - \sigma(g)| \leq \|f - g\|_\infty$, so we can focus on Gâteaux differentiability for a fixed f and h (Shapiro, 1990, Proposition 3.5). Fix some $f, h \in \ell^\infty(\text{gr}A)$ and $t_n \rightarrow 0^+$.

Start with an upper bound for the scaled difference $(\sigma(f + t_n h) - \sigma(f))/t_n$, which, using $s_n = t_n^{-1}$ may be rewritten as $\sigma(s_n f + h) - s_n \sigma(f)$. Consider the inner optimization problem in u . For any $x \in X$ and $\epsilon > 0$, by definition of $U_{(-f)}$ (which collects near-minimizers in u of f) there exists a $u_\epsilon \in U_{(-f)}(x, \epsilon)$ such that $h(u_\epsilon, x) \leq \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x) + \epsilon$ and $f(u_\epsilon, x) \leq \inf_{u \in A(x)} f(u, x) + \epsilon$ and therefore

$$\begin{aligned} \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x) &\geq h(u_\epsilon, x) - \epsilon \\ &= s_n f(u_\epsilon, x) + h(u_\epsilon, x) - s_n f(u_\epsilon, x) - \epsilon \\ &\geq \inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) + s_n \left(\inf_{u \in A(x)} f(u, x) - f(u_\epsilon, x) \right) - \epsilon \\ &\geq \inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) - (s_n + 1)\epsilon. \end{aligned}$$

Thus, for each n , for any $x \in X$,

$$\inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) \leq \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x). \quad (\text{S.1})$$

Next consider the outer optimization problem in x : for each $x \in X$, we may write

$$\begin{aligned} \inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \sigma(f) &= \left(\inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) \right) \\ &\quad + s_n \left(\inf_{u \in A(x)} f(u, x) - \sigma(f) \right) \\ &\leq \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x) + s_n \left(\inf_{u \in A(x)} f(u, x) - \sigma(f) \right), \end{aligned}$$

where the inequality comes from (S.1). Consider this inequality for some $x' \notin X_f^\sigma(\delta)$, for any $\delta > 0$:

$$\begin{aligned} \inf_{u \in A(x')} (s_n f + h)(u, x') - s_n \sigma(f) &\leq \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x', \epsilon)} h(u, x') + s_n \left(\inf_{u \in A(x')} f(u, x') - \sigma(f) \right) \\ &\leq \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x', \epsilon)} h(u, x') - s_n \delta. \end{aligned}$$

Recall that $s_n = t_n^{-1}$ so that s_n diverges as $n \rightarrow \infty$, and that x' that are not δ -maximinimizers cannot be optimal. Then for any $\delta > 0$,

$$\begin{aligned} \limsup_{n \rightarrow \infty} (\sigma(s_n f + h) - s_n \sigma(f)) &= \limsup_{n \rightarrow \infty} \left(\sup_{x \in X_f^\sigma(\delta)} \inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \sigma(f) \right) \\ &\leq \sup_{x \in X_f^\sigma(\delta)} \left(\lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x) \right) \end{aligned}$$

and therefore this inequality holds as $\delta \rightarrow 0$.

To obtain a lower bound, start again with the inner problem. For any $x \in X$, choose an $\epsilon > 0$ and note that for any $u' \in A(x) \setminus U_{(-f)}(x, \epsilon)$,

$$s_n f(u', x) + h(u', x) - s_n \inf_{u \in A(x)} f(u, x) \geq \inf_{u \in A(x)} h(u, x) + s_n \epsilon.$$

Therefore, again recalling that s_n diverges with n , we may restrict attention to the ϵ -minimizers (in u) for each $x \in X$:

$$\begin{aligned} \liminf_{n \rightarrow \infty} \left(\inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) \right) \\ = \liminf_{n \rightarrow \infty} \left(\inf_{u \in U_{(-f)}(x, \epsilon)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) \right) \end{aligned} \quad (\text{S.2})$$

$$\geq \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x) \quad (\text{S.3})$$

and this holds as $\epsilon \rightarrow 0^+$ as well.

Consider again the outer maximization problem. For any $\delta > 0$, define

$$\tilde{h}(\delta) = \sup_{x \in X_f^\sigma(\delta)} \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x, \epsilon)} h(u, x).$$

For each δ there is an $x_\delta \in X_f^\sigma(\delta)$ such that

$$\lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} h(u, x_\delta) \geq \tilde{h}(\delta) - \delta$$

and by construction,

$$\begin{aligned} \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} f(u, x_\delta) &\geq \inf_{u \in A(x_\delta)} f(u, x_\delta) \\ &\geq \sup_{x \in X} \inf_{u \in A(x)} f(u, x) - \delta = \sigma(f) - \delta, \end{aligned}$$

where the second inequality follows from the definition of $X_f^\sigma(\delta)$. Then for each n ,

$$\begin{aligned} \tilde{h}(\delta) &\leq \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} h(u, x_\delta) + \delta \\ &= s_n \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} f(u, x_\delta) + \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} h(u, x_\delta) - s_n \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} f(u, x_\delta) + \delta \\ &\leq \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} (s_n f + h)(u, x_\delta) - s_n \lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} f(u, x_\delta) + \delta \\ &= \left(\lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} (s_n f + h)(u, x_\delta) - s_n \sigma(f) \right) - s_n \left(\lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} f(u, x_\delta) - \sigma(f) \right) + \delta \\ &\leq \left(\lim_{\epsilon \rightarrow 0^+} \inf_{u \in U_{(-f)}(x_\delta, \epsilon)} (s_n f + h)(u, x_\delta) - s_n \sigma(f) \right) + (s_n + 1)\delta \end{aligned}$$

Then letting $\delta \rightarrow 0$ and using (S.2) and (S.3) we have

$$\liminf_{n \rightarrow \infty} (\sigma(s_n f + h) - s_n \sigma(f)) \geq \lim_{\delta \rightarrow 0^+} \tilde{h}(\delta).$$

□

Proof of Theorem 3.2. The function $\gamma : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by $\gamma(x, y) = \max\{x, y\}$ has Hadamard directional derivative

$$\gamma'_{x,y}(h, k) = \begin{cases} h & x > y \\ \max\{h, k\} & x = y \\ k & x < y \end{cases}.$$

Use the equivalence $\sup |f| = \max\{\sup f, \sup(-f)\}$ to rewrite the difference for any n as

$$\begin{aligned} \frac{1}{t_n} \left(\sup_{x \in X} \left| \sup_{u \in A(x)} (f(u, x) + t_n h_n(u, x)) \right| - \sup_{x \in X} \left| \sup_{u \in A(x)} f(u, x) \right| \right) &= \\ \frac{1}{t_n} \left(\max \left\{ \sup_{(u,x) \in \text{gr} A} (f(u, x) + t_n h_n(u, x)), \sup_{x \in X} \left(- \sup_{u \in A(x)} (f(u, x) + t_n h_n(u, x)) \right) \right\} - \right. \\ &\quad \left. - \max \left\{ \sup_{(u,x) \in \text{gr} A} f(u, x), \sup_{x \in X} \left(- \sup_{u \in A(x)} f(u, x) \right) \right\} \right) \\ &= \frac{1}{t_n} (\max\{\mu(f + t_n h_n), \sigma(-f - t_n h_n)\} - \max\{\mu(f), \sigma(-f)\}). \end{aligned}$$

Defining $\tilde{\sigma}(f) = \sigma(-f)$, Lemma B.1 implies that $\tilde{\sigma}'_f(h) = \sigma'_{(-f)}(-h)$. Then the chain rule implies

$$(\gamma(\mu(f), \tilde{\sigma}(f)))'(h) = \gamma'_{(\mu(f), \tilde{\sigma}(f))}(\mu'_f(h), \tilde{\sigma}'_f(h)),$$

and writing the derivatives out is the first result. The condition $\|\psi(f)\|_\infty = 0$ is equivalent to $\mu(f) = \sigma(-f) = 0$. Specializing the general derivative so that $\psi(f) \equiv 0$ implies $A_f^\mu(\epsilon) = U_f(X, \epsilon)$. Lemma B.2, the continuity of $\max\{x, y\}$ and the identity $|x| = \max\{-x, x\}$ imply the second result.

Next consider the functional λ_2 . Write $\sup_x \max\{\sup_u f(u, x), 0\} = \max\{\sup_{(u,x)} f(u, x), 0\}$ and apply the chain rule with $\gamma(x, 0)$. Then the general result for $\lambda_2(f)$ results from reversing the order of the maximum and supremum again. When $\|[\psi(f)]_+\|_\infty = 0$, the derivative is only nonzero if $\sup_{\text{gr}A} f = 0$ so that $A_f^\mu(\epsilon) = U_f(X_0, \epsilon)$, and similar calculations lead to the final result. \square

Lemma B.2. *Let $A_f^\mu(\epsilon)$ and $U_f(\cdot, \epsilon)$ be defined by (15) and (10) in the text. Let*

$$X_f(\epsilon) = \left\{ x \in X : \sup_{u \in A(x)} f(u, x) \geq \sup_{(u,x) \in \text{gr}A} f(u, x) - \epsilon \right\}.$$

Then

$$\lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in A_f^\mu(\epsilon)} h(u, x) = \lim_{\epsilon \rightarrow 0^+} \lim_{\delta \rightarrow 0^+} \sup_{x \in X_f(\epsilon)} \sup_{u \in U_f(x, \delta)} h(u, x).$$

Proof of Lemma B.2. Note that for any $\epsilon > 0$, $A_f^\mu(\epsilon) = \cup_{x \in X_f(\epsilon)} U_f(x, \epsilon)$. Define

$$\mathcal{A} = \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in A_f^\mu(\epsilon)} h(u, x), \quad \mathcal{B} = \lim_{\epsilon \rightarrow 0^+} \lim_{\delta \rightarrow 0^+} \sup_{x \in X_f(\epsilon)} \sup_{u \in U_f(x, \delta)} h(u, x).$$

As families of sets indexed by ϵ , $A_f^\mu(\epsilon)$, $X_f(\epsilon)$ and $U_f(x, \epsilon)$ (for any x) are nonincreasing in diameter as $\epsilon \rightarrow 0$. Therefore suprema evaluated over any of these sets are nonincreasing as $\epsilon \rightarrow 0$.

For any $\epsilon, \delta > 0$, $\mathcal{A} \leq \sup_{x \in X_f(\epsilon)} \sup_{u \in U_f(x, \delta)} h(u, x)$, and by letting $\epsilon, \delta \rightarrow 0$ we have $\mathcal{A} \leq \mathcal{B}$. Since h is bounded, $\mathcal{A} < \infty$ and for all $\eta > 0$ there exist $\epsilon^*, \delta^* > 0$ such that

$$\sup_{x \in X_f(\epsilon^*)} \sup_{u \in U_f(x, \delta^*)} h(u, f) \leq \mathcal{A} + \eta.$$

Then

$$\mathcal{B} \leq \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X_f(\epsilon)} \sup_{u \in U_f(x, \delta^*)} h(u, f) \leq \sup_{x \in X_f(\epsilon^*)} \sup_{u \in U_f(x, \delta^*)} h(u, f) \leq \mathcal{A} + \eta.$$

By letting $\eta \rightarrow 0$ we find $\mathcal{B} \leq \mathcal{A}$. \square

Proof of Theorem 3.3. The result of this theorem follows from several applications of the chain rule for the function evaluated at (almost) every point in X , along with dominated convergence to

move from pointwise convergence to convergence in the L_p norm. The maps and their Hadamard directional derivatives are discussed first and then composed.

First, we note that the condition $f \in \ell^\infty(\text{gr}A)$ and $m(X) < \infty$ imply $\|\psi(f)\|_p < \infty$: since $\|\psi(f)\|_p^p \leq \|f\|_\infty^p m(X)$, the L_p norm of $\psi(f)$ is well-defined.

Let f and h satisfy the assumptions in the statement of the theorem and choose $t_n \rightarrow 0^+$ and h_n such that $\|h_n - h\|_\infty \rightarrow 0$. It can be verified from the definition that:

1. For $x \in \mathbb{R}$ and $a \in \mathbb{R}$, the derivative of $x \mapsto |x|$ in direction a is $\text{sgn}(x) \cdot a$ when $x \neq 0$ and $|a|$ when $x = 0$.
2. For $x \in \mathbb{R}$ and $b \in \mathbb{R}$, the derivative of $x \mapsto [x]_+$ in direction b is b when $x > 0$, $[b]_+$ when $x = 0$ and 0 when $x < 0$.
3. For $x \geq 0$, $\alpha > 0$ and $c \in \mathbb{R}$, the derivative of $x \mapsto x^\alpha$ in direction c is $\alpha c x^{\alpha-1}$.

Suppose that $\|\psi(f)\|_p \neq 0$. For almost all $x \in X$ the chain rule implies that the derivative of $f(\cdot, x) \mapsto |\psi(f)(x)|^p$ in direction $h(\cdot, x)$ is

$$\begin{cases} p \text{sgn}(\psi(f)(x)) |\psi(f)(x)|^{p-1} \psi'_f(h)(x), & \psi(f)(x) \neq 0 \\ 0, & \psi(f)(x) = 0 \end{cases} = p \text{sgn}(\psi(f)(x)) |\psi(f)(x)|^{p-1} \psi'_f(h)(x).$$

Similarly, for almost all $x \in X$ the derivative of $f(\cdot, x) \mapsto [\psi(f)(x)]_+^p$ in direction $h(\cdot, x)$ is

$$p [\psi(f)(x)]_+^{p-1} \times \begin{cases} \psi'_f(h)(x), & \psi(f)(x) > 0 \\ [\psi'_f(h)(x)]_+, & \psi(f)(x) = 0 \\ 0, & \psi(f)(x) < 0 \end{cases} = p [\psi(f)(x)]_+^{p-1} \psi'_f(h)(x).$$

Then the assumed p -integrability of these functions, implied by Assumption **A1** and Proposition 6.10 of Folland (1999), and dominated convergence imply that when $\|\psi(f)\|_p \neq 0$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{t_n} \left(\int_X |\psi(f + t_n h_n)(x)|^p dm(x) - \int_X |\psi(f)(x)|^p dm(x) \right) \\ = \int_X p \text{sgn}(\psi(f)(x)) |\psi(f)(x)|^{p-1} \psi'_f(h)(x) dm(x) \end{aligned}$$

and similarly, when $\|[\psi(f)]_+\|_p \neq 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{t_n} \left(\int_X [\psi(f + t_n h_n)(x)]_+^p dm(x) - \int_X [\psi(f)(x)]_+^p dm(x) \right) = \int_X p [\psi(f)(x)]_+^{p-1} \psi'_f(h)(x) dm(x).$$

When considering the $1/p$ -th power of these integrals, one more application of the chain rule and

the third basic derivative in the above list imply

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\lambda_3(f + t_n h_n) - \lambda_3(f)}{t_n} \\ = \left(\int_X |\psi(f)(x)|^p dm(x) \right)^{(1-p)/p} \int_X \operatorname{sgn}(\psi(f)(x)) |\psi(f)(x)|^{p-1} \psi'_f(h)(x) dm(x) \end{aligned}$$

and

$$\lim_{n \rightarrow \infty} \frac{\lambda_4(f + t_n h_n) - \lambda_4(f)}{t_n} = \left(\int_X [\psi(f)(x)]_+^p dm(x) \right)^{(1-p)/p} \int_X [\psi(f)(x)]_+^{p-1} \psi'_f(h)(x) dm(x).$$

On the other hand, when $\|\psi(f)\|_p = 0$, one can calculate directly that

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{t_n} \left(\left(\int_X |\psi(f + t_n h_n)(x)|^p dm(x) \right)^{1/p} - \left(\int_X |\psi(f)(x)|^p dm(x) \right)^{1/p} \right) \\ = \left(\int_X \left| \lim_{n \rightarrow \infty} \frac{\psi(f + t_n h_n)(x) - 0}{t_n} \right|^p dm(x) \right)^{1/p} \\ = \left(\int_X \left| \lim_{\epsilon \rightarrow 0^+} \sup_{\{u: f(u, x) \geq -\epsilon\}} h(u, x) \right|^p dm(x) \right)^{1/p}. \end{aligned}$$

The case is slightly different for λ_4 because $\|[\psi(f)]_+\|_p = 0$ only implies $\psi(f)(x) \leq 0$ for almost all $x \in X$. Rewrite the difference as in the above display. Rule 2 in the above list of derivative rules implies that only X_0 will contribute asymptotically to the inner integral, and in the limit, using p -integrability and dominated convergence we have the final statement. \square

Lemma B.3. *Let λ be the map defined in (20). Then λ is Hadamard directionally differentiable and under the hypothesis that $\max\{\Pi^+(F)(x), 0\} = L_0(x)$ for all $x \in \mathbb{R}$, its derivative is defined in (21).*

Proof of Lemma B.3. To save space, we have used the notation $\sup_A f = \sup_{x \in A} f(x)$ and $a \vee b = \max\{a, b\}$ in this proof. The map Π^+ has a Hadamard derivative at any (u, x) equal to $h_X(u) + h_Y(x - u)$. Therefore we calculate, for $f \in \ell^\infty(\mathbb{R}^2)$, in the direction of some $h \in \ell^\infty(\mathbb{R}^2)$, the directional derivative

$$\varphi'_f(h) := \lim_{t \rightarrow 0^+} \frac{1}{t} \left(\sup_X |\psi(f + th) \vee 0 - L_0| - \sup_X |\psi(f) \vee 0 - L_0| \right). \quad (\text{S.4})$$

If $\psi(f)(x) < 0$ then for small enough t , $\max\{\psi(f + th)(x), 0\} = 0$. Therefore we may focus on $\{x \in \mathbb{R} : \psi(f)(x) \geq 0\}$. Break the domain into the two subsets $X_0 = \{x \in \mathbb{R} : \psi(f)(x) = 0\}$ and

$X_+ = \{x \in \mathbb{R} : \psi(f)(x) > 0\}$. Write

$$\sup_X |\psi(f + th) \vee 0 - L_0| = \max \left\{ \sup_{X_0} |\psi(f + th) \vee 0 - L_0|, \sup_{X_+} |\psi(f + th) \vee 0 - L_0| \right\},$$

and analogously for the second term in the difference (S.4). This indicates that the easiest way to find the derivative is via the chain rule. We take each one of these two suprema in turn.

Under the null hypothesis that $\psi(f) \equiv L_0$ and on X_0 , where both are zero,

$$\begin{aligned} \lim_{t \rightarrow 0^+} t^{-1} \left(\sup_{X_0} |\psi(f + th) \vee 0 - L_0| - \sup_{X_0} |\psi(f) \vee 0 - L_0| \right) &= \lim_{t \rightarrow 0^+} t^{-1} \sup_{X_0} \psi(f + th) \vee 0 \\ &= \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X_0} \sup_{u \in U_f(x, \epsilon)} [h(u, x)]_+. \end{aligned} \quad (\text{S.5})$$

Meanwhile, on X_+ ,

$$\begin{aligned} \lim_{t \rightarrow 0^+} t^{-1} \left(\sup_{X_+} |\psi(f + th) \vee 0 - L_0| - \sup_{X_+} |\psi(f) \vee 0 - L_0| \right) \\ = \lim_{t \rightarrow 0^+} t^{-1} \left(\sup_{X_+} |\psi(f + th) - L_0| - \sup_{X_+} |\psi(f) - L_0| \right). \end{aligned}$$

Apply Theorem 3.2 using the function $\tilde{f}(u, x) = f(u, x) - L_0(x)$. Note that $U_{\tilde{f}}(x, \epsilon) = U_f(x, \epsilon)$ for all $x \in X_+$. Since the second term in the above expression is zero under the null hypothesis that $\psi(f) \equiv L_0$, Theorem 3.2 implies

$$\lim_{t \rightarrow 0^+} t^{-1} \left(\sup_{X_+} |\psi(f + th) - L_0| - \sup_{X_+} |\psi(f) - L_0| \right) = \lim_{\epsilon \rightarrow 0^+} \sup_{x \in X_+} \left| \sup_{u \in U_f(x, \epsilon)} h(u, x) \right|. \quad (\text{S.6})$$

Let

$$\mu_0(f) = \sup_{X_0} |\psi(f) \vee 0 - L_0|, \quad \mu_+(f) = \sup_{X_+} |\psi(f) \vee 0 - L_0|.$$

Then define the derivatives μ'_{0f} and μ'_{+f} by (S.5) and (S.6). Now, via the chain rule, under the hypothesis that L_0 has been correctly chosen, $\mu_0(f) = \mu_+(f) = 0$. Then using the derivative of the map γ defined in the proof of Theorem 3.2, we have a solution to (S.4):

$$\varphi'_f(h) = \max \{ \mu'_{0f}(h), \mu'_{+f}(h) \}. \quad (\text{S.7})$$

To translate this back to a result on the λ defined in (20), use the definitions that precede (21) in (S.7) and let $h(u, x) = h_X(u) + h_Y(x - u)$. \square

Proof of Theorem 4.1. The limit statements follow directly from Theorem 3.2, Theorem 3.3 above

combined with Theorem 2.1 of Fang and Santos (2019). \square

Proof of Theorem 4.3. We will apply Theorem 3.2 of Fang and Santos (2019) to show the result by verifying that the maps involved satisfy the appropriate regularity conditions. First, Theorems 3.2 and 3.3 show that the λ_j are Hadamard directionally differentiable maps, implying their Assumption 1 is satisfied. Our assumptions **A2** and **A3** imply their Assumptions 2 and 3.

Next, we show that each map is Lipschitz and consistent for each fixed direction h , so that we may apply Lemma S.3.6 of Fang and Santos (2019), which implies their Assumption 4 is satisfied. Start with λ_1 . Fix $f \in \ell^\infty(\text{gr}A)$ and suppose that $h, k \in \ell^\infty(\text{gr}A)$. For any n ,

$$\left| \hat{\lambda}'_{1n}(h) - \hat{\lambda}'_{1n}(k) \right| \leq \sup_{(u,x) \in U_{f_n}(X, a_n)} |h(u, x) - k(u, x)| \leq \|h - k\|_\infty.$$

Next we show that for each h , $|\hat{\lambda}'_{1n}(h) - \lambda'_{1f}(h)| = o_p(1)$. For any $\eta > 0$,

$$P \left\{ \left| \hat{\lambda}'_{1n}(h) - \lambda'_{1f}(h) \right| > \eta \right\} \leq P \left\{ \left| \sup_{(u,x) \in U_{f_n}(X, a_n)} h(u, x) - \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X, \epsilon)} h(u, x) \right| > \eta \right\}.$$

If $(u_n, x_n) \in U_{f_n}(X, a_n)$, then $f_n(u_n, x_n) \geq \sup_{(u,x) \in \text{gr}A} f_n(u, x) - a_n$. This is equivalent to

$$\begin{aligned} f(u_n, x_n) + (f_n(u_n, x_n) - f(u_n, x_n)) &\geq \sup_{(u,x) \in \text{gr}A} f(u, x) + \left(\sup_{(u,x) \in \text{gr}A} f_n(u, x) - \sup_{(u,x) \in \text{gr}A} f(u, x) \right) - a_n \\ &\Rightarrow f(u_n, x_n) \geq \sup_{(u,x) \in \text{gr}A} f(u, x) - a_n - 2\|f_n - f\|_\infty. \end{aligned}$$

That is, $U_{f_n}(X, a_n) \subseteq U_f(X, a_n + 2\|f_n - f\|_\infty)$. Therefore as $n \rightarrow \infty$,

$$\begin{aligned} P \left\{ \sup_{(u,x) \in U_{f_n}(X, a_n)} h(u, x) > \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X, \epsilon)} h(u, x) + \eta \right\} \\ \leq P \left\{ \sup_{(u,x) \in U_f(X, a_n + 2\|f_n - f\|_\infty)} h(u, x) > \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X, \epsilon)} h(u, x) + \eta \right\} \rightarrow 0 \end{aligned}$$

It can similarly be shown that $U_f(X, a_n) \subseteq U_{f_n}(X, a_n + 2\|f_n - f\|_\infty)$. This implies that

$$\lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X, \epsilon)} h(u, x) \leq \sup_{(u,x) \in U_{f_n}(X, a_n)} h(u, x) \leq \sup_{(u,x) \in U_f(X, a_n + 2\|f_n - f\|_\infty)} h(u, x).$$

Then as $n \rightarrow \infty$,

$$\begin{aligned} P \left\{ \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X,\epsilon)} h(u,x) > \sup_{(u,x) \in U_{f_n}(X,a_n)} h(u,x) + \eta \right\} \\ \leq P \left\{ \sup_{(u,x) \in U_{f_n}(X,a_n+2\|f_n-f\|_\infty)} h(u,x) > \sup_{(u,x) \in U_{f_n}(X,a_n)} h(u,x) + \eta \right\} \rightarrow 0. \end{aligned}$$

Next consider λ_2 . Using the fact that $x \mapsto [x]_+$ is Lipschitz with constant 1,

$$\left| \hat{\lambda}'_{2n}(h) - \hat{\lambda}'_{2n}(k) \right| \leq \sup_{(u,x) \in U_{f_n}(X,a_n)} |[h(u,x)]_+ - [k(u,x)]_+| \leq \|h - k\|_\infty.$$

In order to show consistency for fixed h , recall that under the hypothesis that $P \in \mathcal{P}_{00}^I$ there is only one form for λ'_{2f} (that is, the trivial case where it is equal to zero is ruled out by assumption). Note that for any $\eta > 0$

$$\begin{aligned} P \left\{ \left| \hat{\lambda}'_{2n}(h) - \lambda'_{2f}(h) \right| > \eta \right\} &\leq P \left\{ \left| \sup_{(u,x) \in U_{f_n}(\hat{X}_0,a_n)} h(u,x) - \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X_0,\epsilon)} h(u,x) \right| > \eta \right\} \\ &\leq P \left\{ \left| \sup_{(u,x) \in U_{f_n}(X_0 \cap \hat{X}_0,a_n)} h(u,x) - \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X_0 \cap \hat{X}_0,\epsilon)} h(u,x) \right| > \eta \right\} + P \left\{ \hat{X}_0 \triangle X_0 \neq \emptyset \right\}. \end{aligned}$$

On $X_0 \cap \hat{X}_0$, consistency is guaranteed by the argument used above for $\hat{\lambda}'_{1n}$. Consider the probability that $X_0 \triangle \hat{X}_0$ is non-empty. Theorem 4.1 applied to λ_1 implies that $\sup_{x \in X} r_n |\psi(f_n)(x) - \psi(f)(x)|$ is asymptotically tight. Therefore if b_n satisfies the condition in Assumption **A2**, for all $x \in X$, $|\psi(f_n)(x') - \psi(f)(x')| = O_p(r_n) = o_p(b_n)$. First, if $x \in X_0$ then $\psi(f)(x) = 0$ and

$$P \left\{ x \in X_0 \cap \hat{X}_0^c \right\} = P \left\{ |\psi(f_n)(x)| > b_n \right\} = P \left\{ |\psi(f_n)(x) - \psi(f)(x)| > b_n \right\} \rightarrow 0.$$

On the other hand, if $x' \notin X_0$ then $\psi(f)(x') \neq 0$ and $P \left\{ |\psi(f)(x')| \leq b_n \right\} \rightarrow 0$. Then

$$P \left\{ x \in X_0^c \cap \hat{X}_0 \right\} = P \left\{ |\psi(f_n)(x')| \leq b_n \right\} \leq P \left\{ |\psi(f_n)(x') - \psi(f)(x')| + |\psi(f)(x')| \leq b_n \right\} \rightarrow 0.$$

This implies that for all $\epsilon > 0$, $P \left\{ d(\hat{X}_0, X_0) > \epsilon \right\} \rightarrow 0$.

For the case of λ_3 , suppose in addition that $m(X) < \infty$. Then

$$\begin{aligned} |\hat{\lambda}'_{3n}(h) - \hat{\lambda}'_{3n}(k)| &\leq \left(\int_X \left| \sup_{u \in U_{f_n}(x, a_n)} h(u, x) - \sup_{u \in U_{f_n}(x, a_n)} k(u, x) \right|^p dm(x) \right)^{1/p} \\ &\leq \left(\int_X \left| \sup_{u \in A(x)} |h(u, x) - k(u, x)| \right|^p dm(x) \right)^{1/p} \\ &\leq m(X)^{1/p} \|h - k\|_\infty. \end{aligned}$$

The above inequalities follow from Minkowski's inequality (which implies a reverse triangle inequality) and a bound on the L_p norm assuming $m(X)$ is finite. Similarly, for any fixed $\eta > 0$, defining

$$\begin{aligned} X_1 &= \left\{ x \in X : \sup_{u \in U_{f_n}(x, a_n)} h(u, x) > \lim_{\epsilon \rightarrow 0^+} \sup_{u \in U_f(x, \epsilon)} h(u, x) \right\}, \\ X_2 &= \left\{ x \in X : \sup_{u \in U_{f_n}(x, a_n)} h(u, x) < \lim_{\epsilon \rightarrow 0^+} \sup_{u \in U_f(x, \epsilon)} h(u, x) \right\}, \end{aligned}$$

we may write

$$\begin{aligned} &P \left\{ |\hat{\lambda}'_{3n}(h) - \lambda'_{3f}(h)| > \eta \right\} \\ &\leq P \left\{ \left(\int_{X_1} \left(\sup_{u \in U_{f_n}(x, a_n)} h(u, x) - \lim_{\epsilon \rightarrow 0^+} \sup_{u \in U_f(x, \epsilon)} h(u, x) \right)^p dm(x) \right)^{1/p} > \eta \right\} \\ &\quad + P \left\{ \left(\int_{X_2} \left(\lim_{\epsilon \rightarrow 0^+} \sup_{u \in U_f(x, \epsilon)} h(u, x) - \sup_{u \in U_{f_n}(x, a_n)} h(u, x) \right)^p dm(x) \right)^{1/p} > \eta \right\} \end{aligned}$$

Using an analogous argument as for the λ_1 case, this sum is bounded by

$$\begin{aligned} &P \left\{ \left(\int_{X_1} \left(\sup_{u \in U_f(x, a_n + 2\|f_n - f\|_\infty)} h(u, x) - \lim_{\epsilon \rightarrow 0^+} \sup_{u \in U_f(x, \epsilon)} h(u, x) \right)^p dm(x) \right)^{1/p} > \eta \right\} \\ &\quad + P \left\{ \left(\int_{X_2} \left(\sup_{u \in U_{f_n}(x, a_n + 2\|f_n - f\|_\infty)} h(u, x) - \sup_{u \in U_{f_n}(x, a_n)} h(u, x) \right)^p dm(x) \right)^{1/p} > \eta \right\} \rightarrow 0 \end{aligned}$$

by monotone convergence and the dominated convergence theorem.

Finally, because $x \mapsto [x]_+$ is Lipschitz we find

$$|\hat{\lambda}'_{4n}(h) - \hat{\lambda}'_{4n}(k)| \leq \left(\int_X \left| \sup_{u \in A(x)} |h(u, x) - k(u, x)| \right|^p dm(x) \right)^{1/p} \leq m(X)^{1/p} \|h - k\|_\infty.$$

Consistency for fixed h follows from the argument for the λ_3 case. Specifically, subdivide the domain of integration further into $X_1 \cap \hat{X}_0 \cap X_0$, $X_2 \cap \hat{X}_0 \cap X_0$ and $\hat{X}_0 \triangle X_0$. The first two subsets can be dealt with as in the λ_3 case, and the last can be dealt with as in the argument for the λ_2 case.

Bootstrap consistency for λ_j for all j then follows from Theorem 3.2 and Lemma S.3.6 of Fang and Santos (2019). \square

Proof of Theorem 4.4. For $j = 1$ or $j = 3$, under the assumption that $P \in \mathcal{P}_{00}^E$, there are no interior distributions and so for any sequence $P_n \rightarrow P$ such that $P_n \in \mathcal{P}_{00}^E$, $\lambda_j(f(P_n)) = 0$ for all n . Therefore $\lambda'_{jf}(f'(c)) = \lim_{n \rightarrow \infty} r_n(\lambda_j(f(P_n)) - \lambda_j(f(P_0))) = 0$. Now consider λ_j for $j \in \{1, 3\}$. Theorem 3.3 of Fang and Santos (2019) implies that, using $Z_n \xrightarrow{L_c} Z$ to denote weak convergence under $\bigotimes_{i=1}^n P_{c/r_n}$,

$$r_n(\lambda_j(f_n) - \lambda_j(f(P_n))) \xrightarrow{L_c} \lambda'_{jf}(\mathcal{G}_{P_0} + f'(c)) - \lambda'_{jf}(f'(c)) \sim \lambda'_{jf}(\mathcal{G}_{P_0} + f'(c)),$$

The result for the other two statistics follows from Corollary 3.2 of Fang and Santos (2019) if we show the convexity of the derivatives $\lambda'_{jf}(h)$ when $\lambda_j(f) = 0$ for $j = 2$ or $j = 4$. The case that $P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$ is trivially convex, so we restrict our attention to $P \in \mathcal{P}_{00}^I$. It can be verified that for $a, b \in \mathbb{R}$, $[a + b]_+ \leq [a]_+ + [b]_+$. Then for $h, k \in \ell^\infty(U \times X)$,

$$\lambda'_{2f}(\alpha h + (1 - \alpha)k) \leq \left[\alpha \lim_{\epsilon \rightarrow 0^+} \sup_{U_f(X, \epsilon)} h + (1 - \alpha) \lim_{\epsilon \rightarrow 0^+} \sup_{U_f(X, \epsilon)} k \right]_+ \leq \alpha \lambda'_{2f}(h) + (1 - \alpha) \lambda'_{2f}(k).$$

Similarly,

$$\begin{aligned} \lambda'_{4f}(\alpha h + (1 - \alpha)k) &\leq \left(\int_{X_0} |\alpha [\psi'_f(h)(x)]_+ + (1 - \alpha) [\psi'_f(k)(x)]_+|^p dm(x) \right)^{1/p} \\ &\leq \alpha \lambda'_{4f}(h) + (1 - \alpha) \lambda'_{4f}(k). \end{aligned}$$

\square

The next lemma confirms that under assumptions that ensure convergence of the data occurs uniformly over $P \in \mathcal{P}$, the results of Theorems 4.1 and 4.3 also hold uniformly over $P \in \mathcal{P}$. The second result is limited to \mathcal{P}_{00}^I because the estimates $\hat{\lambda}_{jn}$ are constructed under the assumption that $P \in \mathcal{P}_{00}^I$.

Lemma B.4. *Under Assumptions A1-A3 and A5-A8, for $j \in \{2, 4\}$,*

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{g \in BL_1(\mathbb{R})} |\mathbb{E}[g(r_n(\lambda_j(f_n) - \lambda_j(f)))] - \mathbb{E}[g(\lambda'_{jf}(\mathcal{G}_P))]| = 0$$

and for any $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \sup_{g \in BL_1(\mathbb{R})} \left| \mathbb{E} \left[g \left(\hat{\lambda}'_{jn}(r_n(f_n^* - f_n)) \right) \mid \{Z_i\}_{i=1}^n \right] - \mathbb{E} \left[g(\lambda'_{jf}(\mathcal{G}_P)) \right] \right| > \epsilon \right\} = 0.$$

Proof of Lemma B.4. First note that under Assumptions **A5** and **A6** (along with Theorems 3.2 and 3.3), the convergence-in-distribution result of Theorem 2.1 of Fang and Santos (2019) may be rewritten to hold uniformly over $P \in \mathcal{P}$. Specifically, use the uniform continuous mapping theorem developed in Linton, Song, and Whang (2010) in the proof of Theorem 2.1 of Fang and Santos (2019) (in the place of their appeal to Theorem 1.11.1 of van der Vaart and Wellner (1996)) and note that under the null hypothesis, $\lambda_j(f_n) = \lambda_j(f_n) - \lambda_j(f)$.

The second result is similar to Theorem 3.2 of Fang and Santos (2019) but holds uniformly over $P \in \mathcal{P}_{00}^I$ for the specific statistics considered in the lemma. For convenience let $\mathcal{G}_n^* = r_n(f_n^* - f_n)$ and $Z = \{Z_i\}_{i=1}^n$.

Under Assumptions **A3**, **A6** and **A8**, Lemma S.3.1 of Fang and Santos (2019) can be extended to hold uniformly over $P \in \mathcal{P}$ to show that uniformly in $P \in \mathcal{P}$, $\mathcal{G}_n^* \rightsquigarrow \mathcal{G}_P$. This implies that for any closed set $F \in \mathcal{C}(\text{gr}A)$,

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \{ \mathcal{G}_n^* \in F \} \leq \sup_{P \in \mathcal{P}} P \{ \mathcal{G}_P \in F \}. \quad (\text{S.8})$$

This can be seen by writing

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \{ \mathcal{G}_n^* \in F \} \leq \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} (P \{ \mathcal{G}_n^* \in F \} - P \{ \mathcal{G}_P \in F \}) + \sup_{P \in \mathcal{P}} P \{ \mathcal{G}_P \in F \}.$$

If the first term on the right-hand side were strictly positive, then as $n \rightarrow \infty$, for some $\epsilon > 0$ it occurs infinitely often that there is a $P \in \mathcal{P}$ such that $P \{ \mathcal{G}_n^* \in F \} - P \{ \mathcal{G}_P \in F \} \geq \epsilon$. Construct a sequence of such measures $\{P_k\}_{k=1}^\infty$. But then it is possible to find functions that are multiples of functions in $BL_1(\text{gr}A)$ that violate the weak convergence assumption: use functions analogous (the functions should dominate indicator functions here) to those in Addendum 1.12.3 and Theorem 1.12.2 of van der Vaart and Wellner (1996). This contradiction implies that the first term on the right-hand side must be nonpositive, implying (S.8).

Next we show that for fixed h , for every $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \left\| \hat{\lambda}'_{jn}(h) - \lambda'_{jf}(h) \right\|_\infty > \epsilon \right\} = 0. \quad (\text{S.9})$$

Consider λ'_{2f} . As in the proof of Theorem 4.3, for any $\eta > 0$,

$$\sup_{P \in \mathcal{P}_{00}^I} P \left\{ \left| \hat{\lambda}'_{2n}(h) - \lambda'_{2f}(h) \right| > \eta \right\} \leq \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \left| \sup_{(u,x) \in U_{f_n}(X, a_n)} h(u, x) - \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X, \epsilon)} h(u, x) \right| > \eta \right\}.$$

As shown in the proof of that theorem, $U_{f_n}(X, a_n) \subseteq U_f(X, a_n + 2\|f_n - f\|_\infty)$ and $U_f(X, a_n) \subseteq U_{f_n}(X, a_n + 2\|f_n - f\|_\infty)$. Under Assumptions **A5** and **A6**, the other arguments of that theorem and Lemma B.2 imply that

$$\sup_{P \in \mathcal{P}_{00}^I} P \left\{ \left| \sup_{(u,x) \in U_{f_n}(X, a_n)} h(u, x) - \lim_{\epsilon \rightarrow 0^+} \sup_{(u,x) \in U_f(X, \epsilon)} h(u, x) \right| > \eta \right\} \rightarrow 0,$$

implying (S.9) holds for λ'_{2f} . Arguments like those used in Theorem 4.3 for λ'_{4f} hold uniformly over \mathcal{P}_{00}^I as well under Assumption **A5**, implying (S.9) holds for λ'_{4f} as well (note that under Assumption **A5**, the contact set consistency argument made at the end of the proof of Theorem 4.3 holds uniformly over $P \in \mathcal{P}_{00}^I$ as well). In the proof of Theorem 4.3 it is shown that $\hat{\lambda}'_{jn}$ are Lipschitz with bounded Lipschitz constants. Lemma B.5 shows that this and (S.9) imply that for any compact set $K \subset \ell^\infty(\text{gr}A)$, letting K^δ be its δ -enlargement for any $\delta > 0$,

$$\lim_{\delta \rightarrow 0^+} \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \sup_{h \in K^\delta} \|\hat{\lambda}'_{jn}(h) - \lambda'_{jf}(h)\|_\infty > \epsilon \right\} = 0. \quad (\text{S.10})$$

Now fix arbitrary $\epsilon > 0$ and $\eta > 0$. By Assumption **A6**, there is a compact set $K_0 \in \mathcal{C}(\text{gr}A)$ such that $\sup_{P \in \mathcal{P}_{00}^I} P \{ \mathcal{G}_P \notin K_0 \} < \epsilon\eta/2$. Equation (S.8) implies that for any $\delta > 0$,

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \mathcal{G}_n^* \notin K_0^\delta \right\} \leq \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \mathcal{G}_P \notin K_0^\delta \right\} \leq \sup_{P \in \mathcal{P}_{00}^I} P \{ \mathcal{G}_P \notin K_0 \} \leq \epsilon\eta/2. \quad (\text{S.11})$$

Then for each $P \in \mathcal{P}_{00}^I$ and any $\delta_0 > 0$,

$$\begin{aligned} & \sup_{g \in BL_1(\mathbb{R})} \left| \mathbb{E} \left[g(\hat{\lambda}'_{jn}(\mathcal{G}_n^*)) | Z \right] - \mathbb{E} \left[g(\lambda'_{jf}(\mathcal{G}_n^*)) | Z \right] \right| \\ & \leq \sup_{g \in BL_1(\mathbb{R})} \mathbb{E} \left[\left| g(\hat{\lambda}'_{jn}(\mathcal{G}_n^*)) - g(\lambda'_{jf}(\mathcal{G}_n^*)) \right| | Z \right] \\ & \leq \mathbb{E} \left[2I(\mathcal{G}_n^* \notin K_0^{\delta_0}) + \sup_{h \in K_0^{\delta_0}} \|\hat{\lambda}'_{jn}(h) - \lambda'_{jf}(h)\| | Z \right] \\ & \leq 2P \left\{ \mathcal{G}_n^* \notin K_0^{\delta_0} | Z \right\} + \sup_{h \in K_0^{\delta_0}} \|\hat{\lambda}'_{jn}(h) - \lambda'_{jf}(h)\|, \end{aligned} \quad (\text{S.12})$$

where the last term is not conditional on the data Z because $\hat{\lambda}'_{jn}$ only depends on Z .

Markov's inequality and Fubini's theorem (van der Vaart and Wellner, 1996, Lemma 1.2.7) imply that

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} 2P \left\{ P \left\{ \mathcal{G}_n^* \notin K_0^{\delta_0} | Z \right\} > \epsilon \right\} \leq \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} 2\epsilon^{-1} P \left\{ \mathcal{G}_n^* \notin K_0^{\delta_0} \right\} < \eta. \quad (\text{S.13})$$

Furthermore, (S.10) implies that δ_0 can be chosen such that

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \sup_{h \in K^{\delta_0}} \|\hat{\lambda}'_{jn}(h) - \lambda'_{jf}(h)\|_{\infty} > \epsilon \right\} < \eta. \quad (\text{S.14})$$

Finally, using Assumptions **A3**, **A7**, **A8** and the Lipschitz continuity of λ'_{jf} , which can be shown as in the proof of Theorem 4.3, Lemma A.1 of Linton, Song, and Whang (2010) implies

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left\{ \sup_{g \in BL_1(\mathbb{R})} \left| \mathbb{E} [g(\lambda'_{jf}(\mathcal{G}_n^*)) | Z] - \mathbb{E} [g(\lambda'_{jf}(\mathcal{G}_P))] \right| > \epsilon \right\} = 0. \quad (\text{S.15})$$

By combining (S.12), (S.14), (S.13) and (S.15), we have

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \sup_{g \in BL_1(\mathbb{R})} \left| \mathbb{E} [g(\hat{\lambda}'_{jn}(\mathcal{G}_n^*)) | Z] - \mathbb{E} [g(\lambda'_{jf}(\mathcal{G}_P))] \right| > 3\epsilon \right\} < 3\eta.$$

Since the constants are arbitrary we have the second result. \square

Lemma B.5. (Fang and Santos (2019)) Suppose that a Hadamard directional derivative $\lambda'_f : \mathbb{D} \rightarrow \mathbb{E}$ has estimator $\hat{\lambda}'_n : \mathbb{D} \rightarrow \mathbb{E}$ such that for some $\kappa > 0$ and $C_n \in \mathbb{R}_+$, $\|\hat{\lambda}'_n(h) - \hat{\lambda}'_n(k)\|_{\mathbb{E}} \leq C_n \|h - k\|_{\mathbb{D}}^{\kappa}$. Let \mathcal{P} denote a set of probability distributions. Suppose that (a) $C_n = O_P(1)$ uniformly over $P \in \mathcal{P}$, i.e., for all $\epsilon > 0$ there is an M such that $\sup_{P \in \mathcal{P}} P \{C_n > M\} < \epsilon$, and (b) for all $\epsilon > 0$, $\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left\{ \|\hat{\lambda}'_n(h) - \lambda'_f(h)\|_{\mathbb{E}} > \epsilon \right\} = 0$. Then for all compact $K \in \mathbb{D}$, letting K^{δ} for any $\delta > 0$ denote the δ -enlargement of K , for every $\epsilon > 0$ we have

$$\lim_{\delta \rightarrow 0^+} \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left\{ \sup_{h \in K^{\delta}} \|\hat{\lambda}'_n(h) - \lambda'_f(h)\|_{\mathbb{E}} > \epsilon \right\} = 0.$$

Proof of Lemma B.5. This is a straightforward adaptation of Lemma S.3.6 of Fang and Santos (2019), and only needs a few minor modifications. First, the assumption on C_n must be strengthened to hold uniformly over $P \in \mathcal{P}$. Most of the statements are analytic in nature (including the appeal to their Lemma S.3.4) and assumed to hold outer almost surely, and the probabilistic inequalities still hold when taking suprema over $P \in \mathcal{P}$. \square

Proof of Theorem 4.5. For convenience let $\mathcal{G}_n^* = r_n(f_n^* - f_n)$ and let $Z = \{Z_i\}_{i=1}^n$. Consider the subset \mathcal{P}_{00}^I , which is the collection of $P \in \mathcal{P}_0^I$ such that $X_0 \neq \emptyset$. Let $\hat{F}_{n,P}(s) = P \left\{ \hat{\lambda}'_{jn}(\mathcal{G}_n^*) \leq s | Z \right\}$

and $F_P(s) = P \left\{ \lambda'_{jf}(\mathcal{G}_P) \leq s \right\}$ be the distribution functions of bootstrapped and asymptotic statistics. We note that $\hat{q}_{\lambda^*}(1 - \alpha) = \inf\{q \in \mathbb{R} : \hat{F}_{n,P}(q) \geq 1 - \alpha\}$, and define $q_{\lambda'}^P(1 - \alpha) = \inf\{q \in \mathbb{R} : F_P(q) \geq 1 - \alpha\}$. Lemma 10.11 of Kosorok (2008) can be extended to hold uniformly over $P \in \mathcal{P}_{00}^I$, so that given the second result of Lemma B.4, under Assumption **A7**, for any closed $S \subset \mathbb{R}$ with $\min S > 0$,

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \left\{ \sup_{s \in S} |\hat{F}_{n,P}(s) - F_P(s)| > \epsilon \right\} = 0.$$

As in the proof of Theorem S.1.1 of Fang and Santos (2019), we find that for any $\epsilon > 0$,

$$\sup_{P \in \mathcal{P}_{00}^I} P \left\{ |\hat{q}_{\lambda^*}(1 - \alpha) - q_{\lambda'}^P(1 - \alpha)| > \epsilon \right\} \rightarrow 0, \quad (\text{S.16})$$

so that the critical values are consistent uniformly over $P \in \mathcal{P}_{00}^I$. Under the null hypothesis, $r_n \lambda_j(f_n) = r_n(\lambda_j(f_n) - \lambda_j(f))$, and Lemma B.4, Assumption **A7** and (S.16) imply

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} |P \{r_n \lambda_j(f_n) > \hat{q}_{\lambda^*}(1 - \alpha)\} - P \{\lambda'_{jf}(\mathcal{G}_P) > q_{\lambda'}^P(1 - \alpha)\}| = 0.$$

Since the second probability inside the absolute value is less than or equal to α , we have

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_{00}^I} P \{r_n \lambda_j(f_n) > \hat{q}_{\lambda^*}(1 - \alpha)\} \leq \alpha. \quad (\text{S.17})$$

Now consider the subset $\mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$, which implies that $X_0 = \emptyset$, that is, such that $\psi(f)(x) < 0$ for all $x \in X$. The definition of \hat{X}_0 ensures that under the assumption that $X_0 = \emptyset$,

$$\begin{aligned} P \left\{ \hat{X}_0 \neq X \right\} &= P \left\{ \sup_{x \in X} |\psi(f_n)(x)| \leq b_n \right\} \\ &\leq P \left\{ \sup_{x \in X} |r_n(\psi(f_n)(x) - \psi(f)(x))| + \sup_{x \in X} |r_n \psi(f)(x)| \leq r_n b_n \right\} \\ &\rightarrow 0 \end{aligned} \quad (\text{S.18})$$

uniformly over $P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$, where the inequality follows from the fact that $\sup_{x \in X} |\psi(f)(x)| > 0$ and $b_n \rightarrow 0$ and by Assumptions **A5** and **A6**. Let $\tilde{\lambda}'_{jn}$ be defined like their asymptotic counterparts $\tilde{\lambda}'_{jf}$ that were defined in Assumption **A7**:

$$\tilde{\lambda}'_{2n}(h) = \sup_{(u,x) \in U_{f_n}(X, a_n)} [h(u, x)]_+, \quad \tilde{\lambda}'_{4n}(h) = \left(\int_X \sup_{u \in U_{f_n}(x, a_n)} [h(u, x)]_+^p dm(x) \right)^p.$$

A small modification of the proof of the second part of Lemma B.4 (that is, without estimation of X_0) implies that $\tilde{\lambda}'_{jn}(r_n(f_n^* - f_n)) \rightsquigarrow \tilde{\lambda}'_{jf}(\mathcal{G}_P)$ in probability uniformly over $P \in \mathcal{P}_0^I$. Since also (S.18) implies that for all $\epsilon > 0$, $\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I} P \left\{ \left| \hat{\lambda}'_{jn}(\mathcal{G}_n^*) - \tilde{\lambda}'_{jn}(\mathcal{G}_n^*) \right| > \epsilon |Z \right\} = 0$,

we know also that $\hat{\lambda}'_{jn}(\mathcal{G}_n^*) \rightsquigarrow \tilde{\lambda}'_{jf}(\mathcal{G}_P)$ uniformly over $P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$. Under Assumption **A7**, there is some $\underline{q} > 0$ such that $\inf_{P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I} \inf \left\{ q \in \mathbb{R} : P \left\{ \tilde{\lambda}'_{jf}(\mathcal{G}_P) \leq q \right\} \geq 1 - \alpha \right\} \geq \underline{q}$. Using this fact with another small modification of the second part of Lemma B.4 implies that there is a $\underline{q} > 0$ such that

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I} P \left\{ \hat{q}_{\lambda^*}(1 - \alpha) \geq \underline{q} \right\} = 1. \quad (\text{S.19})$$

On the other hand, the first part of Lemma B.4 implies, as in Corollary 4.2 in the main text but uniformly over $P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I$, that for all $\epsilon > 0$, $\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I} P \{ r_n \lambda_j(f_n) > \epsilon \} = 0$. Combining this fact with (S.19) implies that

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0^I \setminus \mathcal{P}_{00}^I} P \{ r_n \lambda_j(f_n) > \hat{q}_{\lambda^*}(1 - \alpha) \} = 0. \quad (\text{S.20})$$

Then (S.17) and (S.20) together imply the result. \square

References

- ABBRING, J. H., AND J. J. HECKMAN (2007): “Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation,” in *Handbook of Econometrics (Chapter 72)*, ed. by J. Heckman, and E. Leamer. North Holland.
- ALIPRANTIS, C. D., AND K. C. BORDER (2006): *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Springer-Verlag, Heidelberg.
- ANDREWS, D. W., AND P. GUGGENBERGER (2010): “Asymptotic Size and a Problem with Subsampling the m out of n Bootstrap,” *Econometric Theory*, 26, 426–468.
- ANDREWS, D. W., AND X. SHI (2017): “Inference Based on Many Conditional Moment Inequalities,” *Journal of Econometrics*, 196, 275–287.
- ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): “Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure,” *Econometrica*, 74, 539–563.
- BICKEL, P. J., AND A. SAKOV (2008): “On the Choice of m in the m Out of n Bootstrap and Its Application to Confidence Bounds for Extreme Percentiles,” *Statistica Sinica*, 18, 967–985.
- BÜCHER, A., J. SEGERS, AND S. VOLGUSHEV (2014): “When Uniform Weak Convergence Fails: Empirical Processes for Dependence Functions and Residuals via Epi- and Hypographs,” *The Annals of Statistics*, 42, 1598–1634.
- CÁRCAMO, J., A. CUEVAS, AND L.-A. RODRÍGUEZ (2020): “Directional Differentiability for Supremum-Type Functionals: Statistical Applications,” *Bernoulli*, 26, 2143–2175.

- CATTANEO, M. D., M. JANSSON, AND K. NAGASAWA (2020): “Bootstrap-Based Inference for Cube Root Asymptotics,” *Econometrica*, 88, 2203–2219.
- CHEN, Q., AND Z. FANG (2019): “Inference on Functionals under First Order Degeneracy,” *Journal of Econometrics*, 210, 459–481.
- CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): “Intersection Bounds: Estimation and Inference,” *Econometrica*, 81, 667–737.
- CHESHER, A., AND A. M. ROSEN (2015): “Counterfactual Worlds,” Cemmap working paper CWP22/15.
- CHETVERIKOV, D., A. SANTOS, AND A. M. SHAIKH (2018): “The Econometrics of Shape Restrictions,” *Annual Review of Economics*, 10, 31–63.
- CHO, J. S., AND H. WHITE (2018): “Directionally Differentiable Econometric Models,” *Econometric Theory*, 34, 1101–1131.
- CHRISTENSEN, T., AND B. CONNAULT (2019): “Counterfactual Sensitivity and Robustness,” *arXiv e-prints*, p. arXiv:1904.00989.
- DANSKIN, J. M. (1967): *The Theory of Max-Min and Its Applications to Weapons Allocation Problems*. Springer-Verlag, New York.
- DAVYDOV, Y. A., M. A. LIFSHITS, AND N. V. SMORODINA (1998): *Local properties of distributions of stochastic functionals*, Translations of Mathematical Monographs. American Mathematical Society, Providence.
- DEHEJIA, R. H., AND S. WAHBA (1999): “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062.
- DÜMBGEN, L. (1993): “On Nondifferentiable Functions and the Bootstrap,” *Probability Theory and Related Fields*, 95, 125–140.
- FAN, Y., E. GUERRE, AND D. ZHU (2017): “Partial Identification of Functionals of the Joint Distribution of “Potential Outcomes”,” *Journal of Econometrics*, 197, 42–59.
- FAN, Y., AND S. S. PARK (2009): “Partial Identification of the Distribution of Treatment Effects and its Confidence Sets,” in *Nonparametric Econometric Methods (Advances in Econometrics, Volume 25)*, ed. by Q. Li, and J. Racine, pp. 3–70. Emerald Group Publishing Limited, Bingley, Series editors Thomas B. Fomby and R. Carter Hill.
- (2010): “Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference,” *Econometric Theory*, 26, 931–951.

- (2012): “Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments,” *Journal of Econometrics*, 167, 330–344.
- FAN, Y., R. SHERMAN, AND M. SHUM (2014): “Identifying Treatment Effects Under Data Combination,” *Econometrica*, 82, 811–822.
- FAN, Y., AND J. WU (2010): “Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and its Confidence Sets,” *Review of Economic Studies*, 77, 1002–1041.
- FANG, Z., AND A. SANTOS (2019): “Inference on Directionally Differentiable Functions,” *Review of Economic Studies*, 86, 377–412.
- FIRPO, S. (2007): “Efficient Semiparametric Estimation of Quantile Treatment Effects,” *Econometrica*, 75, 259–276.
- FIRPO, S., AND G. RIDDER (2008): “Bounds on Functionals of the Distribution of Treatment Effects,” IEPR Working Paper No. 08-09.
- (2019): “Partial Identification of the Treatment Effect Distribution and Its Functionals,” *Journal of Econometrics*, 213, 210–234.
- FOLLAND, G. B. (1999): *Real Analysis: Modern Techniques and Their Applications*. Wiley, New York, 2nd edn.
- FRANK, M. J., R. B. NELSEN, AND B. SCHWEIZER (1987): “Best-Possible Bounds for the Distribution of a Sum — A Problem of Kolmogorov,” *Probability Theory and Related Fields*, 74, 199–211.
- GAUTIER, E., AND S. HODERLEIN (2011): “A Triangular Treatment Effect Model with Random Coefficients in the Selection Equation,” *arXiv e-prints*, p. arXiv:1109.0362.
- GINÉ, E., AND J. ZINN (1991): “Gaussian Characterization of Uniform Donsker Classes of Functions,” *The Annals of Probability*, 19, 758–782.
- HECKMAN, J. J., AND J. HOTZ (1989): “Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training,” *Journal of the American Statistical Association*, 84, 862–874.
- 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, 487–535.
- HONG, H., AND J. LI (2018): “The Numerical Delta Method,” *Journal of Econometrics*, 206, 379–394.
- IMBENS, G. W. (2003): “Sensitivity to Exogeneity Assumptions in Program Evaluation,” *American Economic Review*, 93, 126–132.

- IMBENS, G. W., AND C. F. MANSKI (2004): “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72, 1845–1857.
- KAJI, T. (2019): “Asymptotic Theory of L -Statistics and Integrable Empirical Processes,” *arXiv e-prints*, p. arXiv:1910.07572.
- KIM, J. H. (2014): “Identifying the Distribution of Treatment Effects under Support Restrictions,” *arXiv e-prints*, p. arXiv:1410.5885.
- KOSOROK, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.
- LALONDE, R. J. (1986): “Evaluating the Econometric Evaluations of Training Programs with Experimental Data,” *American Economic Review*, 76, 604–620.
- LEE, S., K. SONG, AND Y.-J. WHANG (2018): “Testing For a General Class of Functional Inequalities,” *Econometric Theory*, 34, 1018–1064.
- LINTON, O., K. SONG, AND Y.-J. WHANG (2010): “An Improved Bootstrap Test of Stochastic Dominance,” *Journal of Econometrics*, 154, 186–202.
- MAKAROV, G. D. (1982): “Estimates for the Distribution Function of a Sum of Two Random Variables when the Marginal Distributions are Fixed,” *Theory of Probability and its Applications*, 26(4), 803–806.
- MASTEN, M. A., AND A. POIRIER (2020): “Inference on Breakdown Frontiers,” *Quantitative Economics*, 11, 41–111.
- ROCKAFELLAR, R. T., AND R. J.-B. WETS (1998): *Variational Analysis*. Springer-Verlag, Berlin.
- ROMANO, J. P., AND A. M. SHAIKH (2012): “On the Uniform Asymptotic Validity of Subsampling and the Bootstrap,” *The Annals of Statistics*, 40, 2798–2822.
- RÖMISCH, W. (2006): “Delta Method, Infinite Dimensional,” in *Encyclopedia of Statistical Sciences*, ed. by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels. Wiley, 3.
- RÜSCHENDORF, L. (1982): “Random Variables with Maximum Sums,” *Advances in Applied Probability*, 14, 623–632.
- SHAPIRO, A. (1990): “On Concepts of Directional Differentiability,” *Journal of Optimization Theory and Applications*, 66, 477–487.
- SHEEHY, A., AND J. A. WELLNER (1992): “Uniform Donsker Classes of Functions,” *The Annals of Probability*, 20, 1983–2030.

- SMITH, J. A., AND P. E. TODD (2001): “Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods,” *American Economic Review*, 91, 112–118.
- (2005): “Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?,” *Journal of Econometrics*, 125, 305–353.
- SOMMERFELD, M., AND A. MUNK (2017): “Inference for empirical Wasserstein distances on finite spaces,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 219–238.
- VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.
- VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.
- WILLIAMSON, R. C., AND T. DOWNS (1990): “Probabilistic Arithmetic I. Numerical Methods for Calculating Convolutions and Dependency Bounds,” *International Journal of Approximate Reasoning*, 4, 89–158.