Uniform inference for bounds on the distribution and quantile functions of treatment effects in randomized experiments

Antonio F. Galvao† Thomas Parker‡

November 26, 2019

Abstract

This paper develops a novel approach to uniform inference for functions that bound the distribution and quantile functions of heterogeneous treatment effects in randomized experiments when only marginal treatment and control distributions are observed and the joint distribution of outcomes is unobserved. These bounds are nonlinear maps of the marginal distribution functions of control and treatment outcomes, and statistical inference methods for nonlinear maps usually rely on smoothness through a type of differentiability. We show that the maps from marginal distributions to bound functions are not differentiable, but uniform test statistics applied to the bound functions — such as Kolmogorov-Smirnov or Cramér-von Mises — are directionally differentiable. We establish the consistency and weak convergence of nonparametric plug-in estimates of the test statistics and show how they can be used to conduct inference for bounds uniformly over the distribution of treatment effects. We also establish the directional differentiability of minimax operators applied to general — that is, not only convex-concave — functions, which may be of independent interest. In addition, we develop detailed resampling techniques to conduct practical inference for the bounds or for the true distribution or quantile function of the treatment effect distribution. Finally, we apply our methods to the evaluation of a job training program.

Keywords: Partial Identification, Uniform Inference, Treatment Effects, Directional Differentiability

JEL codes: C12, C14, C21

*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. Computer programs to replicate all the numerical analyses are available from the authors. 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).

†Department of Economics, University of Arizona, Tucson, USA. E-mail: agalvao@email.arizona.edu
‡Department of Economics, University of Waterloo, Waterloo, Canada. E-mail: tmparker@uwaterloo.ca
1 Introduction

Treatment effects models have provided a valuable method of statistical analysis and causal effects in program evaluation studies. When the treatment effect is heterogeneous, its distribution function (CDF) and related features, for example its quantile function, are often used in order to answer many interesting policy questions. Nevertheless, it is well known in the literature (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 became 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. We refer the reader to Tamer (2010) for a discussion of the importance of partially identified models in econometrics. In particular, the cumulative distribution function and quantile function of the treatment effects distribution are two features of fundamental importance for inference beyond the mean. At each point in the support of the distribution, these two functions are only interval identified without knowledge beyond the marginal distributions of the treatment and control groups. Bounds 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.

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. Fan and Park (2009, 2010) study sharp pointwise bounds on the distribution function and quantile function of treatment effects and corresponding inference, from which they derive sharp bounds on the class of D-parameters (features that vary according to stochastic dominance ordering). In particular, Fan and Park (2010) derive asymptotic properties of estimated bounds on the treatment effect CDF at a point. Fan and Park (2012) discuss estimated bounds on the quantile function of the treatment effect evaluated at a point and establish their asymptotic properties. Firpo and Ridder (2008, 2019); Fan, Guerre, and Zhu (2017) consider bounding other functionals of the distribution of treatment effects using bounds on the treatment effect CDF. There are several works extending these models to more general frameworks, as for example, Fan and Wu (2010). 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 providing 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.

This paper departs from that literature by developing a novel approach for investigating heterogeneous treatment effects uniformly over the distribution and quantile functions in randomized experiments. The methods developed in this paper extend those for the nonparametric estimation of bounds in Fan and Park (2009, 2010, 2012) from the pointwise to the uniform case. Uniform inference has an important role for the analysis of heterogeneity in treatment effects models, and provides a parsimonious strategy for the study of changes over an entire distribution of interest. It is important to notice the advantages of uniform inference over the pointwise counterpart. Angrist, Chernozhukov, and Fernandez-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.

We establish uniform inference results for the bound functions as objects of interest and in a simple example use them to bound the underlying distribution and quantile functions. These bounds are nonlinear maps of the marginal distribution functions of control and treatment outcomes. In order to apply established statistical inference procedures, one must show that the maps are differentiable in a sense that will be made clear below. As a first step, we note that the maps of marginal distribution to bound functions are not differentiable in the required sense. However, uniform test statistics applied to the bound functions — such as Kolmogorov-Smirnov or Cramér-von Mises tests — are directionally differentiable. Second, despite this lack of full differentiability, we establish the asymptotic properties of nonparametric plug-in estimators of these test statistics and invert these tests to conduct uniform inference. This result may be of interest for further research on convolutions other than the difference between two random variables (see, for example, the other examples discussed in Williamson and Downs (1990)). Third, as one unexpected by-product of our main theorem, we establish the directional differentiability of minimax operators applied to general — that is, not necessarily convex-concave — functions. This may be a useful generalization of the convex-concave function case that is prevalent in the literature. The methods proposed in this paper are useful to practitioners, as for example in testing hypotheses uniformly over the CDF and quantiles of the distribution of treatment effects, in constructing uniform confidence bands for these quantities, and in the construction of a dominance test without knowledge of the correlations between the distributions in the comparison.

The uniform results derived in this paper depend on the notion of Hadamard (or compact) differentiability and its directional generalization. There is recent growing literature on applications of this important result — see, e.g., Hong and Li (2016), Cattaneo, Jansson, and Nagasawa
(2017), Masten and Poirier (2017), Chetverikov, Santos, and Shaikh (2018), Cho and White (2018), Christensen and Connault (2019) and Fang and Santos (2019). The directional differentiability of uniform test statistics provides a minimal amount of regularity needed for feasible statistical inference, which we implement using the modified bootstrap methods developed in Fang and Santos (2019). We highlight that the proposed resampling methods are simple to implement in practice. Further results in Fang and Santos (2019) make it easy to show that for certain tests, like those used to construct confidence bands via inversion, our bootstrap procedures control size uniformly across the null region.

Monte Carlo simulations assess the finite-sample properties of the proposed methods. The simulations verify that Kolmogorov-Smirnov-type statistics have accurate empirical size and power against selected local alternatives. Cramér-von Mises type statistics are used to test a stochastic dominance hypothesis using bound functions, similar to the breakdown frontier tests of Masten and Poirier (2017). This test has accurate size at the breakdown frontier and power against alternatives that violate the frontier. All results are improved when the sample size increases and are powerful using a modest number of bootstrap repetitions.

To illustrate the methods, we bound the treatment effect distribution function for the 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). In particular, we use an experimental data set with information from the National Supported Work Program used by Dehejia and Wahba (1999). The interest is to identify the distributional causal effects of job training on future earnings and construct uniform confidence bounds for the CDF of the treatment effects. The main findings of the empirical application 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 and is the distribution function of a random variable constructed via convolution.

The remainder of the paper is organized as follows. Section 2 defines the population bounds for the distribution and quantile functions of the treatment effects, and establishes the directional differentiability of statistics applied to the bound functions. Section 2.6 derives the uniform asymptotic results for the corresponding estimators. Inference procedures are established in Section 3. Section 4 provides Monte Carlo simulations, and an application to job training is discussed in Section 5. Finally, Section 6 concludes. We relegate the proofs of the results to the Appendix.
Notation

Let \( F = (F_0, F_1) \) denote the pair of marginal population distribution functions and \( F_n = (F_{0n}, F_{1n}) \) be the pair of marginal sample distribution functions, respectively, where \( F_{kn}(y) = \frac{1}{n} \sum_{i=1}^{n} 1(Y_{ki} \leq y) \) for \( k \in \{0, 1\} \). Let \( E \) denote the expectation operator. For any set \( T \subseteq \mathbb{R}^d \) let \( C(T) \) denote the space of continuous functions \( f : T \to \mathbb{R} \) and let \( D(T) \) denote the space of cadlag functions \( f : T \to \mathbb{R} \), both equipped with the uniform norm \( \| f \|_\infty = \sup_{t \in T} |f(t)| \). The space of bounded functions on \( T \) is denoted \( \ell_\infty(T) = \{ f : T \to \mathbb{R} \text{ with } \| f \|_\infty < \infty \} \). Given a sequence \( \{f_n\}_n \subset \ell_\infty(T) \) and limiting random element \( f \) we write \( f_n \overset{w}{\to} f \) to denote weak convergence in \( (\ell_\infty(T), \| \cdot \|_\infty) \) in the sense of Hoffmann-Jørgensen (van der Vaart and Wellner, 1996), and we write \( f \sim g \) to mean \( f \) has the same distribution as \( g \). For function \( f \), \( \| f \|_p \) denotes the \( L^p \) norm for \( 1 \leq p \leq \infty \). For \( x \in \mathbb{R}^d \), \( \| x \| \) denotes the usual Euclidean norm. Denote \( \lfloor x \rfloor_+ = \max\{x, 0\} \). A set-valued map (or correspondence) \( S \) that maps elements of \( X \) to the collection of subsets of \( Y \) is denoted \( S : X \rightrightarrows Y \).

2 Mapping marginal distributions to test statistics

In this section, we first describe the population treatment effect bounds and test statistics of interest. Second, we formally review the concept of directional differentiability, and show that the uniform test statistics applied to the test statistic functionals are directionally differentiable. Third, we establish consistency and weak convergence of nonparametric plug-in estimates of the test statistics.

2.1 The test statistics

We consider the distribution and quantile functions of treatment effects in randomized experiments. Suppose a binary treatment is independent of two potential outcomes \( (Y_0, Y_1) \), where \( Y_0 \) denotes outcomes under a control regime and \( Y_1 \) denotes outcomes under a treatment. Suppose that \( Y_0 \) and \( Y_1 \) have marginal distribution functions \( F_0 \) and \( F_1 \) respectively. Suppose that the econometrician is interested in the distribution of the treatment effect \( \Delta = Y_1 - Y_0 \) but unwilling to make any assumptions regarding the dependence between \( Y_0 \) and \( Y_1 \). In this section we study the relationship between the identifiable functions \( F_0 \) and \( F_1 \) and functions that bound the distribution function or quantile function of the unobservable random variable \( \Delta \). This is purely an analytic exercise that does not require the consideration of observable samples. The resulting population objects in this section will play a part in inference using observed data in subsequent sections.

The distribution function of interest is \( F_\Delta(\cdot) = \mathbb{P}\{\Delta \leq \cdot\} \), which is not point-identified because the full bivariate distribution of \( (Y_0, Y_1) \) is unidentified and the econometrician has no knowledge of the dependence between potential outcomes. However, \( F_\Delta \) can be bounded. Define the distribution
Although we have defined four potential estimands and plug-in estimators, inference for all four
bounds

\[ L(\delta) = \sup_{y \in \mathbb{R}} \{ F_1(y) - F_0(y - \delta) \} \quad (1) \]

\[ U(\delta) = \inf_{y \in \mathbb{R}} \{ 1 + F_1(y) - F_0(y - \delta) \} . \quad (2) \]

The functions \( L \) and \( U \), in (1) and (2), respectively, were derived independently by Makarov (1982) (apparently resolving a conjecture posed by Kolmogorov), Rüschendorf (1982), Frank, Nelsen, and Schweizer (1987), and extended by Williamson and Downs (1990) to the convolutions defined by the four basic binary arithmetic operators on \( Y_0 \) and \( Y_1 \). The functions \( \tau_{L}, \tau_{U} \) and \( \tau_{0} \) are pointwise sharp: for any fixed \( \delta \) there exist some \( Y_{0}^{*} \) and \( Y_{1}^{*} \) such that the resulting \( \Delta^{*} = Y_{1}^{*} - Y_{0}^{*} \) has a distribution function \( F^{*} \) such that \( F^{*}(\delta) = L(\delta) \) or \( F^{*}(\delta) = U(\delta) \).

The case of bounds for quantile functions is similar. For any nondecreasing function \( f \) let \( f^{-1}(u) = \inf \{ x : f(x) \geq u \} \) be its generalized inverse. As discussed in Fan and Park (2010), the functions in the relationship \( L \leq F_{\Delta} \leq U \) may be inverted to find that for each quantile \( 0 < \tau < 1 \), \( U^{-1}(\tau) \leq F_{\Delta}^{-1}(\tau) \leq L^{-1}(\tau) \). These inverted bounds are (Williamson and Downs, 1990)

\[ U^{-1}(\tau) = \sup_{u \in (0, \tau)} \{ F_{1}^{-1}(u) - F_{0}^{-1}(u + (1 - \tau)) \} \quad (3) \]

\[ L^{-1}(\tau) = \inf_{u \in (\tau, 1)} \{ F_{1}^{-1}(u) - F_{0}^{-1}(u - \tau) \} . \quad (4) \]

Suppose we observe samples \( \{ Y_{ki} \}_{i=1}^{n_{k}} \) for \( k \in \{ 0, 1 \} \). Let \( L_{n} \) and \( U_{n} \) be the nonparametric plug-in estimates of the bounds based on the empirical distribution functions: for each \( \delta \in \mathbb{R} \), let

\[ L_{n}(\delta) = \sup_{y \in \mathbb{R}} \{ F_{1n}(y) - F_{0n}(y - \delta) \} \quad (5) \]

\[ U_{n}(\delta) = \inf_{y \in \mathbb{R}} \{ 1 + F_{1n}(y) - F_{0n}(y - \delta) \} . \quad (6) \]

Similarly to the distribution function case, plug-in estimates of the \( \tau \)-quantile bounds are

\[ U_{n}^{-1}(\tau) = \sup_{u \in (0, \tau)} \{ F_{1n}^{-1}(u) - F_{0n}^{-1}(u + (1 - \tau)) \} \quad (7) \]

\[ L_{n}^{-1}(\tau) = \inf_{u \in (\tau, 1)} \{ F_{1n}^{-1}(u) - F_{0n}^{-1}(u - \tau) \} . \quad (8) \]

Although we have defined four potential estimands and plug-in estimators, inference for all four
functions is quite similar, involving only notational changes to move from one case to the next. In what follows we focus on uniform inference for $L$, the lower bound of the distribution function of the treatment effect, because the upper bound uses near-identical operations and the quantile function bounds are one (compactly) differentiable transformation away from the distribution function bounds.

There are two immediate applications of uniform inference procedures applied to the difference between empirical and theoretical bound functions. First, for the purposes of testing the null hypothesis $H_0 : L = L_0$ against $H_1 : L \neq L_0$, one would consider the distribution of

$$
\sup_{\delta \in \mathbb{R}} \sqrt{n} |L_n(\delta) - L_0(\delta)| \quad \text{or} \quad \left( \int_{\mathbb{R}} |L_n(\delta) - L_0(\delta)|^p d\delta \right)^{1/p},
$$

corresponding to classical Kolmogorov-Smirnov and (setting $p = 2$) Cramér-von Mises type test statistics. The Kolmogorov-Smirnov test can be used to construct uniform confidence bands by inverting the test statistic distribution, which we discuss below in subsection 3.1. Our second application is a dominance test: suppose that $\{Y_{0i}\}_{i=1}^n$ are observed for $k \in \{0, A, B\}$ where $A$ and $B$ correspond to two different treatments and 0 corresponds to the control group. Then the condition $H_0 : U_A \leq L_B$ implies (but is not implied by) first-order stochastic dominance of treatment $A$ over treatment $B$ regardless of their correlations with the treatment outcomes since $F_A \leq U_A \leq L_B \leq F_B$. One may test for a sufficient condition for stochastic dominance using tests similar to Linton, Song, and Whang (2010):

$$
\sup_{\delta \in \mathbb{R}} \sqrt{n} |U_{An}(\delta) - L_{Bn}(\delta)|_+ \quad \text{or} \quad \left( \int_{\mathbb{R}} |U_{An}(\delta) - L_{Bn}(\delta)|_+^p d\delta \right)^{1/p}.
$$

These functionals are generally close to zero when the null hypothesis $U_A \leq L_B$ is satisfied and greater than zero when it is violated. We return to this example in subsection 3.2 below.

The applications described above lead us to consider the following statistics $\Lambda_j : (\ell^\infty(X))^2 \to \mathbb{R}$:

$$
\begin{align*}
\Lambda_1(f) &= \sup_{x \in X} \left| \sup_{u \in A(x)} (f_1(u) - f_0(u - x)) \right|, \\
\Lambda_2(f) &= \sup_{x \in X} \left[ \sup_{u \in A(x)} (f_1(u) - f_0(u - x)) \right]_+^p \left( \int_X \left[ \sup_{u \in A(x)} (f_1(u) - f_0(u - x)) \right]_+^p dx \right)^{1/p}, \\
\Lambda_3(f) &= \left( \int_X \left[ \sup_{u \in A(x)} (f_1(u) - f_0(u - x)) \right]_+^p dx \right)^{1/p}, \\
\Lambda_4(f) &= \left( \int_X \left[ \sup_{u \in A(x)} (f_1(u) - f_0(u - x)) \right]_+^p dx \right)^{1/p}.
\end{align*}
$$

The choice map $A : X \Rightarrow U$ used in the definitions above allows one to accommodate the case of
quantile function evaluation that takes place over intervals that change with an argument. For example, in the case of the lower bound for the distribution function, \( f \) in these definitions corresponds to \( F_1(y) - F_0(y - \delta) \) and the supremum is evaluated over \( A(\delta) = \mathbb{R} \), while for the lower bound of the quantile function \( f \) is \( F_1^{-1}(u) - F_0(u - \tau) \), and the supremum is evaluated over \( A(\tau) = (0, \tau) \). The use of \( \Lambda_3 \) or \( \Lambda_4 \) also requires the assumption that the integrals exist; we assume \( f \in \ell^\infty(X) \cap L^p(X) \) when using these test statistics.

### 2.2 Directional differentiability

A test statistic such as \( \| \sqrt{n}(L_n - L_0) \|_p \) can be viewed as a perturbation of the nonlinear map from \( F := (F_0, F_1) \in (\ell^\infty(\mathbb{R}))^2 \) to \( \| L \|_p \in \mathbb{R} \). The manner in which asymptotic statistical inference is usually conducted for such nonlinear maps requires the delta method to show that a well-behaved limit distribution exists. The delta method in turn demands that the map from the data distribution to the test statistic is differentiable in some sense. The appropriate notion of differentiability in the empirical processes context is Hadamard (or compact) differentiability (van der Vaart and Wellner, 1996, Section 3.9), and it is very convenient to split the analysis of statistical behavior into two steps: an analytic one demonstrating compact differentiability and a second step using the delta method. However, the \( \Lambda_j \) maps in (9) are not Hadamard differentiable in the standard sense. Shapiro (1990), Dümbgen (1993) and Fang and Santos (2019) discuss Hadamard directional differentiability and show that this weaker notion allows for the application of the delta method. The major result in this subsection is that the \( \Lambda_j \) functionals are Hadamard directionally differentiable.

We start with the definition of Hadamard differentiability and Hadamard directional differentiability, following van der Vaart and Wellner (1996) and Fang and Santos (2019).

**Definition 2.1 (Hadamard differentiability).** Let \( \mathbb{D} \) and \( \mathbb{E} \) be Banach spaces and consider a map \( \lambda : \mathbb{D}_\lambda \subseteq \mathbb{D} \rightarrow \mathbb{E} \).

1. \( \lambda \) is **Hadamard differentiable** at \( f \in \mathbb{D}_\lambda \) tangentially to a set \( \mathbb{D}_0 \subseteq \mathbb{D} \) if there is a continuous linear map \( \lambda' : \mathbb{D}_0 \rightarrow \mathbb{E} \) such that

   \[
   \lim_{n \rightarrow \infty} \left\| \frac{\lambda(f + t_nh_n) - \lambda(f)}{t_n} - \lambda'(h) \right\|_E = 0
   \]

   for all sequences \( \{h_n\} \subseteq \mathbb{D} \) and \( \{t_n\} \subseteq \mathbb{R} \) such that \( h_n \rightarrow h \in \mathbb{D}_0 \) and \( t_n \rightarrow 0 \) as \( n \rightarrow \infty \) and \( f + t_nh_n \in \mathbb{D}_\lambda \) for all \( n \).

2. \( \lambda \) is **Hadamard directionally differentiable** at \( f \in \mathbb{D}_\lambda \) tangentially to a set \( \mathbb{D}_0 \subseteq \mathbb{D} \) if there is a continuous map \( \lambda'_f : \mathbb{D}_0 \rightarrow \mathbb{E} \) such that

   \[
   \lim_{n \rightarrow \infty} \left\| \frac{\lambda(f + t_nh_n) - \lambda(f)}{t_n} - \lambda'_f(h) \right\|_E = 0
   \]
for all sequences \( \{h_n\} \subseteq \mathbb{D} \) and \( \{t_n\} \subseteq \mathbb{R}_+ \) such that \( h_n \to h \in \mathbb{D}_0 \) and \( t_n \to 0^+ \) as \( n \to \infty \) and \( f + t_nh_n \in \mathbb{D}_\lambda \) for all \( n \).

From Definition 2.1 we see that, in case \( \lambda \) is only directionally differentiable, the derivative map \( \lambda' \) need not be linear although \( \lambda' \) is continuous (Shapiro, 1990, Proposition 3.1). The derivative is linear if \( \lambda \) is fully differentiable, and in that case it does not depend on \( f \), while directional derivatives do. The delta method and a corresponding chain rule can be applied to maps that are either Hadamard differentiable or Hadamard directionally differentiable (van der Vaart and Wellner, 1996; Römisch, 2006; Shapiro, 1990).

The next step is to establish directional differentiability of the statistics in (9). Some parts of the map from \( F \mapsto \|L\|_p \) are relatively easy to work with. The map \( \Pi \) defined by \( f \mapsto f_1(s) - f_0(s-t) \) is linear and it can be verified that its Hadamard derivative at \( f \in (\ell^\infty(\mathbb{R}))^2 \) in direction \( h \in (\ell^\infty(\mathbb{R}))^2 \) is \( \Pi'_f(h)(s,t) = \Pi(h)(s,t) = h_1(s) - h_0(s-t) \) for \( h = (h_0, h_1) \in (\ell^\infty(\mathbb{R}))^2 \). Because this map is fully Hadamard differentiable and linear its derivative does not depend on \( f \). The map from distribution to quantile functions has been dealt with before, and is also uniformly well-behaved under a regularity condition.

**Q1** Assume \( \tau \in \mathcal{T} \) where \( \mathcal{T} = (\epsilon, 1-\epsilon) \) where \( \epsilon \in (0, 1/2) \). Suppose there exists an interval \( \mathcal{D} = [F^{-1}_\Delta(\epsilon) - \eta, F^{-1}_\Delta(1-\epsilon) + \eta] \) for some \( \eta > 0 \) such that both marginal distribution functions \( F_0 \) and \( F_1 \) are continuously differentiable with strictly positive derivatives on \( \mathcal{D} \).

This assumption assures well-defined quantile functions and derivatives. If, alternatively, it is assumed that \( \Delta \) lies in a compact set, then analysis for these inverses could be conducted uniformly over \((0, 1)\) instead of \( \mathcal{T} \). See for example equations (39) and (40) and the discussion of inverses on page 96 of Williamson and Downs (1990) and the related Lemma 3.9.23 of van der Vaart and Wellner (1996). Under Assumption **Q1**, the map \( Q : (D(\mathcal{D}))^2 \to (\ell^\infty(T))^2 \) is Hadamard differentiable tangentially to \((C(T))^2\) with Hadamard derivative at \( F \in (D(\mathcal{D}))^2 \) given by \( Q'_F(h)(t) = ((-h_0/f_0) \circ F^{-1}_0(t), (-h_1/f_1) \circ F^{-1}_1(t)) \) cf. Section 3.9.4.2 of van der Vaart and Wellner (1996) — the quantile transformation is Hadamard differentiable at any point where the distribution function is continuous, but continuity over the whole domain is required to discuss uniform results. However, the rest of the chain, that is, between \( F_0(y) - F_1(y-\delta) \) and \( \|L\|_p \), is more complex.

The problematic part of the map \( F \mapsto \|L\|_p \) is related to the pointwise supremum operation in the definition of \( L \). Consider two sets \( U \subseteq \mathbb{R}^j \) and \( X \subseteq \mathbb{R}^k \), a function \( f \in \ell^\infty(U \times X) \) and set-valued map \( A : X \rightharpoonup U \) that serves as a choice set, which we assume is non-empty-valued and continuous. Without loss of generality we focus on the supremum only and appeal to the duality between the supremum and infimum maps. The consideration of general bivariate \( f \) broadens the range of potential applications of our result, although for the specific functions in (9) above, the relevant bivariate functions take the form \( f(u,x) = f_1(u) - f_0(u-x) \). Instead of \( \Lambda_j \) defined in (9),
we consider the following maps $\lambda_j : \ell^\infty(U \times X) \rightarrow \mathbb{R}$:

$$
\lambda_1(f) = \sup_{x \in X} \left( \sup_{u \in A(x)} f(u,x) \right), \quad \lambda_2(f) = \sup_{x \in X} \left( \sup_{u \in A(x)} f(u,x) \right)_+, \\
\lambda_3(f) = \left( \int_X \left( \sup_{u \in A(x)} f(u,x) \right)^p \, dx \right)^{1/p}, \quad \lambda_4(f) = \left( \int_X \left( \sup_{u \in A(x)} f(u,x) \right)^p_+ \, dx \right)^{1/p}.
$$

These equations are essentially the same as (9) above, except that they apply more generally to bivariate functions — this makes it easier to understand the way our results can be generalized, especially since the special structure of the convolution problems does not contribute to any of the results below.

2.3 Problem: a non-differentiable map

In this subsection we discuss one important negative result, the lack of differentiability in the map of a bivariate function to a univariate function resulting from maximizing the bivariate function along one coordinate. This may serve as a warning with regard to naive applications of resampling procedures. Some of the notation will be re-used in later sections.

Let $\psi : \ell^\infty(U \times X) \rightarrow \ell^\infty(X)$ be the value function obtained by optimizing the objective $f$ with respect to $u \in A(x)$ for each $x \in X$ and let $S_f : X \rightrightarrows U$ denote the set-valued map of maximizers of $f$:

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

and define the set of maximizers and $\epsilon$-maximizers of $f$

$$
S_f(x) = \{ u \in A(x) : f(u,x) = \psi(f)(x) \} \\
S_f^+(x,\epsilon) = \{ u \in A(x) : f(u,x) \geq \psi(f)(x) - \epsilon \}.
$$

The value function or envelope map labeled $\psi$ has a long history in optimization and statistics and results on its (pointwise directional) derivatives date to Danskin (1967). See Milgrom and Segal (2002) for a good introduction to this pointwise case. Cárcamo, Cuevas, and Rodríguez (2019, Theorem 2.1) show that for directions $h \in \ell^\infty(U \times X)$,

$$
\psi_f'(h)(x) = \lim_{\epsilon \rightarrow 0^+} \sup_{u \in S_f^+(x,\epsilon)} h(u,x).
$$

Assuming the stronger conditions that $A$ is continuous and compact-valued and that $f$ is continuous on $U \times X$, the maximum theorem implies that $S_f$ is non-empty, compact-valued and upper hemicontinuous (Aliprantis and Border, 2006, Theorem 17.31), and Cárcamo, Cuevas, and Rodríguez
(2019, Corollary 2.2) show that tangentially to $\mathcal{C}(U \times \{x\})$, the derivative of $\psi(f)(x)$ simplifies to

$$\psi'_f(h)(x) = \sup_{u \in S_f(x)} h(u, x).$$

The pointwise directional differentiability of $\psi(f)(x)$ at each $x$ might lead one to suspect that $\psi$ is differentiable more generally as a map from $\ell^\infty(U \times X)$ to $\ell^\infty(X)$ or perhaps tangentially to maps from $\mathcal{C}(U \times X)$ to $\mathcal{C}(X)$. However, this is not true. $S_f$ may in general contain more than one element, and at those points the left-hand and right-hand derivatives of the envelope $\psi(f)(x)$ need not match — see, for example, Figure 1 or Theorem 3 of Milgrom and Segal (2002) for a related phenomenon. In the relatively well-behaved case in which $f$ and $h$ are continuous, the sequence of approximating finite difference functions $(\psi(f + t_n h)(x) - \psi(f)(x))/t_n$ are continuous for each $n$. Because a sequence of continuous functions can not converge uniformly to a discontinuous limit, $\psi$ as defined cannot be compactly differentiable as a map from $\mathcal{C}(U \times X)$ to $\mathcal{C}(X)$.

The lack of uniformity of convergence to $\psi'_f(h)(·)$ appears to be a problem for the use of the delta method for uniform inference. Thinking of the map $F \mapsto \Lambda_j$ as the chain $\Lambda_j = (\|\cdot\|_p \circ \psi \circ \Pi)(F)$, for example, presents a problem because $\psi$ is not compactly differentiable. However, we show below that this issue can be circumvented. In particular, the maps $F \mapsto \Lambda_j$, for $j = 1, \ldots, 4$ are Hadamard directionally differentiable, heuristically speaking, using the alternate chain $\Lambda_j = (\lambda_j \circ \Pi)(F)$. The chain rule applies when the elements in the chain are each differentiable. It does not preclude the possibility that problematic elements of one proposed chain of maps might be combined together to form perhaps more complex, but differentiable, elements of an alternative chain of maps.

The function $L$ (or $U$) has the properties of a CDF: it is zero at the left endpoint of its support, and rises monotonically to one at the right end of its support. This makes it tempting to make bootstrap draws from $L_n$ for inference, as in the standard iid bootstrap. However, the lack of differentiability of $F \mapsto L$ implies that one cannot treat the estimated bound function $L_n$ simply as an empirical CDF and apply standard bootstrap procedures with it. The resampled data generating process will not reproduce the kinks inherent in the map from samples to bounds. More precisely, it cannot be verified that the statistical behavior of $\sqrt{n}(\Lambda^* - L_n)$ obtained by nonparametric bootstrap is consistent because the derivative of $\sqrt{n}(L_n - L)$ does not converge weakly to a well-defined limit process in the usual (Hoffmann-Jørgensen) sense.

### 2.4 Minimax problems

This subsection verifies the differentiability of minimax operators for bivariate functions. This can be considered a generalization of Theorem 5 of Milgrom and Segal (2002), Proposition 4 of Demyanov (2009) or Proposition 2.1 of Shapiro (2008), and similar to Lemma 4.4 of Christensen and Connault (2019) but in a different context. The standard problem proceeds by assuming that
the objective function \( f(u, x) \) is convex-concave — convex in one argument and concave in the other. We actually show the result for maximin problems because this fits most easily with other results in this paper, but the duality between minimization and maximization means it applies to both problems.

Theorem 2.2 below shows that “saddle-point” problems for general bivariate functions are Hadamard directionally differentiable. Note that this cannot be proven using a chain of two directional derivatives for the maximization and minimization steps, because the inner optimization problem is not differentiable as a map to the space of functions, as discussed in subsection 2.3. Nevertheless, the result can be shown using techniques similar to the case of the supremum operator. Theorem 2.2, which builds upon the work of Cárcamo, Cuevas, and Rodríguez (2019), only assumes that the objective function is a bounded function — no convexity is assumed.

**Theorem 2.2.** Suppose that \( U \subseteq \mathbb{R}^j \) and \( X \subseteq \mathbb{R}^k \) and \( f, h \in \ell^\infty(U \times X) \). Let \( A : X \rightrightarrows U \) be non-empty-valued and define \( \sigma : \ell^\infty(U \times X) \to \mathbb{R} \) by

\[
\sigma(f) = \sup_{x \in X} \inf_{u \in A(x)} f(u, x).
\]

For \( \delta, \epsilon > 0 \), define

\[
X(\delta) = \left\{ x \in X : \inf_{u \in A(x)} f(u, x) \geq \sigma(f) - \delta \right\}
\]

and

\[
U(x, \epsilon) = \left\{ u \in A(x) : f(u, x) \leq \inf_{u \in A(x)} f(u, x) + \epsilon \right\}.
\]

Then \( \sigma \) is Hadamard directionally differentiable with derivative at \( f \) in direction \( h \) defined by

\[
\sigma'_f(h) = \lim_{\delta \to 0^+} \sup_{x \in X(\delta)} \left( \lim_{t \to 0^+} \inf_{u \in U(x, \epsilon)} h(u, x) \right),
\]

This theorem may be of independent interest because of the prevalence of minimax problems in economics, although problems usually use convex-concave objective functions. It will be used in the proof of Theorem 2.3 below. There is a practical price to pay for this theoretical generality, which is that finding minimax (or maximin, in our case) points of a general function is much harder to do than in the special case of a convex-concave objective function, in which the set of saddle points is a product set. Compare the reassuring result of Shapiro (2008, Theorem 3.1) for the convex-concave case with, for example, the difficulties that arise in computation discussed in Lin, Jin, and Jordan (2019). In our simulations we resort to grid search conducted over a low-dimensional region and are aided by the imposition of null hypotheses that simplify computation.
2.5 Differentiability of uniform test statistics

Here we show that uniform test statistics applied to bound functions are Hadamard directionally differentiable despite the problem alluded to in subsection 2.3. The next two theorems are the main tools needed to ensure the asymptotic properties of test statistics and the availability of resampling using the procedure of Fang and Santos (2019).

The following theorem shows that under general conditions the supremum-norm maps are Hadamard directionally differentiable. We depart briefly from the notation for maximizers and $\epsilon$-maximizers defined above in (12) and (13) because for the supremum maps it is necessary to consider a larger variety of optimizer sets. For any $\epsilon \geq 0$ define the set of $\epsilon$-maximizers and $\epsilon$-minimizers of a function $f$ over $B \subseteq X$:

$S^+_f(B, \epsilon) = \left\{ x \in B : f(x) \geq \sup_{x \in B} f(x) - \epsilon \right\}$, $S^-_f(B, \epsilon) = \left\{ x \in B : f(x) \leq \inf_{x \in B} f(x) + \epsilon \right\}$. (14)

For functions of two arguments $(u, x)$, we define one more set of approximate optimizers, the set of $\delta$-maximizers of the function (in $x$) $\inf_{u \in A(x)} f(u, x)$:

$S_\eta(\delta) = \left\{ x \in X : \inf_{u \in A(x)} f(u, x) \geq \sup_{x \in X} \inf_{u \in A(x)} f(u, x) - \delta \right\}$. (15)

Theorem 2.3 verifies that the supremum-type maps $\lambda_1$ and $\lambda_2$ are Hadamard directionally differentiable maps of bounded functions to the real numbers. This result comes from transposition of maximum operators along with the result of Theorem 2.2, which appears in the derivative $\eta'_f$ defined in the statement of the theorem. The statistics simplify significantly when the function considered in the map is the zero function, and these cases are separated from the non-zero cases. They are useful when imposing null hypothesis restrictions, as will be seen in an example below.

**Theorem 2.3.** Suppose $U \subseteq \mathbb{R}^j$ and $X \subseteq \mathbb{R}^k$ and let $f, h \in \ell^\infty(U \times X)$. Suppose that $A : X \rightarrow U$ is upper hemicontinuous and non-empty-valued for all $x \in X$. For any $\epsilon, \delta \geq 0$ define the set $S^+_f(A(X), \epsilon)$, the set-valued map $S^-_{(-f)}(A(\cdot), \epsilon)$ and $S_\eta(\delta)$ as in (14) and (15). Define $\zeta(f) = \sup_{(u, x) \in A(X)} f(u, x)$ and $\eta(f) = \sup_{x \in X} \inf_{u \in A(x)} (-f(u, x))$ and

$$\zeta'_f(h) = \lim_{\epsilon \to 0^+} \sup_{(u, x) \in S^+_f(A(X), \epsilon)} h(u, x)$$

and

$$\eta'_f(h) = \lim_{\delta \to 0^+} \lim_{\epsilon \to 0^+} \inf_{u \in S^-_{(-f)}(A(x), \epsilon)} -h(u, x).$$

Then:

1. Suppose $\|f\|_\infty \neq 0$. Then $\lambda_1(f)$ is Hadamard directionally differentiable at $f$ with derivative
and suppose that \( \lambda \) and \( \sup \) focus on non-zero functions.

Corollary 2.4. Suppose that Theorem 2.3. When \( 2.2 \) of Cárdomo, Cuevas, and Rodríguez (2019) we also have the following corollary that is simpler than Theorem 2.3. When \( f \equiv 0 \), there is only a marginal change in the result, so this corollary focuses on non-zero functions.

2. Suppose \( \|f\|_\infty = 0 \). Then \( \lambda_1(f) \) is Hadamard directionally differentiable at \( f \) with derivative in direction \( h \)

\[
\lambda_1(f)(h) = \begin{cases} 
\zeta_f'(h), & \zeta(f) > \eta(f) \\
\max \{ \zeta_f'(h), \eta_f'(h) \}, & \zeta(f) = \eta(f) \\
\eta_f'(h), & \zeta(f) < \eta(f)
\end{cases}
\]

3. Suppose \( \|\{f\}_\pm \|_\infty \neq 0 \). Then \( \lambda_2(f) \) is Hadamard directionally differentiable at \( f \) with derivative in direction \( h \)

\[
\lambda_2(f)(h) = \begin{cases} 
\zeta_f'(h), & \zeta(f) > 0 \\
\left[ \zeta_f'(h) \right]_+, & \zeta(f) = 0 \\
0, & \zeta(f) < 0
\end{cases}
\]

4. Suppose \( \|\{f\}_\pm \|_\infty = 0 \). Then \( \lambda_2(f) \) is Hadamard directionally differentiable at \( f \) with derivative in direction \( h \)

\[
\lambda_2(f)(h) = \begin{cases} 
\left[ \zeta_f'(h) \right]_+, & \zeta(f) = 0 \\
0, & \zeta(f) < 0
\end{cases}
\]

When \( U \) and \( X \) are compact sets, \( (U \times X, \|\cdot\|) \) is a compact metric space. Then using Corollary 2.2 of Cárdomo, Cuevas, and Rodríguez (2019) we also have the following corollary that is simpler than Theorem 2.3. When \( f \equiv 0 \), there is only a marginal change in the result, so this corollary focuses on non-zero functions.

Corollary 2.4. Suppose that \( U \subset \mathbb{R}^j \) and \( X \subset \mathbb{R}^k \) are compact, \( A : X \rightrightarrows U \) be continuous, and suppose that \( \lambda_j(f) \neq 0 \) for \( j \in \{1, 2\} \). Letting \( \zeta(f) = \sup_{(u,x)\in A(X)} f(u,x) \) and \( \eta(f) = \sup_{x\in X} \inf_{u\in A(x)} (-f(u,x)) \). Then \( \zeta \) and \( \eta \) are Hadamard directionally differentiable tangentially to \( C(U \times X) \), and \( \zeta_f'(h) = \max_{(u,x)\in S^+(A(X),0)} h(u,x) \) and \( \eta_f'(h) = \max_{x\in S^-\eta(0)} \min_{u\in S^-(A(x),0)} -h(u,x) \) and \( \lambda_1 \) and \( \lambda_2 \) are Hadamard directionally differentiable tangentially to \( C(U \times X) \) with derivatives

\[
\lambda_1(f)(h) = \begin{cases} 
\zeta_f'(h), & \zeta(f) > \eta(f) \\
\max \{ \zeta_f'(h), \eta_f'(h) \}, & \zeta(f) = \eta(f) \\
\eta_f'(h), & \zeta(f) < \eta(f)
\end{cases}
\]

\[
\lambda_2(f)(h) = \begin{cases} 
\left[ \zeta_f'(h) \right]_+, & \zeta(f) = 0 \\
0, & \zeta(f) < 0
\end{cases}
\]

The next theorem shows that the \( L^p \) test functionals defined in (10) are Hadamard directionally
differentiable. The theorem includes two cases for each functional. The second case corresponds to \( \lambda_j(f) = 0 \) and is useful for describing the behavior the test functional under null hypotheses, which will be useful for inference results later. This is related to the results of Chen and Fang (2019) but different because we deal directly with \( L^p \) norm statistics — in that paper, the authors dealt with the squared \( L^2 \) statistic, which behaves differently. It is interesting to note here that taking the square root of the squared \( L^2 \) statistic results in first-order (directional) differentiability of the map for functions \( f \) with \( \|f\|_p = 0 \). This theorem makes use of the marginal maximizer operator \( \psi \) defined in (11). Although convergence to \( \psi'_f(h) \) may not be uniform, all that the dominated convergence theorem requires is pointwise convergence and integrability, here assumed to be \( p \)-integrability because an \( L^p \) statistic is used. This pointwise convergence is enough to provide the following result.

**Theorem 2.5.** Suppose that \( f, h \in \ell^\infty(U \times X) \) and both are \( p \)-integrable for \( p \geq 1 \). Then \( \lambda_3 \) and \( \lambda_4 \) defined in (10) are Hadamard directionally differentiable at \( f \) with derivative in direction \( h \) given by

\[
\lambda_3f(h) = \begin{cases} 
(\int_X |\psi(f)(x)|^p dx)^{(1-p)/p} \int_X \text{sgn}(\psi(f)(x))|\psi(f)(x)|^{p-1} \psi'_f(h)(x) dx, & \|\psi(f)\|_p \neq 0 \\
(\int_X |\psi'_f(h)(x)|^p dx)^{1/p}, & \|\psi(f)\|_p = 0
\end{cases}
\]

and

\[
\lambda_4f(h) = \begin{cases} 
(\int_X [\psi(f)(x)]^p dx)^{(1-p)/p} \int_X [\psi(f)(x)]^p - 1 \psi'_f(h)(x) dx, & \|\psi(f)\|_+^p \neq 0 \\
(\int_{\{x: \psi(f)(x) = 0\}} [\psi'_f(h)(x)]^p dx)^{1/p}, & \|\psi(f)\|_+^p = 0
\end{cases}
\]

Having established the directional differentiability of the statistics in (10), the next subsection derives the asymptotic statistical properties of tests used for uniform inference.

### 2.6 Limiting behavior of test statistics

Now we apply the analytic results of the last subsection, via the delta method, to the distribution of test statistics given sample data. We assume that we observe a sample of treatment and control observations. We make the following assumptions.

**A1** The observations \( \{Y_{0i}\}_{i=1}^{n_0} \) and \( \{Y_{1i}\}_{i=1}^{n_1} \) are iid samples and independent of each other and are distributed with marginal distribution functions \( F_0 \) and \( F_1 \) respectively. Refer to the pair of distribution functions and their empirical distribution function estimates as \( F = (F_0, F_1) \) and \( F_n = (F_{0n}, F_{1n}) \).

**A2** Let the sample sizes \( n_0 \) and \( n_1 \) increase in such a way that \( n_k/(n_0 + n_1) \to \nu_k \) as \( n_0, n_1 \to \infty \),
where $0 < \nu_k < 1$ for $k \in \{0, 1\}$. Define $n = n_0 + n_1$. When statements such as “as $n \to \infty$” are made below, it is understood that $n_0$ and $n_1$ grow in this balanced way.

These are fairly conventional assumptions in the literature on stochastic dominance comparisons and in papers on pointwise bounds. The independence assumption between samples could be weakened (Klecan, McFadden, and McFadden, 1991; Linton, Maasoumi, and Whang, 2005). This set of assumptions will need to be augmented by Assumption Q1 stated above for the case of quantile bounds, which additionally requires continuity of the densities of $F_0$ and $F_1$.

Under Assumptions A1 and A2, it is a standard result that for $k \in \{0, 1\}$, $\sqrt{n} (F_{kn} - F_k) \sim \mathcal{G}_k$, where $\mathcal{G}_0$ and $\mathcal{G}_1$ are independent $F_0$- and $F_1$-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

$$\sqrt{n} (\mathbb{F}_n - F) \sim \mathcal{G}_F = \begin{bmatrix} \mathcal{G}_0 / \sqrt{\nu_0} \\ \mathcal{G}_1 / \sqrt{\nu_1} \end{bmatrix},$$

where $\mathcal{G}_F$ is a mean-zero Gaussian process with covariance process $\rho_F(x, y) = \text{Diag}(\{\rho_k / \nu_k \}) (x, y)$.

For estimation of the quantile bounds, we require uniformly consistent estimators of the marginal quantile functions and density functions from each sample, a fairly weak condition. This is summarized in the following assumption, which is meant to accompany assumption Q1.

Q2 Assume that for $k \in \{0, 1\}$ there are estimators $\hat{F}_k^{-1}$ such that $\sup_{t \in T} |\hat{F}_k^{-1}(t) - F_k^{-1}(t)| = o_P(1)$ and also $\hat{f}_k$ such that $\sup_{x \in \mathcal{D}} |\hat{f}_k(x) - f_k(x)| = o_P(1)$, as $n \to \infty$. Refer to the pair of quantile functions and their estimates as $F^{-1} = (F_0^{-1}, F_1^{-1})$ and $\hat{F}^{-1} = (\hat{F}_0^{-1}, \hat{F}_1^{-1})$.

The existence of uniformly consistent quantile function estimates is all that is really required for the next theorem — a uniformly consistent density estimator will be required in the next section.

In what follows, the general bivariate functions $f(u, x)$ discussed in the previous subsections are specialized to the special case of functions of the form $f_1(u) - f_0(u - x)$ required for our uniform inference applications. The following result establishes an asymptotic distribution for all the test statistics considered above.

**Theorem 2.6.** 1. Under assumptions A1-A2, for $j \in \{1, \ldots, 4\}$,

$$\sqrt{n} (\Lambda_j(\mathbb{F}_n) - \Lambda_j(F)) \sim \mathcal{H}_j,$$

where

$$\mathcal{H}_j \sim \left( \lambda_j \Pi(F) \circ \Pi \right) (\mathcal{G}_F),$$

and $\mathcal{G}_F$ is the Gaussian process defined in (16).
2. Under assumptions A1, A2, Q1 and Q2, for \( j \in \{1, \ldots, 4\} \),
\[
\sqrt{n} \left( \Lambda_j(\hat{F}^{-1}) - \Lambda_j(F^{-1}) \right) \sim \mathcal{H}_j^{-1},
\]
where
\[
\mathcal{H}_j^{-1} \sim \left( \lambda_j'(F_n) \circ \Pi \circ Q'_{F} \right) (G_{F}).
\]

The distributions of these test statistics are complex — see Fan and Park (2010, 2012) for examples of the pointwise behavior of the bound functions. As such, they are not directly useful for inference. A suitable resampling scheme is the subject of the next section.

3 Inference

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. The resampling strategy we use was proposed by Fang and Santos (2019), and it combines a standard bootstrap procedure with estimates of the directional derivatives.

Once again, we focus on procedures that used to conduct inference for \( L \), and these would need to be modified slightly for \( U \), \( U^{-1} \) or \( L^{-1} \). The desired test statistic \( \Lambda_j(F_n) = (\lambda_j \circ \Pi)(F_n) \) (or \( \Lambda_j(F_n) = (\lambda_j \circ \Pi \circ Q)(F_n) \)) is constructed using sample data and one of the formulas in (9). Then the estimated reference distribution for \( \Lambda_j(F_n) \) is composed of resampled \( \Lambda_j'(\sqrt{n}(F_n^*-F_n)) = (\lambda_j' \circ \Pi)(\sqrt{n}(F_n^*-F_n)) \), where \( \sqrt{n}(F_n^*-F_n) \) is the recentered bootstrap empirical process. First we discuss the our choices for the required estimates, and then explicit resampling instructions will be given below.

The quantile transformation \( Q \) and the differencing/shifting operation \( \Pi \) are both straightforward to use with sample data. Under Assumption Q2, the standard derivative of the inverse map, which is \( Q'_f(h) = -(h/f) \circ F^{-1} \) (van der Vaart and Wellner, 1996, Lemma 3.9.23), can be consistently estimated:
\[
\hat{Q}'_n(h)(t) = \left( -\frac{h_1(\hat{F}_1^{-1})}{f_1(\hat{F}_1^{-1})}(t), -\frac{h_2(\hat{F}_2^{-1})}{f_2(\hat{F}_2^{-1})}(t) \right)
\]
for \( t \in T \). \( \Pi \) is linear, with
\[
\Pi_f'h(u,x) = \Pi(h)(u,x) = h_1(u) - h_0(u-x)
\]
for \( h \in (\ell^\infty(\mathbb{R}))^2 \), and therefore does not depend on \( f \) and needs no estimate.

All the rest of the estimators are similar, and similar to that used, for example, in Linton, Song, and Whang (2010), Chernozhukov, Lee, and Rosen (2013) and Lee, Song, and Whang (2018). We
use $c_n$ to denote a sequence of constants that decrease to zero slowly, meaning that $c_n \to 0^+$ and $\sqrt{n}c_n \to \infty$ as $n \to \infty$. We reuse the term $c_n$ in the definitions, so it could mean different sequences for different estimators. Estimates are written using the functions $f$ or $f_n$. For the case of the distribution function bound estimates, $f = \Pi(F)$, $f_n = \Pi(F_n)$ and $h = \Pi(\sqrt{n}(F_n^\ast - F_n))$, while for quantile function bounds the composition $(\Pi \circ Q)$ takes the place of the $\Pi$ map in the distribution function counterparts.

The supremum norm statistics require that one estimate the relationship between $\zeta(f)$ and $\eta(f)$ (these quantities were defined in Theorem 2.3), specifically to estimate which is larger or if they appear to be the same for $\lambda_1$, and whether $\zeta(f)$ is larger, equal to or smaller than zero for $\lambda_2$. For these derivatives we estimate such relationships using $\hat{\zeta}(f)$, specifically to estimate which is larger or if they must also estimate $\hat{\eta}(f)$. For these cases we compare $\hat{\zeta}(f)$ to zero and $\zeta(f_n)$ to $-c_n$. Inside these estimates there is a second layer of estimation, except in the $\lambda_1$ case with $f \equiv 0$. For these cases one must also estimate

$$\hat{\zeta}(h) = \inf_{x \in S_n(b_n) \cap S_{-f}(A(x), c_n)} h(u, x)$$

and

$$\hat{\eta}(h) = \sup_{x \in S_n(b_n) \cap S_{-f}(A(x), c_n)} -h(u, x),$$

where $b_n$ is another decreasing sequence and $S_{\eta}(\delta)$ was defined in (15). All the statistics in our examples are constructed so that the null hypothesis corresponds to $H_0 : \Lambda_j(F) \leq 0$ against the alternative $H_0 : \Lambda_j(F) > 0$. This simplifies the situation dramatically for $\lambda_1$, as will be seen in an example below. For $\lambda_2$ the situation is also simpler, although one must still estimate $\hat{I}(\zeta = 0) = I(|\zeta(f_n)| \leq c_n)$.

The behavior of the derivatives $\lambda'_{3f}$ and $\lambda'_{4f}$ depend on derivatives of $\psi$ and, in the case of $\lambda_4$, the region where $\sqrt{n}(L_n - L)$ is close to zero. Considering $\psi_n'$ first, recall that the supremum must be taken over one argument for each value of the other argument (luckily, for our applications only one grid is required because $\Pi(f(y, \delta)) = f_1(y) - f_0(y - \delta)$). We discretized the region and calculated maxima over the grid. For each value of in the grid $\{x_i\}$, we calculate the $c_n$-maximizers in $\{u_j\}$, i.e.,

$$S^+_f(A(x_i), c_n) = \left\{u_j \in A(x_i) : f_n(u_j, x_i) \geq \max_{u_j \in A(x_i)} f_n(u_j, x_i) - c_n \right\}$$

and then let

$$\psi_n'(h)(x_i) = \max_{u \in S^+_f(A(x_i), c_n)} h(u, x_i).$$

Meanwhile, $\Lambda_4$ depends on the contact set $X_0 = \{x \in X : f(x) = 0\}$. Linton, Song, and Whang (2010) and Andrews and Shi (2013) discuss uniformity over data generating processes; see also the discussion and example in Section 2.3 of Lee, Song, and Whang (2018). The contact set is so important because it is where $\Lambda_j(\sqrt{n}(F_n^\ast - F))$ converges to a nondegenerate limit distribution.

17
Heuristically, on the set \( \{ x : f(x) > 0 \} \), test statistics diverge and on \( \{ x : f(x) < 0 \} \) test statistics converge in probability to zero. See Linton, Maasoumi, and Whang (2005) for a more detailed discussion. Our contact set estimator for function \( f \) is
\[
\hat{X}_0 = \begin{cases} 
X & \text{if } \min_{x \in X} |f_n(x)| > c_n \\
\{ x \in X : |f_n(x)| \leq c_n \} & \text{otherwise.}
\end{cases}
\] (19)

Two alternatives currently exist in the literature that might be used to deal with the above estimates when discontinuities or kinks are involved. One approach exemplified by the generalized moment selection tests developed in Andrews and Shi (2017), involves recentering functions before using them in a test statistic. There is reason to believe that recentering may result in more powerful test procedures (Andrews and Shi, 2017). The other approach is to numerically evaluate the derivatives, a technique that was developed by Hong and Li (2016). Here we choose to estimate the required functions but future research may show that recentering or numerical derivatives work better for inference.

Given the estimates above, the estimated statistics are, for nonzero \( f \), \( \hat{\Lambda}'_{jn}(h) = (\hat{\lambda}'_{jn} \circ \Pi)(h) \). In the case that \( f \equiv 0 \), which is usually imposed under a null hypothesis, we have simpler forms that can be summarized explicitly (except for \( \hat{\psi}'_n \), which is defined in (18)):
\[
\begin{align*}
\hat{\Lambda}'_{1n}(h) &= \max_{(u,x) \in (U \times X)} |h(u,x)|, \\
\hat{\Lambda}'_{2n}(h) &= \left[ \max_{(u,x) \in S^+_n(A(X),b_n)} h(u,x) \right]_+ I(\zeta(f_n) > -c_n), \\
\hat{\Lambda}'_{3n}(h) &= \left( \int_X |\hat{\psi}'_n(h)(x)|^p \, dx \right)^{1/p}, \\
\hat{\Lambda}'_{4n}(h) &= \left( \int_{\hat{X}_0} \left[ \hat{\psi}'_n(h)(x) \right]^p \, dx \right)^{1/p}.
\end{align*}
\] (20)

The above steps can be summarized in the following routine.

**Resampling routine to estimate the distribution of \( \mathcal{H}_j \) (example for the lower bound \( L \))**

1. Estimate parts of \( \hat{\Lambda}'_{jn} \) if needed, as described above.

Then repeat steps 2-3 for \( r = 1, \ldots, R \):

2. Resample from \( \{ Y_{0i} \} \) and \( \{ Y_{1i} \} \) with replacement and construct resampled empirical CDFs \( \hat{F}_{0n}^* \) and \( \hat{F}_{1n}^* \).

3. Calculate the resampled test statistic \( \hat{\Lambda}^*_r = (\hat{\lambda}'_{jn} \circ \Pi)(\sqrt{n}(\hat{F}_{rn}^* - \hat{F}_n)) \).

Finally,
4. Use \( \{\Lambda^*_r\}_{r=1}^R \) to approximate the distribution of \( H_j \): 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 tests. Reject the null hypothesis if \( \Lambda_j(F_n) \) is larger than \( \hat{q}_{\Lambda^*}(1 - \alpha) \).

The consistency of this resampling procedure can be established using the fact that all the test functionals are Hadamard directionally differentiable and their estimators are sufficiently regular, which is summarized in the next theorem.

**Theorem 3.1.** Let \( BL_1(\mathbb{R}) \) denote the space of bounded Lipschitz functions with level and Lipschitz constant bounded by 1. Then for \( j \in \{1, \ldots, 4\} \),

1. Under assumptions \( A1 \) and \( A2 \), as \( n \to \infty \),

\[
\sup_{f \in BL_1(\mathbb{R})} \left| \mathbb{E} \left[ f \left( \hat{\Lambda}'_{jn} \left( \sqrt{n}(F_n^* - F_n) \right) \right) \right] \right| - \mathbb{E} \left[ f(H_j) \right] = o_P(1),
\]

where \( H_j \) were defined in part 1 of Theorem 2.6.

2. Under assumptions \( A1, A2, Q1 \) and \( Q2 \), as \( n \to \infty \),

\[
\sup_{f \in BL_1(\mathbb{R})} \left| \mathbb{E} \left[ f \left( \hat{\Lambda}'_{jn} \left( \sqrt{n}(F^{-1}_n - F^{-1}_0) \right) \right) \right] \right| - \mathbb{E} \left[ f(H_j^{-1}) \right] = o_P(1),
\]

where \( H_j^{-1} \) were defined in part 2 of Theorem 2.6.

It is also of interest to examine how these tests behave under sequences of distributions locally around distributions that satisfy the null hypothesis. Suppose that the hypothesis \( \Lambda_j(F) = 0 \) is true for \( j \in \{1, 2, 4\} \) — that is, excluding the two-sided \( L^p \) statistic. We consider sequences of distributions \( \{P_n\} \) of the form

\[
\int \left( \sqrt{n}((dF_n)^{1/2} - (dF_0)^{1/2}) - \frac{1}{2}h(dF_0)^{1/2} \right)^2 \to 0
\]

for alternative direction \( h \in L_2(F) \) such that \( \mathbb{E}[h] = 0 \). These sequences have the feature that the empirical process \( \sqrt{n}(F_n - F) \sim G_0 + \int_{(-\infty, \cdot)} h dF \) (Wellner, 1992, p. 257). Examining the form of the derivatives under the hypotheses that \( \Lambda_j(F) = 0 \), we find that for \( j = 1 \) or \( j = 2 \), convexity is straightforward to show. It can be verified that \( \psi'_j(\alpha g + (1 - \alpha)h)(x) \leq \alpha \psi'_j(g)(x) + (1 - \alpha) \psi'_j(h)(x) \). This implies that for \( \Lambda_4 \) an \( L^p \) norm is applied to a nonnegative function, in which case it is a nondecreasing, convex function applied to \( \psi'_j \). Therefore for these three statistics, under this null hypothesis, Corollary 3.2 of Fang and Santos (2019) applies. Therefore these tests control size uniformly: if \( \Lambda_j(F) = 0 \) then

\[
\limsup_{n \to \infty} P_n \left\{ \sqrt{n} \hat{\Lambda}_j(F_n) > \hat{q}_{\Lambda^*}(1 - \alpha) \right\} \leq \alpha.
\]
The above result applies to statistics that are constructed for the simple null hypothesis where the test statistic is equal to zero, because in that case the imposition of the null simplifies the operator into something that is convex. However, it is not apparent how to show that the size of the $\Lambda_2$ test is controlled because of possible negative values of $\psi'(G_F)(\delta)$.

In the next two subsections we illustrate the usefulness of our uniform results with two examples: the construction of uniform confidence bands around the bound functions and a test for stochastic dominance without any dependence information.

### 3.1 Example: uniform confidence bands

We use the lower bound as an example of confidence band construction. Under the hypothesis that the true lower bound $L$ is equal to the hypothesized bound $L_0$, we know that $\sup_\delta |L(\delta) - L_0(\delta)| = 0$, which suggests the statistic

$$\sqrt{n} \sup_\delta |L_n(\delta) - L_0(\delta)| \overset{H_0}{=} \sqrt{n} \left( \sup_\delta |L_n(\delta) - L(\delta)| - \sup_\delta |L_0(\delta) - L(\delta)| \right) = \sqrt{n}(\Lambda_1(F_n - F) - \Lambda_1(F_0 - F)).$$

This null hypothesis implies that for each $c$,

$$P \left\{ \sup_\delta \sqrt{n}|L_n(\delta) - L_0(\delta)| \leq c \right\} = P \left\{ \sqrt{n}|L_n(\delta) - L_0(\delta)| \leq c \text{ for all } \delta \right\} = P \left\{ \Lambda_n(\delta) - c/\sqrt{n} \leq L_0(\delta) \leq \Lambda_n(\delta) + c/\sqrt{n} \text{ for all } \delta \right\}.$$

The null of correct specification also implies that $\Lambda_{10}'$ does not require an estimate when used with the bootstrap. Theorem 3.1, specialized using the null hypothesis, implies that

$$P \left\{ \sqrt{n} \sup_{(y, \delta)} |\Pi(F^*_n - F_n)(y, \delta)| \leq c \left| \{Y_i\}_{i=1}^n \right\} \right\} \rightarrow P \{ \mathcal{H}_1 \leq c \}.$$

We can find a critical value $c_{1-\alpha}$ of this distribution using the bootstrap by estimating

$$c_{1-\alpha} = \min \left\{ c : P \left\{ \sqrt{n}\Lambda_{1n}'(F^*_n - F_n) \leq c \left| \{Y_i\}_{i=1}^n \right\} \geq 1 - \alpha \right\},$$

which amounts to simulating, for $r = 1, \ldots, R$,

$$\Lambda^*_r = \sqrt{n} \sup_{(y, \delta)} |F^*_n(y) - F^*_0n(y - \delta) - F_1n(y) + F_0n(y - \delta)|$$

and finding the $(1 - \alpha)^{th}$ quantile of the resulting bootstrap distribution $\{\Lambda^*_r\}_{r=1}^R$. Simulation evidence presented in the next section verifies that the coverage probability of such intervals is...
3.2 Example: stochastic dominance

Suppose that we wish to compare two outcome distributions using a stochastic dominance criterion, but we have no information on the dependence between potential outcomes between the distributions, labeled $A$ and $B$. If we can assume that treatments are perfectly observed, the methods represented by Linton, Song, and Whang (2010) allow us to test the stochastic dominance hypothesis $H_0 : A \succeq^{FOSD} B$ against the alternative that $A$ does not dominate $B$. If we suppose outcomes under control conditions and under treatments $A$ and $B$ are represented by random variables $(Y_0, Y_A, Y_B)$, the treatment effect distributions $\Delta_A = Y_A - Y_0$ and $\Delta_B = Y_B - Y_0$ cannot be point identified. However, we can construct a bound on dominance relations under which there are no possible dependence structures such that $A$ dominates $B$. Specifically, if $U_A(\delta) \leq L_B(\delta)$ for all $\delta$ then $A \succeq^{FOSD} B$ for all types of dependence between the random variables. If there exists some $\delta^*$ such that $U_A(\delta^*) > L_B(\delta^*)$ then there is a distribution of $(Y_0, Y_A, Y_B)$ such that $\Delta_A$ does not dominate $\Delta_B$. This represents not so much a direct test of the hypothesis that $A$ dominates $B$ as, using a phrase from Masten and Poirier (2017), a frontier at which of the notion that $A$ dominates $B$ breaks down.

To this end, suppose that the null hypothesis is $H_0 : U_A \leq L_B$ (uniformly in $\delta$). Under this null, the test statistic $\Lambda = (\int_{\mathcal{D}} [U_A(\delta) - L_B(\delta)]_+^2 d\delta)^{1/2}$ is equal to zero. Assume that we observe three independent samples of iid observations of size $n_0$, $n_A$ and $n_B$, let $n = n_0 + n_A + n_B$ and suppose that the sample sizes asymptotically satisfy $n_k/n \to \nu_k \in (0, 1)$ for $k \in \{0, A, B\}$. Given the observed data, we can calculate the statistic

$$\hat{\Lambda} = \sqrt{n} \left( \int_{\mathcal{D}} [U_{An}(\delta) - L_{Bn}(\delta)]_+^2 d\delta \right)^{1/2}$$

to test the hypothesis. For a sequence of constants $c_n \to 0^+$ such that $\sqrt{n}c_n \to \infty$ as $n \to \infty$, let $\mathcal{D}_0 = \{\delta \in \mathcal{D} : |U_{An}(\delta) - L_{Bn}(\delta)| \leq c_n\}$ and if it is empty is set equal to $\mathcal{D}$. Then the theory above can be extended in a straightforward way to the result that $\sqrt{n} \left( \hat{\Lambda} - \Lambda \right) \sim \mathcal{H}$ for a random variable similar to those discussed in Theorem 2.6. Defining, for $r = 1, \ldots, R$,

$$\Lambda^*_r = \sqrt{n} \left( \int_{\mathcal{D}_0} \left[ \psi_{An}^r \{\Pi(F_n^* - F_n)(y, \delta)\} - \psi_{Bn}^r \{\Pi(F_n^* - F_n)(y, \delta)\} \right]_+^2 d\delta \right)^{1/2},$$

a small extension of Theorem 3.1 to the case of two treatment effects implies that critical values based on resampling $\Lambda^*$ statistics results in a test with correct size at the breakdown frontier. This requires two estimates in addition to the estimate of $\mathcal{D}_0$, one for the derivative of $\psi_A(\Pi(F))\delta = \inf_y (F_A(y) - F_0(y - \delta))$ and one for the derivative of $\psi_B(\Pi(F))\delta = \sup_y (F_B(y) - F_0(y - \delta))$. A simulation in the next section illustrates the accurate size and power of this testing strategy.
4 Simulation experiments

First, some details about the computation of the bounds using sample data. 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, bounds computed from two samples could take unique values at all possible $Y_{1i} - Y_{0j}$ combinations, and computing the function on all $n_1 \times n_0$ points could be computationally prohibitive. Therefore, it is practical to compute the functions on a reasonably fine grid of $\{\delta_k\}_{k=1}^{K}$.

Define the support of a bound as the region where it is strictly inside the unit interval. 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. When computing both bounds on the same grid, a grid of points over $\{\min_i Y_{1i} - \max_j Y_{0j}, \max_i Y_{1i} - \min_j Y_{0j}\}$ is sufficient to capture the supports of both bounds. Take the lower bound as an example of how to calculate the support of one bound from two samples. The lower bound is the maximum of the difference empirical process $F_{1n}(\cdot) - F_{0n}(\cdot - \delta)$, which takes steps of size $+1/n_1$ at $Y_{1i}$ observations and steps of size $-1/n_0$ at $Y_{0j} + \delta$ shifted observations. It is relatively easy to find the upper endpoint of the support. This is the smallest value for which the lower bound is equal to 1. If for some (large) $\delta$, $Y_{0j} + \delta \geq Y_{1i}$ for all $i$ and $j$, then the shape of $F_{1n}(y) - F_{0n}(y - \delta)$ rises monotonically to one, then falls, as $y$ increases. The $\delta$ that satisfy this condition are $\delta \geq \max_{i,j}\{Y_{1i} - Y_{0j}\} = \max_i Y_{1i} - \min_j Y_{0j}$. The minimum of the support, which is the maximum value of $\delta$ such that $\mathbb{L}_n(\delta) = 0$, is more difficult to find. The function $F_{1n}(y) - F_{0n}(y - \delta)$ is always equal to zero for some $y$, but zero is the maximum only when $F_{n1}$ first-order stochastically dominates $F_{0n}(\cdot - \delta)$. Therefore, the smallest $\delta$ in the support of $\mathbb{L}_n$ is the smallest $\delta$ that shifts the $F_{0n}$ distribution function enough so that it is weakly dominated by $F_{n1}$. This is most easily found by computing the minimum vertical distance between quantile functions for both samples on a common set of quantile levels, that is, $\min_k\{\hat{Q}_{1n}(\tau_k) - \hat{Q}_{0n}(\tau_k)\}$. If an exact lower bound for the grid is not so important, one could also bound the grid using the looser $\min_{i,j}\{Y_{1i} - Y_{0j}\} = \min_i Y_{1i} - \max_j Y_{0j}$. Similarly, the support of the upper bound is between $\min_i Y_{1i} - \max_j Y_{0j}$ and $\max_k\{\hat{Q}_{1n}(\tau_k) - \hat{Q}_{0n}(\tau_k)\}$.

It may also be of interest to note that given two samples, one will always be able to reject the hypothesis of a constant treatment effect unless the distributions are perfect shift-translates of one another. If $\Delta = \delta^*$ with probability one, then the value $\delta^*$ should pass through the bounds for all quantile levels. However, the estimated upper bound $U_n(\delta) < 1$ for all $\delta$ less than $\max_k\{\hat{Q}_{1n}(\tau_k) - \hat{Q}_{0n}(\tau_k)\}$ (maximum taken over the common set of quantile levels), while the lower bound function is strictly greater than zero for all $\delta$ greater than $\min_k\{\hat{Q}_{1n}(\tau_k) - \hat{Q}_{0n}(\tau_k)\}$. This means the hypothesis of a constant $\Delta$ is not rejected only when the maximum and minimum quantile difference are equal to each other, that is, when one is a vertical translate of the other (or the CDFs are horizontal translates).
We illustrate the examples from the previous section with some simulation experiments. The simulated observations in the first experiment are normally distributed and we shift the location of one distribution to examine size and power properties. First, we used this normal location experiment to choose a sequence \( \{a_n\} \) used to approximate the set of \( \epsilon \)-maximizers in the estimation of the derivative \( \psi' \): we decided on the form \( a_n = \sqrt{c \log(\log(n))/n} \) (recall \( n = n_A + n_B \)) and chose \( c = 0.2 \) from simulations that examined size and power. For contact set estimation we chose the sequence \( b_n = 4 \log(\log(n))/\sqrt{n} \) based on the simulation results of Linton, Song, and Whang (2010), who concentrated on estimating the contact set in similar experiment.

Figure 4 verifies that our proposed method provides accurate coverage probability for uniform confidence bands and power against local alternatives. It shows the results of an experiment in which 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 \( L(\delta) = 2\Phi(\delta/2) - 1 \) for \( \delta > 0 \) and zero otherwise (Frank, Nelsen, and Schweizer, 1987, Section 4). The upper bound is symmetric so it is sufficient to examine only the lower bound test. 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 1,000 simulation repetitions. Alternatives are local location-shift alternatives with the mean of one distribution set to zero while the other ranges from \(-5/\sqrt{n}\) to \(5/\sqrt{n}\).

Although the power curves look very similar, it is important to remember that they represent power against these local alternatives and not fixed alternatives. Similarly, the low point on these curves (the point closest to the tests’ nominal size) does not occur at \(-1\), but at \(-1/\sqrt{n}\). As discussed at the end of Section 3, asymptotic size is controlled uniformly with this resampling procedure.

In a second experiment we simulated uniformly distributed samples and test for the condition that implies stochastic dominance of \( A \) over \( B \). 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 \( U_A - L_B \leq 0 \) as in the theoretical example of the last section. Uniform distributions are used because when \( \mu = 1 \), \( U_A - L_B \equiv 0 \) (Frank, Nelsen, and Schweizer, 1987, Section 4) so that value of \( \mu = 1 \) represents a sort of least-favorable null. 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 Figure 4. 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. Functions were evaluated on a grid with step size 0.01 and 1,000 simulation repetitions were used for each sample size.

The results of this test are similar to what one would expect from the stochastic dominance tests in Linton, Song, and Whang (2010), except that here the test is applied to bounds and represents
a breakdown frontier for the hypothesis that treatment $A$ dominates $B$, without any knowledge of how the treatments and control may be correlated. This shows how much information is lost by dropping the assumption of full treatment observability.

5 The treatment effect distribution of job training on wages

This section illustrates the usefulness of our proposed 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 Program (NSW), 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).
Figure 2: Power curves for the second (stochastic dominance) experiment. The horizontal axis shows local parameter values around the boundary of the null region where $\mu = 1$, which has been normalized to zero in this figure. Tests are $\Lambda_4$ tests with sample size $n \in \{100, 500, 1000\}$ and bootstrap repetitions $R \in \{499, 999, 1999\}$ respectively, using the bootstrap algorithm described in the text. 1,000 simulation repetitions were used to find empirical rejection probabilities.

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, 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.\(^1\) We 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. Also, we consider male workers only. Summary statistics for the two parts of the data are presented in Table 1.

To provide a more complete overview of the data, we also compute the empirical CDFs of the

\(^1\)The 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.
| Treatment Group | Control Group |
|-----------------|--------------|
| Mean            | Mean         |
| Median          | Median       |
| Min.            | Min.         |
| Max.            | Max.         |
| Earnings (1978) | 6,349.1      |
|                 | 4,232.3      |
|                 | 0.0          |
|                 | 60,307.9     |
| (7,867.4)       | (5,483.8)    |

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

treatment and control groups in Figure 5. 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, $F_{n,tr}(0) \approx 0.24$ and $F_{n,co}(0) \approx 0.35$. In this application we focus on the CDF of the treatment effects because of the apparent near-zero densities at the upper tail of the distributions, a condition that violates the regularity assumptions for inference on quantile functions (this could be avoided by making inferences over a restricted set of quantiles).

![Figure 3: Empirical CDFs of treatment and control observations in the experimental NSW data.](image)

There are 260 control group observations and 185 treatment group observations, and the treatment group outcomes stochastically dominate the control group outcomes at first order.

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 (5)–(6). 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 5 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 5 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 $\epsilon$ greater than zero but smaller than the next smallest outcome (about $45), F_{n, tr}(0) - F_{n, co}(-\epsilon) = 0.24$, which explains the jump in the lower bound estimate near zero (it really for a point in the grid just above zero). Likewise, for the same $\epsilon, F_{n, tr}(-\epsilon) - F_{n, co}(0) = -0.35$, 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 is the same as $1 - F_{n, co}(-\delta)$ and the lower bound to the right of zero is the same as $F_{n, tr}(\delta)$.

Taking the lower bound as an example, for each $\delta > 0$ find the closest observation from the control sample $y_{i, co}$, and set $y^*(\delta) = y_{i, co} + \epsilon$, leading to the supremum $F_{n, tr}(y_{i, co} + \epsilon)$ at every point where $y_{i, co} + \epsilon < y_{j, tr}$ 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 $1 - F_{n, tr}(0) + F_{n, co}(0)$ to 1 as $\delta$ goes from 0 to the right.

Figure 4: Bound functions and their uniform confidence bands. These confidence bands were constructed by inverting Kolmogorov-Smirnov-type test statistics as described in the text.
A first application of these bound functions is to construct a uniform confidence band for the true distribution function of treatment effects. This is shown in Figure 5. 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 $P$ is the collection of bivariate distributions that have marginal distributions $F_{tr}$ and $F_{co}$, then

$$\lim_{n \to \infty} \inf P \{ F_\Delta(\delta) \in CB(\delta) \text{ for all } \delta \} \geq 1 - \alpha.$$

This confidence band is likely conservative, since

$$P \{ \exists \delta : F_\Delta(\delta) \not\in CB(\delta) \}$$

$$= P \left\{ \{ \exists \delta : F_\Delta(\delta) < L_n(\delta) - c_{L,1-\alpha/2}/\sqrt{n} \} \cup \{ \exists \delta : F_\Delta(\delta) > U_n(\delta) + c_{U,1-\alpha/2}/\sqrt{n} \} \right\}$$

$$\leq P \left\{ \{ \exists \delta : L(\delta) < L_n(\delta) - c_{L,1-\alpha/2}/\sqrt{n} \} \cup \{ \exists \delta : U(\delta) > U_n(\delta) + c_{U,1-\alpha/2}/\sqrt{n} \} \right\}.$$ 

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. However, also note that the technique of Imbens and Manski (2004) cannot be used to tighten these functional bounds, because here the parameter could violate the null hypothesis at both sides of the confidence band.

There are other features plotted in this figure. 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 — however, see the discussion of constant treatment effects at the beginning of this section — 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 uniform across all possible joint distributions. 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 5: 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.

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 5. The results show that the uniform bands are farther from the bound estimates than the set of pointwise confidence intervals. However, because each confidence interval has $1 - \alpha$ coverage probability, we know that asymptotically, $\alpha$ percent of these intervals will not cover the true bounds (and we have no knowledge of which intervals do or do not cover the true bound). Moreover, the pointwise intervals are not always available at the extremes of the samples due to the lack of observations in these regions.
6 Conclusion

This paper develops new uniform statistical inference methods for investigating heterogeneous treatment effects in randomized experiments. We provide methods to conduct uniform inference for bounds of either the cumulative distribution function or quantile function of the treatment effects distribution. When the dependence between control and treatment outcomes is unspecified, it is known that at specific points in the treatment effect distribution, the distribution and quantile functions of effects are interval identified. The bounds on the distribution and quantile functions are nonlinear maps of observable marginal distribution functions of control and treatment outcomes. They are not differentiable (in a functional sense) but statistics utilized to conduct uniform inference using estimated bound functions are directionally differentiable. Despite this lack of full differentiability, we formally establish the asymptotic properties of nonparametric plug-in estimators of the statistics, which can be used for testing or inverted to construct uniform confidence bands. We also develop a detailed resampling technique to conduct practical inference for the bounds or for the true distribution or quantile function of the treatment effect distribution. Finally, to illustrate the methods, we provide an application to the evaluation of a job training program.
A Appendix

Proof of Theorem 2.2. It can be verified using $\inf f = \sup (-f)$ and the reverse triangle inequality 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 $s_n = t_n^{-1}$ we rewrite as $\sigma(s_n f + h) - s_n \sigma(f)$. First consider the inner optimization problem in $u$. For any $x \in X$ and $\epsilon > 0$ define $\tilde{h}(x, \epsilon) = \inf_{u \in U(x, \epsilon)} h(u, x)$. There exists a $u_\epsilon \in U(x, \epsilon)$ such that $h(u_\epsilon, x) \leq \tilde{h}(x, \epsilon) + \epsilon$ and $f(x, u_\epsilon) \leq \inf_{u \in A(x)} f(u, x) + \epsilon$ and therefore

$$\tilde{h}(x, \epsilon) \geq h(u_\epsilon, x) - \epsilon$$

$$= s_n f(u_\epsilon, x) + h(u_\epsilon, x) - s_n f(u_\epsilon, x) - \epsilon$$

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

$$\geq \inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) - (s_n + 1) \epsilon.$$

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

$$\inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) \leq \lim_{\epsilon \to 0^+} \inf_{u \in U(x, \epsilon)} h(u, x). \quad (21)$$

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

$$\inf_{u \in A(x')} (s_n f + h)(u, x') - s_n \sigma(f) = \left( \inf_{u \in A(x')} (s_n f + h)(u, x') - s_n \inf_{u \in A(x')} f(u, x') \right)$$

$$+ s_n \left( \inf_{u \in A(x')} f(u, x') - \sigma(f) \right)$$

$$\leq \lim_{\epsilon \to 0^+} \inf_{u \in U(x', \epsilon)} h(u, x') + s_n \left( \inf_{u \in A(x')} f(u, x') - \sigma(f) \right),$$

where the last inequality comes from (21). This means that for any $\delta > 0$, for any $x' \not\in X(\delta)$,

$$\inf_{u \in A(x')} (s_n f + h)(u, x') - s_n \sigma(f) \leq \lim_{\epsilon \to 0^+} \inf_{u \in U(x', \epsilon)} h(u, x') + s_n \left( \inf_{u \in A(x')} f(u, x') - \sigma(f) \right)$$

$$\leq \lim_{\epsilon \to 0^+} \inf_{u \in U(x', \epsilon)} h(u, x') - s_n \delta,$$
so that for any $\delta > 0$,
\[
\limsup_{n \to \infty} (\sigma(s_n f + h) - s_n \sigma(f)) = \limsup_{n \to \infty} \left( \sup_{x \in X(\delta)} \inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \sigma(f) \right)
\leq \sup_{x \in X(\delta)} \left( \liminf_{\epsilon \to 0^+} \inf_{u \in U(x, \epsilon)} h(u, x) \right)
\]
and therefore this inequality holds as $\delta \to 0$.

A lower bound for this expression can be obtained the same way, switching the inequalities in the inner and outer optimization problems. Start again with the inner problem. For any $x \in X$, choose an $\epsilon > 0$ and note that for any $u' \in A(x) \setminus U(x, \epsilon)$,
\[
s_n f(u', x) + h(u', x) - s_n \inf_{u \in A(x)} f(u, x) \geq \inf_{u \in A(x)} h(u, x) + s_n \epsilon.
\]
Therefore for any $x \in X$,
\[
\lim_{\epsilon \to 0} \inf_{u \in U(x, \epsilon)} h(u, x) \leq \liminf_{\epsilon \to 0} \inf_{u \in U(x, \epsilon)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x)
\]
and
\[
\liminf_{n \to \infty} \left( \liminf_{\epsilon \to 0} \inf_{u \in U(x, \epsilon)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) \right) = \liminf_{n \to \infty} \left( \inf_{u \in A(x)} (s_n f + h)(u, x) - s_n \inf_{u \in A(x)} f(u, x) \right). \quad (22)
\]
Consider again the outer maximization problem. For any $\delta > 0$, define
\[
\tilde{h}(\delta) = \sup_{x \in X(\delta)} \lim_{\epsilon \to 0} \inf_{u \in U(x, \epsilon)} h(u, x).
\]
For each $\delta$ there is an $x_\delta \in X(\delta)$ such that
\[
\lim_{\epsilon \to 0} \inf_{u \in U(x_\delta, \epsilon)} h(u, x_\delta) \geq \tilde{h}(\delta) - \delta
\]
and
\[
\lim_{\epsilon \to 0} \inf_{u \in U(x_\delta, \epsilon)} f(u, x_\delta) \geq \sigma(f) - \delta.
\]
Then for each \( n \),

\[
\tilde{h}(\delta) \leq \lim_{\epsilon \to 0^+} \inf_{u \in U(x_\delta, \epsilon)} h(u, x_\delta) + \delta
\]

\[
\leq s_n \lim_{\epsilon \to 0^+} \inf_{u \in U(x_\delta, \epsilon)} f(u, x_\delta) + \lim_{\epsilon \to 0^+} \inf_{u \in U(x_\delta, \epsilon)} h(u, x_\delta) - s_n \lim_{\epsilon \to 0^+} \inf_{u \in U(x_\delta, \epsilon)} f(u, x_\delta) + \delta
\]

\[
\leq \lim_{\epsilon \to 0^+} \inf_{u \in U(x_\delta, \epsilon)} (s_n f + h)(u, x_\delta) - s_n \lim_{\epsilon \to 0^+} \inf_{u \in U(x_\delta, \epsilon)} f(u, x_\delta) + \delta
\]

\[
= \left( \lim_{\epsilon \to 0^+} \inf_{u \in U(x_\delta, \epsilon)} (s_n f + h)(u, x_\delta) - s_n \inf_{u \in U(x_\delta, \epsilon)} f(u, x_\delta) - \sigma(f) \right) + \delta
\]

Then letting \( \delta \to 0 \) and using (22) we have

\[
\lim_{\delta \to 0} \tilde{h}(\delta) \leq \lim_{n \to \infty} \inf_{\gamma(x, y) = \max \{x, y\}} (s_n f + h) - s_n \sigma(f).
\]

\[\Box\]

**Proof of Theorem 2.3.** First, the map \( \gamma : \mathbb{R}^2 \to \mathbb{R} \) defined by \( \gamma(x, y) = \max \{x, y\} \) has directional derivative

\[
\gamma'_{x,y}(h_x, h_y) = \begin{cases} 
  h_x & x > y \\
  \max \{h_x, h_y\} & x = y \\
  h_y & x < y
\end{cases}
\]

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

\[
\frac{1}{t_n} \left( \sup_{x \in X} \sup_{u \in A(x)} (f(u, x) + t_n h_n(u, x)) \right) - \frac{1}{t_n} \left( \sup_{x \in X} \sup_{u \in A(x)} f(u, x) \right) = 
\]

\[
\frac{1}{t_n} \left( \sup_{(u, x) \in A(X)} \left( f(u, x) + t_n h_n(u, x) \right), \sup_{x \in X} \left( - \sup_{u \in A(x)} f(u, x) + t_n h_n(u, x) \right) \right) - 
\]

\[
- \frac{1}{t_n} \left( \sup_{(u, x) \in A(X)} f(u, x), \sup_{x \in X} \left( - \sup_{u \in A(x)} f(u, x) \right) \right) 
\]

\[
= \frac{1}{t_n} \left( \max \{\zeta(f + t_n h_n), \eta(f + t_n h_n)\} - \max \{\zeta(f), \eta(f)\} \right).
\]

It has been established (Cúrcamo, Cuevas, and Rodríguez, 2019) that \( \zeta \) has directional derivative

\[
\zeta'_f(h) = \lim_{\epsilon \to 0^+} \sup_{(u, x) \in S_f^+ (X, \epsilon)} h(u, x)
\]

33
Theorem 2.2 shows that, using \( S_\eta(\delta) \) defined in (15),
\[
\eta_f'(h) = \lim_{\delta \to 0^+} \sup_{\delta} \lim_{\epsilon \to 0^+} \inf_{S_{\epsilon f}(x,\epsilon)} -h(u, x).
\]

Then the chain rule \((\gamma(\zeta(f), \eta(f)))' = \gamma'_{(\zeta(f), \eta(f))}(\zeta_f'(h), \eta_f'(h))\) implies the first result in the statement. When \( \|f\|_\infty = 0 \), \( f \) is (essentially) zero and we can calculate directly
\[
\lim_{n \to \infty} \frac{\|f + t_n h\|_\infty - \|f\|_\infty}{t_n} = \|h\|_\infty.
\]

Next consider the functional \( \lambda_2 \), and suppose that \( \|f\|_+ \not= 0 \). Write \( \sup_x \max\{\sup_x f(u, x), 0\} = \max\{\sup_{(u,x)} f(u, x), 0\} \) and define \( \hat{\gamma} : \mathbb{R} \to \mathbb{R} \) by \( \hat{\gamma}(x) = \max\{x, 0\} \), which has the derivative \( \hat{\gamma}'(x) = h \) for \( x > 0 \), \( \hat{\gamma}'(x) = [h]_+ \) and \( \hat{\gamma}'(h) = 0 \) for \( x < 0 \). Then the result for \( \lambda_2(f) \), \( \|f\|_+ \not= 0 \) results from reversing the order of the maximum and supremum again in the final statement. When \( \|f\|_+ \not= 0 \), the derivative is only nonzero if \( \sup_{A(X)} f = 0 \), and similar calculations result in the final result. \( \blacksquare \)

Proof of Theorem 2.5. In this proof, the word “derivative” always refers to a Hadamard directional derivative. 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 together second.

Let \( f \) and \( h \) satisfy the assumptions in the statement of the theorem and choose \( t_n \to 0^+ \) and \( h_n \) such that \( \|h_n - h\|_\infty \to 0 \), implying that \( h_n \) are also \( p \)-integrable. By Lemma S.4.9 of Fang and Santos (2019), for each \( x \in X \), the derivative of \( \psi(f)(x) \) in direction \( h(x) \) is \( \psi_f'(h)(x) = \sup_{u \in S_f(x)} h(u, x) \). 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 \( x \not= 0 \) 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 \( x \mapsto x^c \) in direction \( c \) is \( \alpha cx^{\alpha - 1} \).

First assume that \( \|\psi(f)\|_p \not= 0 \). Then 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
\[
\left\{
\begin{array}{ll}
p\text{sgn}(\psi(f)(x))|\psi(f)(x)|^{p-1}\psi'_f(h)(x), & \psi(f)(x) \not= 0 \\
0, & \psi(f)(x) = 0
\end{array}
\right.
= p\text{sgn}(\psi(f)(x))|\psi(f)(x)|^{p-1}\psi'_f(h)(x).
\]

34
Similarly, when \(|\psi(f)|_p \neq 0\), 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 and dominated convergence imply that when \(\|\psi(f)\|_p \neq 0\),

\[
\lim_{n \to \infty} \frac{1}{t_n} \left( \int_X |\psi(f + t_n h_n)(x)|^p \, dx - \int_X |\psi(f)(x)|^p \, dx \right) = \int_X p\text{sgn}(\psi(f)(x))|\psi(f)(x)|^{p-1}\psi'_f(h)(x) \, dx
\]

and similarly, when \(|\psi(f)|_p \neq 0\),

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

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

\[
\lim_{n \to \infty} \frac{\lambda_3(f + t_n h_n) - \lambda_3(f)}{t_n} = \left( \int_X |\psi(f)(x)|^p \, dx \right)^{(1-p)/p} \int_X \text{sgn}(\psi(f)(x))|\psi(f)(x)|^{p-1}\psi'_f(h)(x) \, dx
\]

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

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

\[
\lim_{n \to \infty} \frac{1}{t_n} \left( \left( \int_X |\psi(f + t_n h_n)(x)|^p \, dx \right)^{1/p} - \left( \int_X |\psi(f)(x)|^p \, dx \right)^{1/p} \right) = \left( \int_X \sup_{u \in A(x)} h(u, x) \, dx \right)^{1/p}
\]

\[
= \|\psi(h)\|_p.
\]

The case is slightly different for \(\lambda_4\) because \(|\psi(f)|_p \neq 0\) only implies \(\psi(f)(x) \leq 0\) for almost all
Rule 2 in the above list of derivative rules implies that only the region \( \{ x \in X : \psi(f)(x) = 0 \} \) will contribute asymptotically to the inner integral, and in the limit, using \( p \)-integrability and dominated convergence we have the derivative \((\int_{X} \psi(f)(x) dx)^{1/p} = \lim_{n \to \infty} \left( \int_{X} \frac{[\psi(f + t_n h_n)(x)]_{+} - 0}{t_n} dx \right)^{1/p} \) as in the statement of the theorem.

**Proof of Theorem 2.6.** The limiting statements follow directly from the Hadamard directional differentiability established in Theorems 2.3 and 2.5 and the delta method, for example, Theorem 2.1 of Fang and Santos (2019).

**Proof of Theorem 3.1.** Theorems 2.3 and 2.5 show that the \( \Lambda_j \) are Hadamard directionally differentiable. These theorems, bootstrap consistency for the \( \Lambda_j \) functionals follows from the condition that the \( \Lambda_j \) satisfy Assumption 4 of Fang and Santos (2019), because this implies that Theorem 3.2 of Fang and Santos (2019) holds for the functionals. Below we verify that their Assumption 4 holds for all four functionals, using Lemma S.3.6 of Fang and Santos (2019), which is mainly to show that each functional is Lipschitz continuous with Lipschitz constant equal to one. We suppose that \( h_1, h_2 \in \ell^\infty(X) \), and also that they are \( p \)-integrable if an \( L^p \) statistic is used.

The standard functionals \( \hat{\Lambda}_{ln}'(h) = ||h||_\infty \) and \( \hat{\Lambda}_{3n}(h) = ||h||_p \) satisfy \( \hat{\Lambda}_{jn}(h_1 + h_2) \leq \hat{\Lambda}_{jn}(h_1) + \hat{\Lambda}_{jn}(h_2) \) by the subadditivity of the supremum for \( \hat{\Lambda}_{ln} \) and Minkowski’s inequality for \( \hat{\Lambda}_{3n} \). Then these two functionals satisfy the reverse triangle inequality as well using a standard argument. For \( \hat{\Lambda}_1 \) the above calculations are equivalent to \( ||h_1||_\infty - ||h_2||_\infty|| \leq ||h_1 - h_2||_\infty \). In the case of potentially nonzero \( f \), the derivative estimates depend on estimating the relationships between \( \zeta(f) \), \( \eta(f) \) and (potentially) 0, which are discussed in Lemma S.4.2 of Fang and Santos (2019). The estimated derivatives of \( \zeta \) and by extension \( \eta \) are dealt with in Lemma S.4.8 of Fang and Santos (2019). Then \( \hat{\Lambda}_1 \) satisfies Assumption 4 of Fang and Santos (2019) and the conditions of Lemma S.3.6 of Fang and Santos (2019) are satisfied.

Now consider the derivative \( \hat{\Lambda}_{2n} \). There is only one difference between this and \( \hat{\Lambda}_{1n} \) that needs to be addressed, which is the positive part function \([\cdot]_+\). Note that for any set \( X \), the inequality
\[
\max\{a + b, 0\} \leq \max\{a, 0\} + \max\{b, 0\}
\]
implies
\[
\sup_{x \in X} [h_1(x)]_+ = \sup_{x \in X} [h_1(x) - h_2(x) + h_2(x)]_+ \leq \sup_{x \in X} [h_1(x) - h_2(x)]_+ + \sup_{x \in X} [h_2(x)]_+
\]
and symmetrically with the labels switched. This implies that
\[
\sup_{x \in X} [h_1(x)]_+ - \sup_{x \in X} [h_2(x)]_+ \leq \sup_{x \in X} [h_1(x) - h_2(x)]_+ \leq \sup_{x \in X} |h_1(x) - h_2(x)|
\]
and
\[
-(\sup_{x \in X} [h_1(x)]_+ - \sup_{x \in X} [h_2(x)]_+) \leq \sup_{x \in X} [h_2(x) - h_1(x)]_+ \leq \sup_{x \in X} |h_1(x) - h_2(x)|
\]
so that
\[
\|[h_1]_+ - [h_2]_+\| \leq \|h_1 - h_2\|_{\infty}.
\]
Using this with the other parts discussed for \(\hat{\Lambda}'_{1n}\) implies the result for \(\hat{\Lambda}'_{1n}\).

For \(\hat{\Lambda}'_3\) and \(\hat{\Lambda}'_4\) Minkowski’s inequality can be used to show \(\|h_1\|_p - \|h_2\|_p\| \leq \|h_1 - h_2\|_p\). The estimator \(\hat{\psi}'_{n}\) requires that for each point one applies Lemma S.4.8 of Fang and Santos (2019). For the one-sided \(\hat{\Lambda}'_{4n}\), use the fact that
\[
\left\| \left( \int_{D_0} [h_1(x)]_+^p dx \right)^{1/p} - \left( \int_{D_0} [h_2(x)]_+^p dx \right)^{1/p} \right\| \leq \|[h_1]_+ - [h_2]_+\|_p
\]
and the inequality \(|\max\{a, 0\} - \max\{b, 0\}| \leq |a - b|\) implies
\[
\leq \|h_1 - h_2\|_p
\]
and that \(\hat{\Lambda}'_{4n}\) is also Lipschitz. Then \(\|\hat{\Lambda}'_{jn}(h) - \Lambda'_j(h)\| = o_P(1)\) can be shown by showing the consistency of the estimator \(\hat{D}_0\). This can be shown for the estimator defined in (19) for the sequence of constants defined there — see Linton, Song, and Whang (2010), for example. ■
References

ABBRING, J., AND 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., AND K. BORDER (2006): Infinite Dimensional Analysis: A Hitchhiker’s Guide. Springer-Verlag, Heidelberg.

ANDREWS, D., AND X. SHI (2013): “Inference Based on Conditional Moment Inequalities,” Econometrica, 81, 609–666.

— (2017): “Inference Based on Many Conditional Moment Inequalities,” Journal of Econometrics, 196, 275–287.

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNANDEZ-VAL (2006): “Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure,” Econometrica, 74, 539–563.

BICKEL, P., 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.

CÁRCAMO, J., A. CUEVAS, AND L.-A. RODRÍGUEZ (2019): “Directional Differentiability for Supremum-Type Functionals: Statistical Applications,” arXiv e-prints, p. arXiv:1902.01136.

CATTANEO, M., M. JANSSON, AND K. NAGASAWA (2017): “Bootstrap-Based Inference for Cube Root Consistent Estimators,” Unpublished working paper.

CHEN, Q., AND Z. FANG (2019): “Inference on Functionals under First Order Degeneracy,” Journal of Econometrics, forthcoming.

CHERNOZHUKOV, V., S. LEE, AND A. ROSEN (2013): “Intersection Bounds: Estimation and Inference,” Econometrica, 81, 667–737.

CHESHER, A., AND A. ROSEN (2015): “Counterfactual Worlds,” Cemmap working paper CWP22/15.

CHETVERIKOV, D., A. SANTOS, AND A. SHAIKH (2018): “The Econometrics of Shape Restrictions,” Annual Review of Economics, 10, 31–63.

CHO, J., AND H. WHITE (2018): “Directionally Differentiable Econometric Models,” Econometric Theory, 34, 1101–1131.

CHRISTENSEN, T., AND B. CONNAULT (2019): “Counterfactual Sensitivity and Robustness,” Unpublished working paper.
Danskin, J. (1967): *The Theory of Max-Min and Its Applications to Weapons Allocation Problems*. Springer-Verlag, New York.

Dehejia, R., and S. Wahba (1999): “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062.

Demyanov, V. (2009): “Minimax: Directional Differentiability,” in *Encyclopedia of Optimization*, ed. by C. Floudas, and P. Pardalos. Springer, Boston.

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. 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.

Frank, M., R. 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 Effects Model with Random Coefficients in the Selection Equation,” Unpublished working paper.

Heckman, 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. 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 (2016): “The Numerical Delta Method,” *Journal of Econometrics*, 206, 379–394.

Imbens, G. (2003): “Sensitivity to Exogeneity Assumptions in Program Evaluation,” *American Economic Review*, 93, 126–132.

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

Kim, J. (2014): “Identifying the Distribution of Treatment Effects under Support Restrictions,” Unpublished working paper.

Klecan, L., R. McFadden, and D. McFadden (1991): “A Robust Test for Stochastic Dominance,” Unpublished working paper.

LaLonde, R. (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.

Lin, T., C. Jin, and M. Jordan (2019): “On Gradient Descent Ascent for Nonconvex-Concave Minimax Problems,” *arXiv e-prints*, p. arXiv:1906.00331.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005): “Consistent Testing for Stochastic Dominance under General Sampling Schemes,” *Review of Economic Studies*, 72, 735–765.

Linton, O., K. Song, and Y.-J. Whang (2010): “An Improved Bootstrap Test of Stochastic Dominance,” *Journal of Econometrics*, 154, 186–202.

Makarov, G. (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., and A. Poirier (2017): “Inference on Breakdown Frontiers,” Unpublished working paper.
Milgrom, P., and I. Segal (2002): “Envelope Theorems for Arbitrary Choice Sets,” *Econometrica*, 70, 583–601.

Rockafellar, R., and R.-B. Wets (1998): *Variational Analysis*. Springer-Verlag, Berlin.

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.

——— (2008): “Asymptotics of Minimax Stochastic Programs,” *Statistics & Probability Letters*, 78, 150–157.

Smith, J., and P. 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.

Tamer, E. (2010): “Partial Identification in Econometrics,” *Annual Review of Economics*, 2, 167–195.

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

Wellner, J. (1992): “Empirical Processes in Action: A Review,” *International Statistical Review*, 60, 247–269.

Williamson, R., and T. Downs (1990): “Probabilistic Arithmetic I. Numerical Methods for Calculating Convolutions and Dependency Bounds,” *International Journal of Approximate Reasoning*, 4, 89–158.