Bounds in continuous instrumental variable models

Florian F Gunsilius*

MIT

December 1, 2019

Abstract

Partial identification approaches have seen a sharp increase in interest in econometrics due to improved flexibility and robustness compared to point-identification approaches. However, formidable computational requirements of existing approaches often offset these undeniable advantages—particularly in general instrumental variable models with continuous variables. This article introduces a computationally tractable method for estimating bounds on functionals of counterfactual distributions in continuous instrumental variable models. Its potential applications include randomized trials with imperfect compliance, the evaluation of social programs and, more generally, simultaneous equations models. The method does not require functional form restrictions *a priori*, but can incorporate parametric or nonparametric assumptions into the estimation process. It proceeds by solving an infinite dimensional program on the paths of a system of counterfactual stochastic processes in order to obtain the counterfactual bounds. A novel "sampling of paths"- approach provides the practical solution concept and probabilistic approximation guarantees. As a demonstration of its capabilities, the method provides informative nonparametric bounds on household expenditures under the sole assumption that expenditure is continuous, showing that partial identification approaches can yield informative bounds under minimal assumptions. Moreover, additional monotonicity assumptions lead to considerably tighter bounds, which constitutes a novel assessment of the identificatory strength of such nonparametric assumptions in a unified framework.

1 Introduction

In recent years, a trend in the literature in econometrics has been to obtain bounds on causal effects in general instrumental variable models (e.g. Chesher & Rosen 2017, Demuynck 2015, Kitagawa 2009, Manski 1990, Torgovitsky 2016). The arguments put forward in favor this partial identification approach are higher flexibility and robustness compared to point-identification approaches (Chesher & Rosen 2017). However, general and widely applicable partial identification approaches are often too complex for practical

*I want to thank Susanne Schennach, Toru Kitagawa, and Whitney Newey as well as Arie Beresteanu, Ken Chay, Tim Christensen, Adam McCloskey, Anna Mikusheva, Francesca Molinari, Nikolaus Schweizer, Eduardo Souza-Rodrigues, and Alex Torgovitsky for helpful discussions and feedback. I also thank the audiences at Boston College, Brandeis, Brown, Cornell, Georgetown, Harvard/MIT, Johns Hopkins, MIT, Michigan, Penn State, Rotterdam, Tilburg, UPenn, UPitt, the University of Amsterdam, the 3rd Econometrics workshop at Notre Dame, and the Conference on Counterfactuals with Economic Restrictions at Western for very helpful comments. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University. All errors are mine.

applications, which limits their broader use in general models, as noted in Beresteanu, Molchanov & Molinari (2012). Existing methods have therefore almost exclusively focused on the case of a binary treatment (e.g. Fan, Guerre & Zhu 2017, Lafférs 2015, Mogstad, Santos & Torgovitsky 2018) or have been intractable in practical settings, in particular when the endogenous variable has many points in its support. A practical method that provides a flexible approach for partial identification in general instrumental variable models with potentially continuous endogenous variables has so far been unavailable.

This paper introduces such a method for obtaining bounds on functionals of counterfactual distributions in instrumental variable models with arbitrary unobserved heterogeneity and potentially continuous variables. The method does not require structural assumptions per se, but allows to incorporate them rather intuitively into the model. This way, the researcher can include assumptions about particular individual behavior. The basic idea for this is to represent the instrumental variable model as a system of stochastic processes. The method then proceeds by solving an infinite dimensional program on the paths of these processes, which provides the bounds on the desired causal effects. Its potential applications include randomized trials with imperfect compliance, the evaluation of social programs and, more generally, simultaneous equations models. Further, it allows the researcher to only focus on a minimal model since it accounts for arbitrary forms of unobserved heterogeneity by default.

A novel "sampling of paths"-approach provides a convergent sequence of approximate solutions to these infinite dimensional programs. The idea for this approach is to sample a set of (shape-preserving) basis functions and to solve the (then semi-infinite dimensional) optimization problems on this sample. This introduced randomness is crucial, because it permits the derivation of probabilistic approximation guarantees of the unobserved heterogeneity by using large deviations results (Vapnik 1998, van der Vaart & Wellner 2013). In particular, a researcher can use these results to gauge how well the semi-infinite dimensional program approximates the infinite dimensional program for a given sample of basis functions. Furthermore, the sampling approach allows to determine the identificatory content of (nonparametric) shape restrictions in the model by sampling paths under these additional restrictions.

As a demonstration of its capabilities, the method estimates bounds on expenditure differences using the 1995/1996 UK family expenditure survey. This problem is well suited for demonstrating the method's capabilities in practice as it (i) is nonlinear with continuous variables, (ii) allows to gauge if the program actually obtains reasonable results[1], and (iii) provides a setting not directly related to causal inference, showing the scope of potential applications. In particular, the method provides the first informative nonparametric bounds on household expenditures under the sole assumptions that expenditure on goods is continuous with respect to the budget set, corroborating the nonparametric and semi-nonparametric approaches in Imbens & Newey (2009), Blundell, Chen & Kristensen (2007), de Nadai & Lewbel (2016), and Song (2018), which assume monotonicity or additive separability in the unobservables in the first- or second stage. This demonstrates that it is possible to obtain informative bounds on the causal effects of interest, even in general nonlinear problems and without introducing many functional form restrictions. Moreover, when including monotonicity assumptions in the model, the counterfactual bounds become more informative. This constitutes the first instance where the identificatory strength of monotonicity assumptions can be assessed in a unified manner. The practical estimation method is hence a first step to determine the identificatory strength

[1] In particular, the method should produce results which show that food is a necessity- and leisure is a luxury good, as these are well-established economic facts. Anything else would indicate the method's insufficiency. This application is actually more challenging than a Monte-Carlo approach, as the method needs to replicate known facts on real data under minimal assumptions (Advani, Kitagawa & Słoczyński 2019). A priori, it is not even clear that *any* approach can deliver informative enough bounds to check its validity. The fact that this method does provide informative bounds is a testament to its potential usefulness.

of frequently made (non-) parametric assumptions on the model.

The focus in this article lies on estimation. Still, large sample results are derived, for each bound separately. They prove directional Hadamard differentiability (Bonnans & Shapiro 2013, Shapiro 1991) of the value functions which correspond to the counterfactual bounds, and provide the respective large sample-distributions. This also makes recently developed bootstrap methodologies (Dümbgen 1993, Fang & Santos 2018, Hong & Li 2018) applicable in this setting.[2] Based on these large sample results, one can obtain coverage results of the identified set by relying on established results from the literature on partial identification such as Imbens & Manski (2004) and Stoye (2009).[3] In combination with these inference results, the proposed method therefore provides a practical approach for estimation and inference in general instrumental variable models.

2 Setup

This article deals with general instrumental variable models with no *a priori* assumptions on the functional relations between the variables and unrestricted unobserved heterogeneity. These instrumental variable models have many (almost) equivalent representations (Imbens 2019, Pearl 1995*b*), for instance as nonseparable triangular models of the form

$$\begin{aligned} Y &= h(X, W) \\ X &= g(Z, W) \qquad Z \perp\!\!\!\perp W, \end{aligned} \tag{1}$$

where, Y is the outcome variable, X is the endogenous treatment in the sense that it depends on the unobservable variable W, and Z is the instrument.[4] Throughout, Y, X, and Z lie in the unit interval $[0, 1]$, while W is infinite dimensional.[5] Alternative ways to represent the instrumental variable model are directed acyclic graphs (Pearl 1995*a*) and the counterfactual notation $P_{Y(x)}$ and $P_{X(z)}$ (Rubin 1974). The latter representation is rather useful in the current setting.

In fact, the main idea for the approach is to interpret the counterfactual laws $P_{Y(x)}$ and $P_{X(z)}$ of the instrumental variable model as laws of *counterfactual stochastic processes* Y_x and X_z, where the randomness is induced by the unobservable W. Each element $w \in \mathcal{W}$ indexes one path $Y_x(w)$ of the process. This representation allows to optimize over the unobserved heterogeneity W directly instead of the laws $P_{Y(x)}$ and $P_{X(z)}$ induced by it, which is the key for making the approach feasible.

It has been shown in Gunsilius (2019) that model (1) without structural assumptions is too general in the case when X is continuous: it can replicate any data generating process for which $P_{X|Z=z}$ is absolutely continuous with respect to Lebesgue measure. This directly implies that without restricting the functional forms h and g, any approach for estimating sharp bounds will give trivial bounds. In particular, completely unrestricted functions h and g correspond to a set of stochastic processes defined in $[0, 1]^{[0,1]}$, i.e. all possible

[2] One challenge is the computational burden of resampling methods in this setting, which can be substantial for complex problems. The derived large sample distributions can potentially be helpful in such cases.

[3] The results for inference and estimation of Chernozhukov, Lee & Rosen (2013) do not seem to be applicable in this setting as they deal with optimization problems with a Euclidean parameter space. In contrast, the proposed optimization method is defined over an infinite dimensional path space.

[4] $Z \perp\!\!\!\perp W$ means that Z is independent of W, i.e. $P_{Z,W} = P_Z P_W$. This model is often written with two separate unobservable variables V and U in the second- and first stage (e.g. Imbens & Newey 2009). This is an equivalent model to (1) as one can define the two dependent variables U and V on the same probability space and combine them to one variable W.

[5] The requirement of all random variables to the unit interval is a normalization of the observable random variables is made due to practical considerations and for theoretical convenience. Theoretically, one can extend the results by allowing for multivariate Y, X, and Z on compact sets, in which case the counterfactual processes Y_X and X_Z would actually be random fields. In practice, the variables are currently required to be univariate, however, as the practical solution approach is subject to a curse of dimensionality as no sparsity or factor-model assumptions are placed on the stochastic processes.

paths mapping the unit interval to the unit interval, equipped with the standard Borel σ-algebra. From a mathematical perspective, $[0,1]^{[0,1]}$ is too big and the Borel σ-algebra is too coarse in the sense that many subspaces of interest like $C([0,1])$, the space of continuous functions on $[0,1]$, are not even measurable in the Borel σ-algebra on $[0,1]^{[0,1]}$ (Bauer 1996, Corollary 38.5).

It is here where the stochastic process representation is useful for the first time, as it allows to introduce very weak but easy to understand assumptions which restrict the unobservable heterogeneity W. In fact, in order to obtain a theoretical and practical useful method, one needs to restrict the space of all possible paths. A rather weak but restrictive enough assumption for this is to only consider stochastic processes that possess almost surely *càdlàg* paths, i.e. paths that are continuous on the right with limits on the left. The standard space for these paths is the Skorokhod space $D([0,1])$ under the Skorokhod metric

$$d_S(Y_x, Y'_x) \coloneqq \inf_{\lambda \in \Lambda} \left\{ \max\{ \|\lambda - I\|_\infty, \|Y_x - \lambda Y'_x\|_\infty \} \right\},$$

where Λ is the set of all strictly increasing and continuous functions $\lambda : [0,1] \to [0,1]$, $I : [0,1] \to [0,1]$ is the identity function on $[0,1]$, and $\|\cdot\|_\infty$ is the supremum norm (Billingsley 1999, Section 12). One sufficient and necessary condition for a function $f : \mathcal{X} \to \mathbb{R}$ to lie in the Skorokhod space is that its extended modulus of continuity $\omega'_f(\delta)$ satisfies $\omega'_f(\delta) \to 0$ for any $\delta \downarrow 0$ (Billingsley 1999, p. 123).[6] The following assumption holds throughout.

Assumption 1 (Skorokhod space)**.** *Y, X, and Z take values in $[0,1]$. Moreover, for every $\delta > 0$ with $\delta \to 0$ it holds that*

$$\lim_{\delta \to 0} \sup_{w \in \mathcal{W}} \omega'_{Y_x(w)}(\delta) = 0 \qquad \textit{and} \qquad \lim_{\delta \to 0} \sup_{w \in \mathcal{W}} \omega'_{X_z(w)}(\delta) = 0.$$

Assumption 1 only allows paths $Y_x(w)$ and $X_z(w)$ that have finitely many discontinuities which exceed any given real number. This assumption is weak enough in all practical settings. For instance in the setting of randomized controlled trials with imperfect compliance, where Z is the initially randomly assigned dose of the treatment and X is the actual dose taken, each path $X_z(w)$ defines one response profile of a hypothetical participant in this trial, for instance the never taker has a path which is zero for every dose z. Then Assumption 1 restricts the model to response profiles which jump finitely often with the initially assigned dose—this allows to model drop-outs for instance, which are paths $X_z(w)$ which drop to zero at a certain $z \in [0,1]$ and stay zero for all $z' \geq z$. $\mathcal{W}$ denotes the support of W.

Under this assumption, one can formally introduce the stochastic process representation of model (1). Throughout the article, these stochastic processes are defined for "deterministic time", even though X and Z are random variables. This is achieved by assuming that X and Z have a fixed (i.e. deterministic) support in $[0,1]$. This is without loss of generality in this model; in fact, all of the results in this article go through almost verbatim under the assumption that X and Z have *random* support in $[0,1]$. The only difference will be the slight formal complication of working with stochastic processes in random time.[7]

Proposition 1 (Stochastic process representation)**.** *For fixed supports $\mathcal{X}$ and $\mathcal{Z}$ in $[0,1]$, model (1) is equivalent to a system of counterfactual stochastic processes Y_x and X_z on $[0,1]^{[0,1]}$ equipped with the Borel σ-algebra on the cylinder sets and with corresponding laws $P_{Y(x)}$ and $P_{X(z)}$. These processes are measurable*

[6] $\omega'_f(\delta) \coloneqq \inf_{\{x_i\}} \max_{i \in \mathbb{N}} \sup_{x, x' \in [x_{i-1}, x_i)} |f(x) - f(x')|$, where the infimum extends over all δ-partitions $\{x_i\}_{i \in \mathbb{N}}$, i.e. partitions where the values x_i are at a distance of at least $\delta > 0$ from each other, see Billingsley (1999, p. 122). This definition extends the modulus of continuity for continuous functions to càdlàg functions.

[7] In this case, one would need to replace Kolmogorov's extension theorem by the random version provided in Theorem 1 of Hu (1988).

in $D([0,1])$ under Assumption 1. The exclusion restriction implies that (i) $X_z(w)$ is a stopping time for the process $Y_{X_z(w)}(w)$ and (ii) the laws $P_{Y|X=x}$ and $P_{X|Z=z}$ induce a joint law

$$P_{[Y,X]^*(z)}(A_y, A_x) = \int_{A_x} P_{Y(x)}(A_y) P_{X(z)}(dx), \qquad A_y \in \mathscr{B}_{\mathcal{Y}}, A_x \in \mathscr{B}_{\mathcal{X}},$$

which corresponds to a joint counterfactual process $[Y,X]^*_z$ in $([0,1]\times[0,1])^{[0,1]}$.[8] *Furthermore, $[Y,X]^*_z$ can be defined on $D([0,1])$ with codomain $[0,1]^2$ and is measurable with respect to the filtration $\mathscr{F}_z^{[Y,X]}$ of all events of $[Y,X]^*_z$ that have happened before $z \in \mathcal{Z} \subset [0,1]$. The independence restriction $Z \perp\!\!\!\perp W$ allows to compare the counterfactual process $[Y,X]^*_z$ to the stochastic process $[Y,X]_z$ corresponding to the observable joint law $P_{Y,X|Z=z}$.*

Proposition 1 states that the counterfactual probability measures $P_{Y(x)}$ and $P_{X(z)}$ are the laws of Y_x and X_z, where the randomness is captured by some general probability space $(\Omega, \mathcal{B}_\Omega, P)$. In model (1), this randomness stems from the unobservable confounder W. Since W is unobserved, one can identify the abstract probability space $(\Omega, \mathcal{B}_\Omega, P)$ with the probability space $([0,1], \mathscr{B}_{[0,1]}, P_W)$, i.e. to identify the support $\mathcal{W}$ of W with the unit interval.[9] Then the elements $w \in \mathcal{W}$ index the paths $Y_x(w)$ and $X_z(w)$ of the respective stochastic process.

This representation allows one to set up an infinite dimensional program for obtaining bounds in these general instrumental variable models. The idea for this is standard in the counterfactual partial identification literature[10]: one maximizes (for an upper bound) or minimizes (for a lower bound) some functional of the counterfactual law $P_{Y(x)}$ under the restriction that the counterfactual laws $P_{Y(x)}$ and $P_{X(z)}$ together replicate the observable distribution $F_{Y,X|Z=z}$. The innovation of this article is that by considering $P_{Y(x)}$ and $P_{X(z)}$ as the laws of corresponding counterfactual processes Y_x and X_z, one can make this approach tractable in practice even in the setting where all variables are continuous.

Proposition 1 implies that it is equivalent to find an optimal measure P^*_W on all paths of the processes $Y_x(w)$ and $X_z(w)$ to obtaining an optimal combination of the laws $P_{Y(x)}$ and $P_{X(z)}$. Therefore, instead of optimizing over combinations of laws $P_{Y(x)}$ and $P_{X(z)}$, one optimizes over the index set $\mathcal{W} \subset [0,1]$ of the corresponding stochastic processes. Intuitively, optimizing over the unobserved heterogeneity directly permits the approximation of the structure of the unobservable heterogeneity by sampling elements $w \in \mathcal{W} \subset [0,1]$.

The method works for general linear and nonlinear operators $K : (\mathcal{W}, \mathscr{B}_{\mathcal{W}}) \to (\mathcal{Y}\times\mathcal{X}, \mathscr{B}_{\mathcal{Y}} \otimes \mathscr{B}_{\mathcal{X}})$. In many settings, one can write the objective function as a linear operator of the form[11]

$$\begin{aligned} KP_W(A_y, A_x) &:= E_{P_W}[K(Y_x, A_y, A_x)] \\ &= \int K(Y_x(w), A_y, A_x) P_W(dw), \qquad A_y \in \mathscr{B}_{\mathcal{Y}}, A_x \in \mathscr{B}_{\mathcal{X}}. \end{aligned} \tag{2}$$

For fixed A_x, A_y this operator reduces to a functional, a one-dimensional quantity, which is currently required

[8] $\mathscr{B}_{\mathcal{Y}}$ denotes the Borel σ-algebra of all open subsets of $\mathcal{Y}$.

[9] This follows from standard isomorphism results, see for instance Bogachev (2007, Theorem 9.2.2). It is in fact possible to construct these explicit isomorphisms between general Borel measures on Polish spaces and Lebesgue measure on the unit interval based on the approach laid out in Kuratowski (1934). One standard application of these isomorphisms is the construction of Wiener measure, see for instance Hess (1982). The difference to other existing identification results in the literature is that W is always infinite dimensional in this article: only the cardinality of $\mathcal{W}$ is important, which needs to be such that each path $Y_x(w)$ and $X_z(w)$ in the Skorokhod space—a Polish space under the Skorokhod metric—can be indexed by one w. By the above isomorphism results, this is already possible if $\mathcal{W} \subset [0,1]$.

[10] See the discussion in section 6 for a brief overview of applications of linear programming in econometrics.

[11] Throughout, the integral operator and its kernel are denoted by the same letter K.

by the method.[12] Sharp upper and lower bounds on the functional $KP_W(A_y, A_x)$ can then be obtained as solutions to the following infinite dimensional linear programs:

$$\begin{aligned} &\textit{minimize/maximize}_{P_W \in \mathscr{P}^*(\mathcal{W})} \quad KP_W(A_y, A_x) \\ &\text{s.t.} \;\; F_{Y,X|Z=z}(y,x) = P_W(Y_x \leq y, X_z \leq x) \end{aligned} \tag{3}$$

for all $(y, x, z) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ and fixed A_y, A_x. $F_{Y,X|Z=z}$ is the distribution function corresponding to the observable law $P_{Y,X|Z=z}$.

$\mathscr{P}^*(\mathcal{W})$ denotes the set of all measures P_W on $(\mathcal{W}, \mathscr{B}_{\mathcal{W}})$ which induce paths on the Skorokhod space defined in Assumption 1. All additional functional form assumptions on the model will be made by shrinking $\mathscr{P}^*(\mathcal{W})$. For instance, making continuity assumptions between Y and X or X and Z in (1) translates to continuity assumptions on the paths Y_x or X_z. In this case, $\mathscr{P}^*(\mathcal{W})$ contains only measures P_W on $\mathcal{W}$ which put positive probability on continuous paths.[13]

3 Method

This section introduces a probabilistic approach to solve the problems (3) approximately in practice with probabilistic finite sample guarantees for the validity of the approximation.[14] The method proceeds by sampling bases of the path space and solving semi-infinite dimensional analogues of (3) over this sample of basis functions. The randomness introduced by sampling the paths permits the use of standard large deviation results (Vapnik 1998, Chapter 5) which provides probabilistic guarantees for the approximation of (3) via a random sample.

The idea is to approximate the infinite dimensional problems (3) by their semi-infinite analogues

$$\begin{aligned} &\textit{min/max}_{\hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W})} \quad \frac{1}{l}\sum_{i=1}^{l} K(\tilde{Y}_x^{\kappa}(i), A_y, A_x)\frac{d\hat{P}_W}{d\hat{P}_0}(i) \\ &\text{s.t.} \quad \left\| F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^{\kappa}(i), \tilde{X}_z^{\kappa}(i))\frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^2)} \leq \varepsilon \end{aligned} \tag{4}$$

for some small $\varepsilon > 0$ and all $z \in [0,1]$, and where the L^2-norm is taken with respect to y, x.[15] Underlying the approximations (4) is a sample of size l of a number of κ basis functions which approximate the path

[12] Thus, quantile effects cannot be established directly through the form of the objective functions, as they require knowledge of the complete counterfactual distribution. It is, however, possible to approximate bounds on the counterfactual quantile functions by solving the optimization problems (3) for many different values y and a fixed value x_0. This will give an approximation of the upper- (for maximization) and lower- (for minimization) "envelopes" of the counterfactual distribution $F_{Y(x_0)}(y)$. Respective bounds on the quantile distributions can then be obtained by inverting these envelopes.

[13] Since under Proposition 1 all processes are progressively measurable with respect to their natural filtration, one can even introduce dynamic assumptions this way, like mixing properties.

[14] One can in fact only approximate the solutions to (3) (Kappen 2007, p. 152) since one cannot solve the problem over all possible paths on a space.

[15] Rewriting the programs (4) in their Lagrangian form later on will reveal that ε fulfills the same purpose as a penalty term/Lagrange multiplier for the constraint. In this sense, ε is a penalty parameter of the program which needs to be chosen appropriately. See the next section for a discussion. Overall, the relaxation of the linear constraints to an L^2-constraint is made exclusively based on finite sample considerations. In the population $\varepsilon = 0$.

space of the processes Y_x and X_z. These approximations are denoted by

$$\tilde{Y}_x^\kappa(i) \coloneqq \sum_{j=1}^{\kappa} \beta_j(i) b_j(x) \qquad \text{and} \qquad \tilde{X}_z^\kappa \coloneqq \sum_{j=1}^{\kappa} \alpha_j(i) a_j(z) \tag{5}$$

for coefficients β, α and basis functions $b(x), a(z)$. Of particular convenience are basis functions used in Sieves estimation such as (trigonometric) polynomials, splines, wavelets, and neural networks (Chen 2007). This articles works with a shape-preserving wavelet basis, which is defined as

$$\varphi_{\kappa j}(x) \coloneqq 2^{\frac{\kappa}{2}} \varphi(2^\kappa x - j) \qquad \kappa, j \in \mathbb{Z}$$

with $\varphi : \mathbb{R} \to [0,1]$ of the form

$$\varphi(x) \coloneqq \begin{cases} x+1 & \text{if } -1 \leq x \leq 0 \\ 1-x & \text{if } 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$

Based on this the notation for the paths sampled via this wavelet basis for dilation κ is

$$\tilde{Y}_x^\kappa(i) \coloneqq \sum_{j=-\infty}^{\infty} \alpha_j(i) \varphi_{\kappa j}(x) \quad \text{and} \quad \tilde{X}_z^\kappa(i) \coloneqq \sum_{j=-\infty}^{\infty} \beta_j(i) \varphi_{\kappa j}(z), \tag{6}$$

where the sums in the definition are both finite since the unit interval is bounded. This wavelet basis preserves shapes in the sense that an approximation of a monotone (convex) function via this basis will itself be monotone (convex), see Anastassiou & Yu (1992*b*) and Anastassiou & Yu (1992*a*); this is an important feature when introducing shape assumptions into the model in practice.

The dependence of $\tilde{Y}_x(i)$ and $\tilde{X}_z(i)$ on the index i shows that the problems (4) are indeed semi-infinite dimensional. In particular, the index i now runs over finitely many elements l and replaces the variable $w \in \mathcal{W}$ on the state space of P_W. In this respect, the term $\frac{d\hat{P}_W}{d\hat{P}_0}(i)$ is fundamental. In particular, the *empirical sampling law* $\hat{P}_0 \in \hat{\mathscr{P}}^*(\mathcal{W})$ is one *representative* law on the paths of stochastic processes, which is used for sampling the basis functions. The optimization is then over all $\hat{P}_W$ which are *absolutely continuous* with respect to $\hat{P}_0$, so that $\frac{d\hat{P}_W}{d\hat{P}_0}(i)$ is the Radon-Nikodym derivative. This construction arises naturally, as the empirical sampling law $\hat{P}_0$ determines the universe of all l paths, over which an optimal law $\hat{P}_W$ will be chosen to solve the programs. $\hat{P}_W$ must by construction be absolutely continuous with respect to $\hat{P}_0$ as it can only place positive mass on the i sampled paths, which have been determined via $\hat{P}_0$. In other words, $\hat{\mathscr{P}}^*(\mathcal{W})$ is the set of all probability measures which do not put positive measure on paths other than the l paths sampled via $\hat{P}_0$. This implies a natural assumption on the data-generating process in the population.

Assumption 2 (Representer P_0 of $\mathscr{P}^*(\mathcal{W})$)**.** *The sampling law P_0 is a representative law of $\mathcal{P}^*(\mathcal{W})$ in the sense that (i) $P_0 \in \mathcal{P}^*(\mathcal{W})$ and (ii) every $P_W \in \mathcal{P}^*(\mathcal{W})$ is absolutely continuous with respect to P_0 with β-Hölder continuous ($\beta > \frac{1}{2}$) Radon-Nikodym derivative satisfying $\sup_{w \in \mathcal{W}} \frac{dP_W}{dP_0}(w) \leq C_{RN} < +\infty$ for a fixed constant C_{RN}.*[16]

Assumption 2 is the theoretical analogue to the fact that all $\hat{P}_W$ are absolutely continuous with respect to $\hat{P}_0$ by construction. Since by the finite dimensional construction all measures $\hat{P}_W$ are automatically

[16] A function $f : \mathcal{X} \to \mathbb{R}$ is β-Hölder continuous if $\sup_{x \neq x' \in \mathcal{X}} \frac{|f(x) - f(x')|}{\|x - x'\|^\beta} < +\infty$ (Folland 2013, p. 138).

absolutely continuous with respect to $\hat{P}_0$, theoretical measures P_W which are not absolutely continuous to P_0 can never be detected. This implies that Assumption 2 is non-testable. It is an assumption on $\mathscr{P}^*(\mathcal{W})$ which necessarily follows from the approximation argument. Below, we will argue that it is similar in spirit to smoothness assumptions when estimating probability densities by kernel density estimators, but on measures with an infinite dimensional support.

In practice, it is through P_0 that the researcher introduces functional form restrictions into the model. For instance, if one wants to assume continuity in the relation between Y and X, one will choose a P_0 which only puts positive measure on continuous paths. $\hat{P}_0$ will then only sample continuous paths in practice. This way, one can theoretically introduce any form of functional form restriction into the model. These restrictions will rule out certain paths (for instance the assumption of an increasing relationship between Y and X will rule out all paths which are non-increasing) in the model and therefore restrict the unobservable heterogeneity captured by $\mathcal{W}$ over which the programs need to optimize.

One also needs to impose some regularity on the objective function.

Assumption 3 (Regularity of objective function)**.** *For every fixed $A_y \in \mathscr{B}_{\mathcal{Y}}$ and $A_x \in \mathscr{B}_{\mathcal{X}}$, the kernel $K(Y_x(w), A_y, A_x)$ is either*

(i) bounded and Lipschitz continuous in $Y_x(w)$ with Lipschitz constant $L < +\infty$ or

(ii) takes the form of an indicator function, i.e. $\mathbb{1}_G(Y_x(w))$ for some event G.

The following theorem gives probabilistic guarantees of for the approximation of the infinite programs (3) by the semi-infinite programs (4). Denote by V^* and V_* the value functions of(3) and by $\tilde{V}^*_{l,\kappa}$ and $\tilde{V}_{*,l,\kappa}$ the value functions of (4).

Theorem 1 (Probabilistic approximation via the "sampling of paths"-approach)**.** *Under Assumptions 1 – 3, with probability $1-\rho$,*

$$\max\{|V^* - \tilde{V}^*_{l,\kappa}|, |V_* - \tilde{V}_{*,l,\kappa}|\} < C \sup_{w\in[0,1]} \omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1}) + \sqrt{\frac{\log\left(\frac{\bar{C}}{\rho}\right)}{\bar{D}l}} + o(1). \tag{7}$$

$0 < C < +\infty$ is a constant depending on κ and $0 < \bar{C} < +\infty$ and $0 < \bar{D} < 2$ are constants depending on $\kappa, \beta, \varepsilon, F_{Y,X|Z=z}, C_{RN}$ and L. The $o(1)$-term appears from a logistic approximation of the indicator functions from the constraint (and potentially the objective function).

Theorem 1 jointly provides a lower bound on the number of sampled paths l as well as the number of terms in the wavelet decomposition κ for the semi-infinite program to provide a good approximation to the infinite dimensional program for a requested probability ρ. The result follows from large deviation results for empirical processes with functions of finite entropy numbers (van der Vaart & Wellner 2013, chapter 2.14). The main assumptions for this are the Lipschitz continuity of the kernel $K(Y_x, A_y, A_x)$ and especially the and Hölder continuity of the Radon-Nikodym densities $\frac{dP_W}{dP_0}(w)$. The constants $\bar{C}$ and $\bar{D}$ are a result of these large deviation results, which depend mostly on the smoothness parameters involved.[17]

The introduction of the randomness via sampling the paths permits the use of statistical procedures for the purpose of function approximation. The idea is to perform statistical estimation *with respect to the*

[17] Under stronger assumptions—like a finite Vapnik-Chervonenkis dimension (Vapnik 1998)—on the Radon-Nikodym derivatives and K, one can achieve an even better convergence by using the results in Vapnik (1998, chapter 5). These assumptions are often unreasonably strong (in particular the assumption that the densities have finite VC dimension). The focus in this article hence lies on the weaker assumption which give slightly worse approximation guarantees.

counterfactual elements via sampling. In this respect, Assumption 2 is important for obtaining the rate of convergence in Theorem 1: if all Radon-Nikodym derivatives $\frac{dP_W}{dP_0}(w)$ are smooth and have support in all of $[0,1]$, one does not need many draws l to approximate the optimization problem. For more concentrated densities, the approximation can be much worse in the sense that one needs a substantially larger sample l to approximate the optimal density. The approach proposed via Theorem 1 can hence be seen as a nonparametric estimation of a probability density with *infinite dimensional support* and where one has complete control over the data sample.

The method proposed here differs from existing approaches for partial identification in infinite dimensional spaces that rely on specifying the identified set via a set of inequalities (Chesher & Rosen 2017). By exploiting the optimization structure, the proposed method has a straightforward way of "selecting the binding inequalities". Intuitively, instead of having to specify the complete set of inequalities describing the identified set (via Artstein's inequality for instance), the method "puts a probability measure on the set of inequalities" and samples from these. This, coupled with the optimization problem, allows the program to sample the "probably binding" constraints in the optimization and reduces the infinite dimensional—and unsolvable—problem of specifying all inequality constraints in the infinite dimensional space down to a finite set.

Theorem 1 complements existing results on the approximation of infinite dimensional (linear) programs by semi-infinite dimensional programs by explicitly introducing randomness via sampling. While articles which employ probabilistic arguments exist in the general mathematical literature on function approximation (e.g. Girosi 1995, Pucci de Farias & Van Roy 2004), Theorem 1 samples *the state space of the problem*, whereas other approaches sample constraints (Pucci de Farias & Van Roy 2004); moreover, Theorem 1 does so without the introduction of dynamic assumptions (Kappen 2007). For the types of problems considered here, the randomness introduced is actually required, because one can never optimize over the complete path space directly and is hence forced to sample. In this respect, note that standard solution concepts for infinite dimensional programs work on a Euclidean state space (Anderson & Nash 1987), so that sampling is not actually necessary for solving these programs.

4 Inference

The statistical randomness follows from approximating the population distribution $F_{Y,X|Z}$ by a finite-sample estimator $\hat{F}_{Y,X|Z;n}$, potentially smoothed via some bandwidth h_n, where n denotes the size of this sample.[18] This subsection introduces large sample results which enable the researcher to perform inference on the solution of the programs (4). These results are only derived for each bound separately and for fixed l and ε. In order to derive inference bounds on the whole identified set, one can use well-established results from the literature, such as Imbens & Manski (2004) and Stoye (2009) as the outcomes of interest are univariate and hence form an interval.

Even though the theoretical programs (4) have relaxed constraints, it could potentially still be the case that for very small $\varepsilon > 0$ there exist data-generating processes $F_{Y,X|Z}$ that no $P_W \in \mathscr{P}^*(\mathcal{W})$ can replicate; this is especially true in the case where many additional form restrictions are imposed on the model. In this case, however, the data-generating process directly introduces testable assumptions on the model (Gunsilius

[18] The theoretical results in this section are derived for the standard empirical cumulative distribution function $\hat{F}_{Y,X|Z;n}$. They extend straightforwardly to smoothed estimators $\hat{F}_{Y,X|Z;h_n}$. For this, all one has to do in the proofs is to replace the classical Glivenko-Cantelli and Donsker theorems by analogous versions for smoothed empirical processes. These results (and the corresponding weak assumptions) are contained in Giné & Nickl (2008) for instance.

2019), as the model under the respective assumptions is not able to replicate the observed data. In this sense, obtaining bounds is complementary to testing the instrumental variable model. Since the focus of this article is on estimation of bounds, it is convenient to introduce an assumption on the data-generating process which guarantees that the constraint is non-empty. In the following, $\mathcal{F}_{Y,X|Z=z}$ denotes a set of all conditional cumulative distribution functions on $[0,1]^2$ satisfying certain assumptions the researcher is comfortable to assume for the given data-generating process. This set is equipped with the $L^\infty([0,1]^2)$-norm.

Assumption 4 (Non-emptiness of the constraint set)**.** *For given $F_{Y,X|Z=z} \in \mathcal{F}_{Y,X|Z=z}$ there exists a ball $\mathcal{B}_r \in \mathcal{F}_{Y,X|Z=z}$ of radius $r > 0$ such that the constraint set $\tilde{\mathcal{C}} :=$*

$$\left\{ \hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W}) : \left\| F'_{Y,X|Z=z}(y,x) - \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(i), \tilde{X}_z^\kappa(i)) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^2)} \leq \varepsilon \right\}$$

is non-empty for some small $\varepsilon > 0$ and all $F'_{Y,X|Z=z} \in \mathcal{B}_r$, $z \in [0,1]$.

Assumption 4 is deliberately high-level, because (i) specific assumptions on the data-generating process $\mathcal{F}_{Y,X|Z=z}$ and the model $\mathscr{P}^*(\mathcal{W})$ usually come from economic theory, (ii) an empty constraint for a given data-generating process corresponds to the existence of testable implications on the model, and (iii) Assumption 4 is only required to obtain regular and well-behaved asymptotic results, but not for any other results in this article. It is also straightforward to derive a low-level sufficient condition on $\mathcal{F}_{Y,X|Z=z}$ implying Assumption 4 in the case where only Assumption 1 but no other shape restrictions are imposed on $\mathscr{P}^*(\mathcal{W})$. For instance, when the set $\mathcal{F}_{Y,X|Z=z}$ consists only of distribution functions $F_{Y,X|Z=z}$ which are laws to stochastic processes $[Y,X]_z$ whose paths have an extended modulus of continuity $\omega'_{[Y,X]_z}(\delta)$ satisfying $\limsup_{\delta\to 0} \omega'_{[Y,X]_z}(\delta) = 0$, then Assumption 4 is fulfilled. This follows directly from the fact that this condition on the extended modulus of continuity implies that almost all paths of the observable process $[Y,X]_z$ lie in the Skorokhod space defined by Assumption 1, which is the assumption made on the latent process $[Y,X]^*_z$.[19]

From now on $\hat{F}_{Y,X|Z;n}$ denotes the conditional empirical distribution function, the terms $\tilde{V}_{*,l,\kappa}(F_{Y,X|Z})$ and $\tilde{V}^*_{l,\kappa}(F_{Y,X|Z})$ denote the minimal and maximal value functions of (4), and $\hat{\tilde{V}}_*(\hat{F}_{Y,X|Z;n})$ and $\hat{\tilde{V}}^*(F_{Y,X|Z;n})$ their empirical counterparts. The first result concerns the consistency of these value functions. The idea is to use the Glivenko-Cantelli theorem (van der Vaart 2000, Theorem 19.1) which provides the convergence $\hat{F}_{Y,X|Z;n}$ to $F_{Y,X|Z}$ in $L^\infty([0,1]^2)$-norm. Then the Berge Maximum theorem (Aliprantis & Border 2006, Theorem 17.31) provides the consistency results for the value function.

Proposition 2 (Consistency)**.** *Under Assumptions 1 – 4 it holds that*

$$|\hat{\tilde{V}}_*(\hat{F}_{Y,X|Z=z;n}) - \tilde{V}_{*,l,\kappa}(F_{Y,X|Z=z})| \to 0 \qquad \textit{and} \qquad |\hat{\tilde{V}}^*(\hat{F}_{Y,X|Z=z;n}) - \tilde{V}^*_{l,\kappa}(F_{Y,X|Z=z})| \to 0$$

for almost all $z \in \mathcal{Z}$ almost surely as $n \to \infty$.

The derivation of the large sample distribution of $\hat{\tilde{V}}^*$ and $\hat{\tilde{V}}_*$ follows form Donsker's theorem (van der Vaart 2000, Theorem 19.3) in combination with standard sensitivity arguments in optimization problems, in particular Theorem 4.26 and Proposition 4.47 in Bonnans & Shapiro (2013), and the functional delta method (Shapiro 1991, Theorem 2.1). The formal result is captured in the following

[19] Proposition 2 in Gunsilius (2019) contains a smoothness assumption on $F_{Y,X|Z}$ which guarantees that the theoretical constraint in (3)—and therefore the constraint in (4) for every $\varepsilon > 0$ and large enough sample l—is satisfied when the paths of the counterfactual processes Y_x and X_z have almost surely continuous paths.

Proposition 3 (Large sample distribution)**.** *Under Assumptions 1 – 4 it holds that*

$$\sqrt{n}(\hat{\tilde{V}}_*(\hat{F}_{Y,X|Z=z;n}) - \tilde{V}_{*,l,\kappa}(F_{Y,X|Z=z})) \Rightarrow d\tilde{V}_{*,F_{Y,X|Z=z}}(\mathbb{G}_{F_{Y,X|Z=z}}) \qquad \textit{and}$$
$$\sqrt{n}(\hat{\tilde{V}}^*(\hat{F}_{Y,X|Z=z;n}) - \tilde{V}^*_{l,\kappa}(F_{Y,X|Z=z})) \Rightarrow d\tilde{V}^*_{F_{Y,X|Z=z}}(\mathbb{G}_{F_{Y,X|Z=z}}),$$

for almost all $z \in \mathcal{Z}$ *as* $n \to \infty$*.*

Here, $d\tilde{V}_{*,F_{Y,X|Z=z}}(F'_{Y,X|Z=z})$ is the directional Hadamard derivative of $\tilde{V}_*$ at $F_{Y,X|Z=z}$, $\mathbb{G}_{F_{Y,X|Z=z}}$ is a Brownian bridge with covariance function

$$\text{Cov}_{\mathbb{G}_{F_{Y,X|Z=z}}} = F_{Y,X|Z=z}(\min\{y,y'\},\min\{x,x'\}) - F_{Y,X|Z=z}(y,x)F_{Y,X|Z=z}(y',x')$$

for all $(y,x),(y',x') \in [0,1]^2$, $z \in \mathcal{Z}$, and "$\Rightarrow$" denotes weak convergence.

For nonlinear objective functionals KP_W, the directional Hadamard derivative takes the form

$$d\tilde{V}_{*,F_{Y,X|Z=z}}(F'_{Y,X|Z=z}) = \delta_{F_{Y,X|Z=z}} L(\hat{P}_1, \lambda(\hat{P}_W), F_{Y,X|Z=z})(F'_{Y,X|Z=z}),$$

where $L(\cdot,\cdot,\cdot)$ denotes the Lagrangian of the program (4), $\lambda(\hat{P}_1)$ denotes the respective Lagrange multiplier (which is unique for given $\hat{P}_1$ by Proposition 4.47 in Bonnans & Shapiro 2013), and $\delta_{F_{Y,X|Z=z}} L$ is the directional derivative of L in its third argument in direction $F'_{Y,X|Z=z}$ at $F_{Y,X|Z=z}$. In many cases, one can say a little more about the form of the Lagrangian under Assumption 3, as the objective functional then becomes linear in P_W. In this case, the directional Hadamard derivative takes the of an optimization over an inner product between the Lagrange multiplies λ and the direction $F'_{Y,X|Z=z}$ in some instances (Bonnans & Shapiro 2013, Theorem 4.27).[20]

Proposition 3 implies that the rate of convergence is parametric, which is a sensible result as the quantity of interest is univariate and uniformly bounded. Moreover, even though the large sample distribution of the value functions is not a standard Brownian bridge process, it still has a relatively common form from a purely statistical perspective, as it takes the form of the first-order directional Hadamard derivative of the value function taken at $F_{Y,X|Z=z}$ in directions $F'_{Y,X|Z=z} \in \mathcal{F}_{Y,X|Z=z}$. In addition, there are several results in the literature (Dümbgen 1993, Fang & Santos 2018, Hong & Li 2018) which establish bootstrap methods for estimating this type of large sample distribution in practice. In particular, they deal with general directional Hadamard differentiability (Shapiro 1991), which conforms with Proposition 3, so that these subsampling/bootstrap results are directly applicable to the problems (4). These bootstrap-type arguments are convenient in models with a light computational burden mostly. In more complex models one should use the analytically derived large sample theory.

Together, Theorem 1 and Proposition 3 provide a general overview of the approximative behavior of the programs (4). In conjunction with these subsampling results and the existing inference results for partial identification (Imbens & Manski 2004, Stoye 2009), the proposed method covers partial identification, practical estimation, and inference of bounds on functionals in general instrumental variable models.

[20] Proposition 3 works equally well with a smoothed estimator $\hat{F}_{Y,X|Z=z;h_n}$ of $F_{Y,X|Z=z}$ and bandwidth h_n. The only important requirement is that the respective empirical process converges to a Brownian bridge. In the smoothed case, these results—which allow for bandwidths obtained via cross-validation—have been derived in Giné & Nickl (2008) for instance.

5 Implementation

The programs (4) are *semi-infinite* programs (Anderson & Nash 1987), which naturally reduce to finite dimensional problems in practice by approximating the space $[0,1]^3$ where Y, X, and Z live. One can do this in two general ways. The first is to simply evaluate $\hat{F}_{Y,X|Z=z;n}$ on the values taken by the sample $(Y_i, X_i, Z_i)_{i=1,\dots,n}$. The second is to evaluate $\hat{F}_{Y,X|Z=z;n}$ on a finite grid that spans $[0,1]^3$. This article focuses on the latter part as a grid approach gives more flexibility with respect to the computational requirements: one can make the grid coarser or finer, depending on the available memory.[21] Throughout this section, the index ι captures the degree of approximation of the grid. For instance, $\iota = 11$ means that this approximation decomposes the unit interval into 11 points $0, 0.1, 0.2, \dots, 0.9, 1$, which will be taken to be equidistant without loss of generality. Also, all three intervals for Y, X and Z are decomposed in the same way, so that ι is the only necessary parameter controlling the approximation.

The practical implementation deviates from the theoretical approach in that it uses a smoothed variant $\hat{F}_{Y,X|Z=z;h_n}$ of the empirical conditional cumulative distribution function, where the bandwidth is determined via cross-validation. Heuristically, it seems as the introduced smoothness gives more robust results compared to the standard empirical cumulative distribution function.[22]

Under a given finite approximation, the programs take the form

$$\begin{aligned} &\underset{\mu \geq 0,\ \vec{1}'\mu \leq 1}{\text{minimize/maximize}} \quad \tilde{K}'\mu \\ &\text{s.t.} \quad \|\tilde{\Theta}\mu - \hat{F}_{Y,X|Z;h_n}\|_2^2 \leq \varepsilon \end{aligned} \tag{8}$$

where μ is a $l \times 1$ vector which corresponds to the Radon-Nikodym derivative $\frac{d\hat{P}_W}{d\hat{P}_0}(i)$ with row-dimension equal to the number of sampled paths l[23], $\vec{1}$ denotes the vector of the same dimension as μ containing all ones, $\tilde{K}'$ is a $1 \times l$ vector, and $\|\cdot\|_2$ denotes the Euclidean norm.[24] $\tilde{\Theta}$ is a $\iota^3 \times l^2$-matrix which maps the realization of the stochastic processes to the distribution $\hat{F}_{Y,X|Z;h_n}$. The $L^2([0,1]^2)$-norm from (4) reduces to the Euclidean norm due to the approximation of $[0,1]^3$ by a finite grid.

In practice, it is convenient to write the programs (8) in their penalized form as

$$\underset{\mu \geq 0, \vec{1}'\mu \leq 1}{\text{minimize/maximize}} \quad \tilde{K}'\mu + \frac{\lambda}{2}\|\tilde{\Theta}\mu - \hat{F}_{Y,X|Z;h_n}\|_2^2 \tag{9}$$

for some penalty λ corresponding to the original constraint qualification $\varepsilon > 0$. Intuitively, a larger λ corresponds to a tighter ε. The choice of the Euclidean norm $\|\cdot\|_2$ for the constraint is convenient, as (9)

[21] This is also a difference to the computational approach in Kitamura & Stoye (2018), who set up their problem based on the observed realizations in the data and not a grid. This is a direct consequence of their finitary approach. In contrast, the infinite approach in this article allows arbitrary discretization of Y, X, and Z and is therefore also a potential complementary approach to Kitamura & Stoye (2018) in their setting.

[22] For the practical estimation of $\hat{F}_{Y,X|Z=z;h_n}$ the method uses the "np"-package in *R* (Hayfield & Racine 2008) with a standard cross-validated bandwidth.

[23] Note that all elements in μ must lie in $[0,1]$, as $\frac{d\hat{P}_W}{d\hat{P}_0}(i)$ is defined on the *finite and discrete space* of l paths which were sampled by some $\hat{P}_0$. This means that $\frac{d\hat{P}_W}{d\hat{P}_0}(i)$ can only put non-negative probabilities of at most one on the occurrence of each path. Intuitively, this follows from the fact that $\frac{d\hat{P}_W}{d\hat{P}_0}(i)$ is a probability mass function. These bounds on μ are included as additional constraints using $\vec{1}$.

[24] A' denotes the transpose of the matrix A.

can be rewritten as

$$\begin{aligned}\min_{\mu\geq 0,\vec{1}'\mu\leq 1} \quad & \frac{\lambda}{2}\mu'\tilde{\Theta}'\tilde{\Theta}\mu - \left(\lambda\tilde{\Theta}'\hat{F}_{Y,X|Z;h_n} - \tilde{K}\right)'\mu + \frac{\lambda}{2}\left(\hat{F}_{Y,X|Z;h_n}\right)'\hat{F}_{Y,X|Z;h_n} \\ \min_{\mu\geq 0,\vec{1}'\mu\leq 1} \quad & \frac{\lambda}{2}\mu'\tilde{\Theta}'\tilde{\Theta}\mu - \left(\lambda\tilde{\Theta}'\hat{F}_{Y,X|Z;h_n} + \tilde{K}\right)'\mu + \frac{\lambda}{2}\left(\hat{F}_{Y,X|Z;h_n}\right)'\hat{F}_{Y,X|Z;h_n}.\end{aligned} \tag{10}$$

The programs (10) are quadratic due to the Euclidean norm used and can easily be solved. This article uses the alternating direction method of multipliers (ADMM) (Boyd, Parikh, Chu, Peleato & Eckstein 2011 and Parikh & Boyd 2014) for optimization. This algorithm is known to converge rather quickly to reasonable approximations of the optimum, which makes it a perfect tool for this purpose. The algorithm requires two more parameters, the augmented Lagrangian parameter ρ and an over-relaxation parameter α, which control the convergence of the ADMM algorithm to the optimum. In practice, an over-relaxation parameter of $\alpha = 1.7$ and an augmented Lagrangian parameter of ρ between 100 and 500 leads to fast and robust convergence.

The computational bottleneck in a practical implementation is the construction of the matrix $\tilde{\Theta}$, whose dimension grows polynomially with the granulation of the grid ι and the number of paths sampled l. In the case where (3) and *a fortiori* (4) and (8) are have linear objective functions, it is convenient to let $\tilde{\Theta}$ take the form of a binary sparse matrix: for each point $(y_\iota, x_\iota, z_\iota) \in [0,1]^3$ in the grid a given combination of paths $Y_x(i)$ and $X_z(i)$ either gets assigned a 0 if they jointly "do not go through" the intervals $[0, y_\iota] \times [0, x_\iota]$ for given values z_i or a 1 if they jointly do.[25] This sparseness is helpful as sparse matrices can be stored efficiently. In addition, the process of setting up $\tilde{\Theta}$ can be parallelized, which abates the computational costs even further if the researcher has access to several cores.

In many cases, however, a researcher might only have access to computational resources with very limited working memory. In such situations, it is still possible to apply the proposed method by a "sampling trick" which trades off memory requirements for time. In particular, the idea is to iteratively (i) sample with replacement a relatively small initial number l_0 of paths (depending on the available memory), (ii) optimize the programs (10) on this sample, (iii) obtain the value functions *as well as* the optimizers μ, (iv) *drop* all paths which were assigned a probability of (close to) 0 by the optimizer μ, (v) sample another relatively small number l_s, add these paths to the already existing paths and go back to (ii). The idea of this "sampling trick" is that paths which were assigned a probability of (close to) 0 by the optimal μ do not matter for the optimal value. By dropping these paths before sampling new ones, the memory requirements do not grow or only grow modestly in practice—at the additional cost of having to run this optimization for many iterations.[26]

When applying this sampling trick, the solution will be expressed as a *solution path* over the sampling iterations. This solution path in general will be erratic due to the nature of the sampling approach, but has the added benefit over the "static" direct method where all relevant paths are sampled immediately that

[25] In the case where the indicator function in the constraint (3) is approximated by a logistic function, the matrix $\tilde{\Theta}$ is no longer binary sparse.

[26] Discarding elements ex-post in optimization routines is not new. For instance Wu, Fang & Lin (2001) use discards in solving the general capacity problem on Euclidean state space. Currently, the approach is rather naïve as it performs a "random grid search" over the infinite dimensional path space. A more efficient implementation uses ideas from sequential MCMC approaches (see Schweizer 2012 for a recent overview): in the first iteration, randomly sample paths. Then optimize and discard all paths which were assigned a probability of zero. When sampling new paths, do not just sample randomly, but sample a fraction of paths which are close to the paths that were assigned a positive probability and sample another fraction of paths randomly to find other, different binding constraints. This approach is currently implemented for fixed fractions. An interesting question in this setting is to optimize this procedure, i.e. to optimize the fraction of randomly sampled paths compared to paths which are close to others. Intuitively, there seems to be an "exploration-exploitation-tradeoff", which one could optimize.

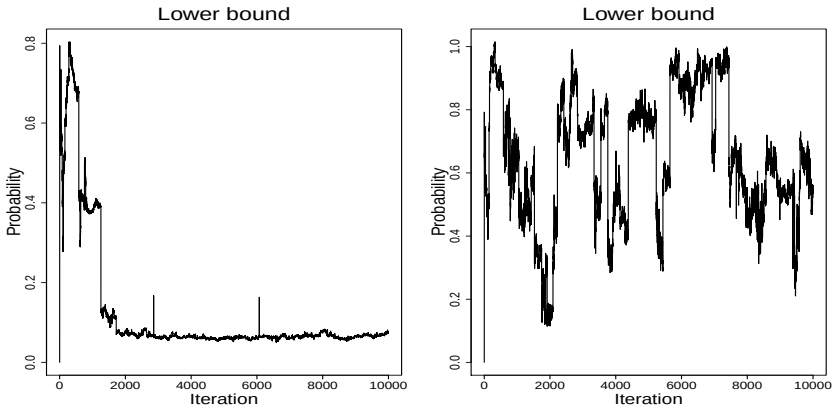

Figure 1: Depiction of convergence of the solution path for estimating $F_{X=0.75}(0.75)$ for $\lambda = 100$ (left) and $\lambda = 600$ (right) for a coarse approximation of length 5 of the unit interval, i.e. a decomposition $0, 0.25, 0.5, 0.75, 1$. 16 new paths are sampled at each iteration. The average of the values over the last 500 iterations in the left panel is nontrivial at 0.065.

one can gauge if the solutions "converge" to some stable limit after a "burn-in" period. This convergence relies on the choice of the penalty parameter λ. The larger λ is chosen, the more it forces the optimizer μ to replicate the observable $\hat{F}_{Y,X|Z;h_n}$. In fact, in the limit $\lambda \to +\infty$, the program forces to replicate the constraint perfectly. If λ is too low, the program ignores the constraint, which will always result in trivial bounds.

This implies that there exists a range of λ-values for which constraint and objective function are balanced. Figure 1 depicts the behavior of the solution paths of this "sampling trick" in a stylized setting of the household expenditure application in the next section, where the coarseness of the approximation of the unit interval makes it possible to sample *all possible* paths in order to see whether convergence still occurs if one is able to sample the "whole universe of paths". The left panel depicts a solution path which looks like it converges, while the right panel depicts a case of non-convergence for estimating a lower bound on $F_{Y(X=0.75)}(0.75)$.

In this form, these solution paths are reminiscent of the solution paths of regularized linear programs such as LASSO. The difference, however, is that the paths induced by this program are for a *fixed* λ, while the actual LASSO solution paths are traced out while varying λ. In order to have an analogue of the LASSO solution paths in the current method, one would have to solve the program for many different values of λ, which would generate a *system of solution paths*. Then one could choose the largest lambda for which the corresponding solution path converges to a stable value.[27]

6 Discussion

The setting for the method introduced in this article is that of a general instrumental variable model which can be represented as a nonseparable triangular model. This article is hence closely related to the literature on (partial) identification in these models, in particular to Imbens & Newey (2009). There, the authors

[27] It is an intriguing question how to determine an appropriate λ by data-driven methods. Such a data-driven method might open up the way for solving other infinite dimensional programs on path spaces in statistics and mathematics via a "sampling of paths"-approach.

introduce a flexible partial- and point identification approach in these models under a strict monotonicity assumption of the first stage in the instrumental variable model. In contrast, the current approach does not require monotonicity or other functional form restrictions and does not rely on a control function approach.

This method is also connected to the literature on causal inference with instruments (Angrist, Imbens & Rubin 1996, Balke & Pearl 1994, Imbens & Angrist 1994), simultaneous equation models (e.g. Blundell & Matzkin 2014, Matzkin 2003), and partial identification in demand estimation and welfare analysis (Dette, Hoderlein & Neumeyer 2016, Hausman & Newey 2016, Kitamura & Stoye 2018).[28] In particular, the setting in this article can be seen as an infinite analogue to the finitary approach in Kitamura & Stoye (2018). This infinite dimensional analogue is likely to be helpful in two settings which Kitamura & Stoye (2018) cannot address: first, the proposed method allows to incorporate structural nonparametric assumptions such as continuity into the model. This can be helpful in demand analysis, where smoothness assumptions allow to work with restrictions on the Slutsky matrix directly (Hausman & Newey 2016). Second, it allows for a grid-approach and smoothing for the practical estimation, which makes the method applicable in settings with a massive number of observations. In contrast, it seems as if Kitamura & Stoye (2018) need to include all data-points in their analysis, which makes the computational burden too high in data sets with many observations.

From a theoretical perspective, the most closely related article is Chesher & Rosen (2017) which introduces a general framework for partial identification using the Aumann integral and Artstein's inequality. This approach is situated within a general model that can incorporate any structural assumptions, and is theoretically more general than the method presented here, which works within the setting of nonseparable triangular models. What distinguishes the proposed method from their approach is the applicability in practice. In fact, results relying on Artstein's inequality run into severe curses of dimensionality for endogenous variables with more than a few points in their support, as the number of inequalities describing the identified region grows very rapidly (Beresteanu, Molchanov & Molinari 2012). This is a main reason for the often unsurmountable challenges current approaches based on specifying the identified set via inequalities face in many models of interest. In contrast, this article proposes to solve an infinite dimensional program, which handles the complexity of the model automatically. Another intuitive reason for the tractability of the proposed method even in models with continuous variables is its exclusive focus on estimating functionals of counterfactual distributions (a one-dimensional quantity), i.e. the "endpoints" of the identified set; in contrast, Chesher & Rosen (2017) also identify the model parameters such as the production functions (an infinite dimensional quantity), i.e. the complete identified set.[29]

This article is not the first to propose an optimization approach for obtaining bounds in partial identification models in a counterfactual setting, see for instance Aguiar, Allen & Kashaev (2019), Chiburis (2010), Demuynck (2015), Hansen, Heaton & Luttmer (1995), Honoré & Tamer (2006), Honoré & Lleras-Muney (2006), Manski (2007), Manski (2014), Molinari (2008), Norets & Tang (2013), Lafférs (2015), Kamat (2017), Torgovitsky (2016), Kitamura & Stoye (2018), Mogstad, Santos & Torgovitsky (2018), Tebaldi, Torgovitsky & Yang (2019), but it is the first to present an infinite dimensional optimization problem for doing so. The infinite dimensionality is key in order to solve models with a continuous endogenous variable theoretically as well as practically. Furthermore, in combination with the "sampling of paths"- approach, it allows researchers to solve complex but finitary models by considering random variables with many points

[28] By introducing Slutsky-type assumptions on the stochastic processes introduced in this article, one can potentially apply the proposed method to estimate bounds in welfare analysis in the most general setting, complementing the results in Hausman & Newey (2016). See the discussion in the conclusion.

[29] This fact was pointed out by Arie Beresteanu and Francesca Molinari.

in their support as continuous. Garen (1984) for instance treats years of schooling as a continuous variable for similar reasons.

The most closely related articles in terms of underlying ideas are Balke & Pearl (1994) and Balke & Pearl (1997) which provide tight upper and lower bounds when all variables are binary. In particular, the proposed method in this article reduces to their approach in the binary case. These results strengthen the original results in Robins (1989), Manski (1990), and Manski (1997), who found upper and lower bounds on causal effect also in the setting where the treatment is binary. They also enabled Kitagawa (2009) to derive closed-form solutions for sharp bounds on causal effects for a continuous outcome and binary treatment and instrument. Recently, Russell (2017) derived sharp bounds on causal effects for discrete outcome and treatment using Artstein's inequality and optimal transport theory similar to Galichon & Henry (2011).

The proposed infinite dimensional optimization program is similar in form to—but more general in its state-space dimensionality and objective function than—the *general capacity problem* (Anderson & Nash 1987, Anderson, Lewis & Wu 1989, Lai & Wu 1992), which itself is a generalization of Choquet's capacity theorem (Choquet 1954). Interestingly, two other existing articles dealing with partial identification in general models (Beresteanu, Molchanov & Molinari 2012, Galichon & Henry 2011) both directly deal with Choquet's capacity theory and Aumann integrals. Their focus is somewhat different as the first article assumes discrete treatment, while the second deals with partially identifying unknown parameters in a structural setting, but the connection via capacity theory seems worth mentioning.

The proposed "sampling of paths"-approach allows the researcher to introduce assumptions on the model by choosing an appropriate sampling distribution on the paths of the stochastic processes. In practice, the optimization over probability measures on the paths of stochastic processes then exclusively considers probability measures which possess a Radon-Nikodym derivative with respect to this proposed sampling measure. This solution approach is reflective of the ELVIS estimator (Schennach 2014), which exponentially tilts a user-chosen probability distribution to satisfy given population moment conditions. In the proposed method, the applied researcher can introduce structural assumptions into the model (e.g. allowing for only increasing paths, equivalent to a monotonicity assumption on the instrumental variable model) by setting the probability of certain paths (e.g. all non-increasing paths) to zero *a priori*. The method then automatically only optimizes over the sampled paths. Similarly, the ELVIS estimator also requires a reference measure whose support must contain the support of the tilted probability measures. Notwithstanding this connection, both approaches are complementary: the proposed method works with problems without the necessity to introduce moment conditions, whereas the ELVIS estimator efficiently addresses models that can be written in terms of moment restrictions which contain unobservables.

Other identification results in nonseparable triangular models often focus on the production function h and require monotonicity assumptions (e.g. Chesher 2003, Chernozhukov & Hansen 2005, Shaikh & Vytlacil 2011, Torgovitsky 2015, d'Haultfœuille & Février 2015) or presuppose some other general structural relationship (Florens, Heckman, Meghir & Vytlacil 2008). Recently, Heckman & Pinto (2018) introduced the concept of unordered monotonicity in the case where endogenous variable is discrete and unordered.

7 Demonstration

This section presents practical settings for testing the method. The first is a small Monte-Carlo simulation, where the focus lies on the sensitivity of the method to the choice of the number of basis functions κ and the penalty term λ. The second is an application to real data, where the method manages to obtain informative

bounds under minimal assumptions.

7.1 Simulation

The idea for generating the data in this simulation exercise is to (i) generate paths of stochastic processes Y_x and X_z for a given set $\{z_i\}_{i=1,\dots,m} \in [0,1]$, (ii) combine the two processes to a joint process $[Y,X]_z^*$, (iii) randomly sample points z and corresponding points (y,x) induced by the paths of $[Y,X]_z^*$ to obtain the data (Y,X) for the given set of $\{z_i\}$. The processes Y_x and X_z are Gaussian processes with mean 0.5 and a squared exponential covariance kernel of the form $k_{SE}(z-z') \coloneqq \sigma^2 \exp\left(-\frac{(z-z')^2}{\ell^2}\right)$, where the length parameter ℓ is 0.5 for Y_x and 0.2 for X_z, and the variance parameter σ^2 is 0.2 for Y_x and 0.15 for X_z. All paths of these processes are restricted to lie within $[0,1]$. 2500 paths were sampled, generating 5000 random data points from it on a relatively coarse grid based on a dyadic approximation of order 3, i.e. grid points at $\frac{k}{8}$ for $k = 0,\dots,8$. The values z were drawn uniformly on the unit interval.

The goal is to estimate $E[Y(0.5)]$, which is equal to 0.5 by construction. $\hat{F}_{Y,X|Z=z;h_n}$ is estimated by kernel density methods with a cross-validated bandwidth using the *np*-package in *R*. The only assumption on the paths made for estimation is continuity. The hat-functions defined in (6) form the basis for different levels of κ and λ. Figure 2 depicts the convergence of the bounds in the "sampling trick" approach. The left

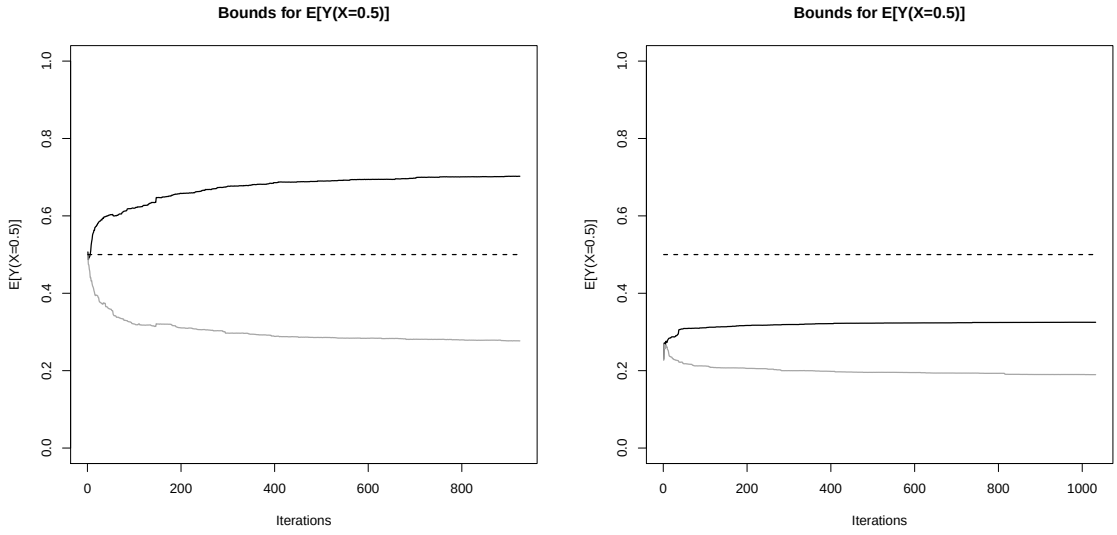

Figure 2: Convergence of upper (black) and lower (gray) bounds on $E[Y(X=0.5)] = 0.5$. The penalty term is $\lambda = 5$ for an equidistant approximation of the unit interval by 9 points. At each iteration the method samples 10 new continuous paths. In the left panel the paths are constructed by summing over the wavelet basis with $\kappa = 1,\dots,5$. In the right panel the paths are constructed for fixed $\kappa = 8$.

panel depicts the convergence of the bounds on $E[Y(X=0.5)]$ for $\lambda = 5$ and a sieve basis which consists of a sum of the basis functions for $\kappa = 1,\dots,5$. The method seems to converge nicely to bounds which contain the true value and are actually reasonably tight, especially for such a coarse grid and the fact that no functional form restrictions are imposed on the model except continuity. The right panel depicts the same thing, only there the paths $Y_x(w)$ and $X_z(w)$ are constructed for one form of basis functions corresponding to $\kappa = 8$. Clearly, the method obtains biased results in this case, as the true value does not lie between the

upper- and lower bound. Intuitively, this bias stems from the fact that the paths generated for fixed $\kappa = 8$, despite being continuous, are not smooth and very erratic, while the paths generated in the simulation are exceptionally smooth due to the choice of the squared exponential covariance kernel and relatively large length parameters ℓ. This bias can be understood as a "nonparametric misspecification bias" by sampling paths which are very different from the paths in the true data generating process. Fortunately, the left panel shows that already summing over only a few different basis functions removes this bias in this stylized example, which most likely also holds in more realistic settings (Chen 2007).

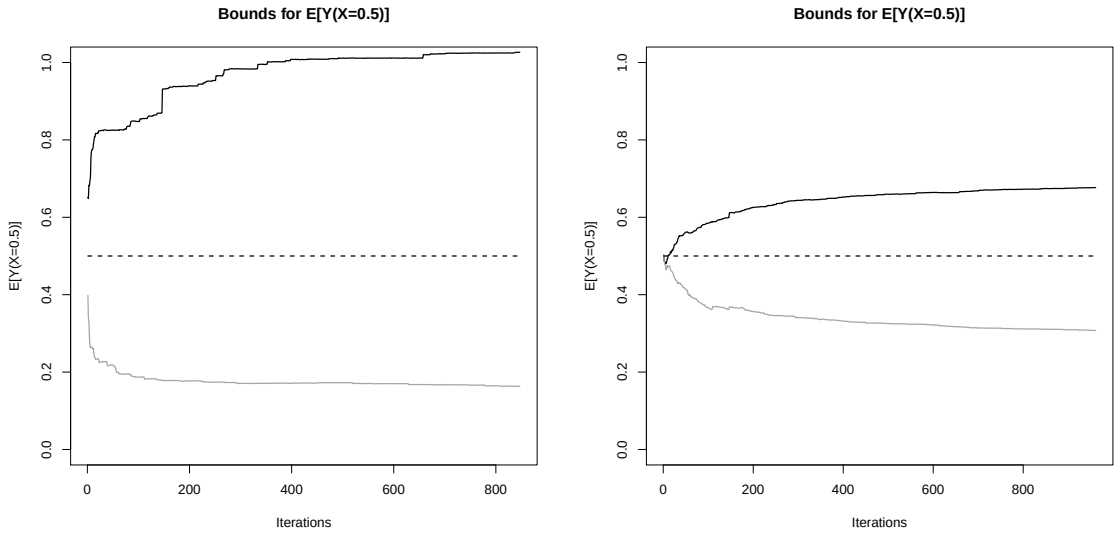

Figure 3: Convergence of upper (black) and lower (gray) bounds on $E[Y(X = 0.5)] = 0.5$. The penalty terms are $\lambda = 0.1$ (left panel) and $\lambda = 50$ (right panel) for an equidistant approximation of the unit interval by 9 points. At each iteration the method samples 10 new continuous paths. The paths are constructed by summing over the wavelet basis with $\kappa = 1, \ldots, 5$.

Figure 3 depicts the sensitivity of the method to choosing different penalty terms λ in this setting. In the left panel $\lambda = 0.1$ while in the right $\lambda = 50$. The bounds in the left case are significantly wider than on the right, but both estimators are unbiased. Interestingly, and in contrast to Figure 1, even larger λ does not lead to erratic behavior of the paths. This is most likely a result of the smoothness of the simulated paths in combination with a very high signal-to-noise ratio of the simulated data, which implies that there always exists an optimal μ in the finite dimensional problems (9). In contrast, when the signal-to-noise ratio is low, which is often the case in real-world applications, then the behavior of the solution paths will become too erratic as depicted in Figure 1. Another interesting fact is that the bounds do not tighten much when going from $\lambda = 5$ to $\lambda = 50$, which most likely is a result of the coarse approximation of the unit interval.

7.2 Application

As a demonstration of its capabilities, the method estimates bounds on expenditure differences using the 1995/1996 UK family expenditure survey. This problem is well suited as it (i) is nonlinear with continuous variables (Blundell, Chen & Kristensen 2007, Imbens & Newey 2009), (ii) allows to gauge if the program actually obtains reasonable results, and (iii) provides a setting not directly related to causal inference,

showing the scope of the proposed method. Therefore, in the following the focus will be on the outcomes food and leisure.

Analogous to Blundell, Chen & Kristensen (2007) and Imbens & Newey (2009), the outcome of interest Y will be the share of expenditure on a commodity and X will be the log of total expenditure, scaled to lie in the unit interval. The instrument used in this setting is gross earnings of the head of the household, which assumes that the way the head of the household earns the money is (sufficiently) independent of the household's expenditure allocation; this instrument is used in both Blundell, Chen & Kristensen (2007) and Imbens & Newey (2009). All three variables are inherently continuous which makes this problem a nice setting for demonstrating the practical implementation of the method.

The sample is restricted to the subset of married and cohabiting couples where the head of the household is aged between 20 and 55, and couples with 3 or more children are excluded. Also excluded are households where the head of the household is unemployed in order to have the instrument available for each observation. The final sample comprises 1650 observations.[30] Table 1 gives a summary of the outcome distributions,

	Leisure	Food
mean	0.143	0.182
standard deviation	0.110	0.0723
skewness	1.83	0.686
min	0.00122	0.00946
lower quartile	0.0689	0.131
median	0.111	0.176
upper quartile	0.183	0.225
max	0.831	0.657

Table 1: Summary statistics for the outcome distributions

showing that the relative expenditure on leisure is much more skewed towards zero but with a higher variance than the distribution for food.

The only shape restriction on the instrumental variable model is continuity, i.e. h and g are continuous functions in X and Z, respectively. This is a natural assumptions since Engel curves are usually believed to be continuous. It implies that almost all of the paths of Y_x and X_z lie in $C([0,1])$, the space of continuous functions. No other assumptions are upheld. The most general current approaches either require continuity and strict monotonicity of g (Imbens & Newey 2009) or of h (Blundell, Chen & Kristensen 2007) in the unobservable W. In contrast, the proposed method does not require any monotonicity assumptions and hence intuitively gives an indication of how much information is available in the data to solve this problem. Surprisingly, there seems to be a substantial amount of information, as the obtained bounds indicate that food is a necessity- and leisure is a luxury good without any assumptions on the model besides continuity. Furthermore, when introducing monotonicity assumptions *in the observable variables*, the bounds become significantly tighter, showing the identificatory strength of these assumptions in this setting.

Figure 4 depicts the solution paths for obtaining bounds on the counterfactual difference $F_{Y(X=0.75)}(0.15) - F_{Y(X=0.25)}(0.15)$ for a reasonably fine approximation of the unit interval into 17 equidistant points (which corresponds to a dyadic approximation of order 4).

Here the penalization parameter $\lambda = 1$, which seems to provide reasonable convergence to the solution, in particular for the food data. Remarkably, the estimated bounds in this setting are qualitatively informative

[30] The data used in both Blundell, Chen & Kristensen (2007) and Imbens & Newey (2009) is based on the 1994/1995 UK family expenditure survey. Under the same restriction they end up with 1655 observations.

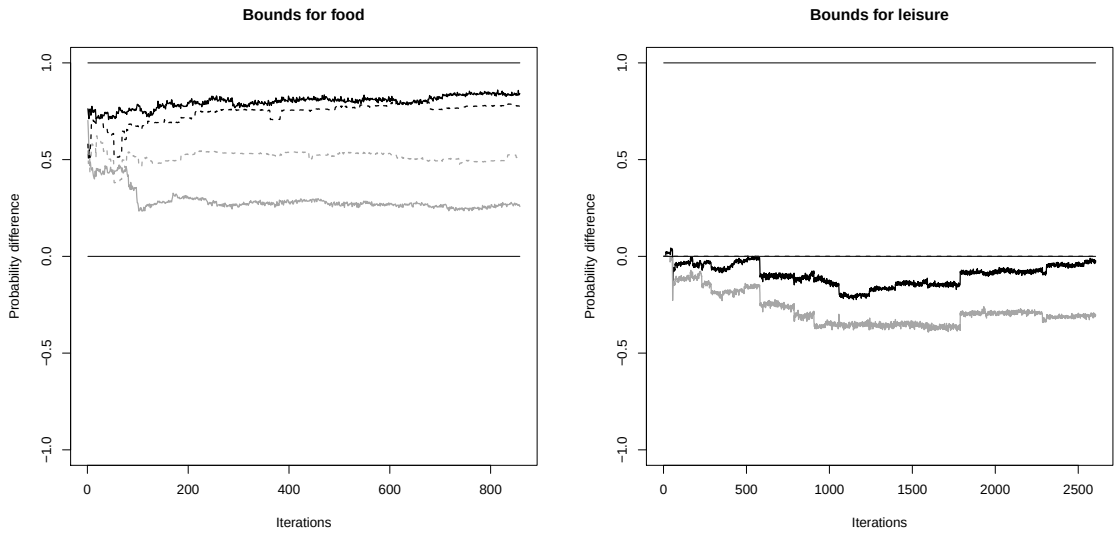

Figure 4: Convergence of upper (black) and lower (gray) bounds on $F_{Y(X=0.75)}(0.15) - F_{Y(X=0.25)}(0.15)$ for Y being the relative spending on food (left panel) and leisure (right panel). The solid lines indicate the bounds under no additional assumptions on the model, the dashed lines indicate the bounds under the additional assumption that Y is decreasing (food) or increasing (leisure) in overall expenditure X and that X is increasing in income of the head of the household Z. The penalty term is $\lambda = 1$ for an approximation of the unit interval by 17 points. At each iteration the method samples 25 new paths, of which 6 satisfy the respective monotonicity requirement.

for the problem. The left panel depicts the households' expenditures on food and the right depicts their expenditures on leisure. The solid lines are the upper- and the lower bound for a model without further assumptions, while the dashed lines are the upper- and lower bounds for a model with the additional assumption that Y is increasing for leisure and decreasing for food in overall expenditure X and that X is increasing in income of the head of the household Z.

Consider the left panel first, which depicts the expenditures on food. Here, the general bounds seem to converge and the average values of the bounds over the last 200 iterations are 0.84 and 0.26 for the general upper- and lower bound, and 0.78 and 0.50 for the corresponding upper- and lower bounds for the monotone model. All four bounds are positive, which indicates that families that spend a lot in general ($X = 0.75$) and spend up to 15% on food ($Y \in [0, 0.15]$) would spend much more on food relatively to overall expenditure if they spent much less overall ($X = 0.25$). Put differently, families are much more likely to lie in higher quantiles for expenditure on food if they lie in the lower quartile in overall spending than families that lie in the upper quartile in overall spending, which is the defining characteristic of a necessity good at this given level $y^* = 0.15$.

It is rather striking that even the model without monotonicity assumptions produces bounds which reflect this fact via a positive lower bound. In this regard, note that the monotonicity assumptions do not only tighten the bounds, but also shift up the lower bound, indicating that monotonicity has a stronger identificatory content in this setting. In particular, they imply that more families (between 50% and 78%) in the upper quartile of overall spending ($X = 0.75$) spend only up to 15% on food compared to families in the lower quartile of overall spending ($X = 0.25$). As mentioned, without monotonicity assumptions, these differences can be as high as 84% and as low as 26%, but all positive.

The results for expenditure on leisure for this given scenario are similar, but more erratic. For a clear indication of a luxury good at the given levels, one would expect both bounds to be negative. In fact, these would imply that families in the upper quartile on overall spending (i.e. $X = 0.75$) who spend up to 15% of their overall expenditure on leisure ($Y \in [0, 0.15]$) are very likely to spend even less on leisure, relatively, if they had a negative shock to overall spending ($X = 0.25$). Put differently, families should be more likely to spend only up to 15% of overall expenditure on leisure ($Y \in [0, 0.15]$) if they lie in the lower quartile in overall spending ($X = 0.25$) than families that lie in the upper quartile in overall spending ($X = 0.75$). The obtained results do reflect this circumstance at this level. In particular, the averages of the last 500 iterations of the general bounds are -0.032 and -0.31, implying that typically more families (up to 31%) in the lower quartile of overall spending ($X = 0.25$) spend only up to 15% of their overall on leisure compared to families in the upper quartile of overall spending.

The overall convergence for the leisure data is more erratic compared to food, which is rather illuminating for the purpose of this article. In particular, the main culprit for the poor performance is most likely the fact that the data is highly skewed towards zero in the leisure case, while the grid placed uniformly over the unit interval is too coarse to "measure" the behavior of the data around zero. Note that the method did not manage to sample enough monotone paths which correspond to the respective events of interest do to this reason, so that no bounds for monotonicity exist in this setting. In order to circumvent this, one can simply put a non-uniform grid on $[0, 1]$ which has more points close to zero and fewer points further away.

One can perform this exercise for different levels of y^* in order to gauge the behavior of the bounds at different quantiles. Figure 5 depicts the same exercise for $y = 0.25$, i.e. looking at families that spend up to a quarter of the overall expenditures on food and leisure. Here, the method managed to sample enough relevant monotonic paths in the leisure setting despite the coarser grid around zero. These bounds converge to some value in the more general bounds. One important fact to point out is the behavior of the solution path in the leisure case. At a certain point around the 1000 iterations mark a path is sampled which forces the solution path to jump up dramatically, i.e. a path which contains a lot of information for the given problem. However, the good news here is that over time the paths approach the previous level as more and more paths are drawn and add to the information from the one path. This is an important property to note, as it shows that the solution paths do not monotonically increase in the number of samples, and that even if paths are sampled which contain a lot of information for the respective program, will the method still converge to sharp bounds in the long run. In particular, it is the choice of the penalty term which forces to solution path to approach zero again after the jump.

The results are qualitatively similar to the ones for $y^* = 0.15$, which is not surprising as one is now comparing more families, i.e. all families that spend up to a quarter on food/leisure. The more families one compares, the less pronounced the effects become. The following table provides an overview of the results at the levels $y^* = 0.15, 0.25, 0.5, 0.75$, which shows exactly this. At $y = 0.75$ one is basically comparing all existing families, so that one does not obtain any effects, another sanity check for the method.

Overall, these estimation results are remarkably informative from a qualitative perspective. Recall that the instrumental variable model allows for general unobserved heterogeneity, in particular measurement error in the treatment variable X, which indicates that the ratio of information to noise in the data for answering these questions is rather high. These qualitative results not only corroborate the theoretical predictions for expenditure, but also the previous results obtained in Blundell, Chen & Kristensen (2007), Imbens & Newey (2009), and Song (2018). During their estimation process Imbens & Newey (2009) and Song (2018) assume a univariate and strictly monotonic production function $g(z, W)$ between X and W for all z and use a control

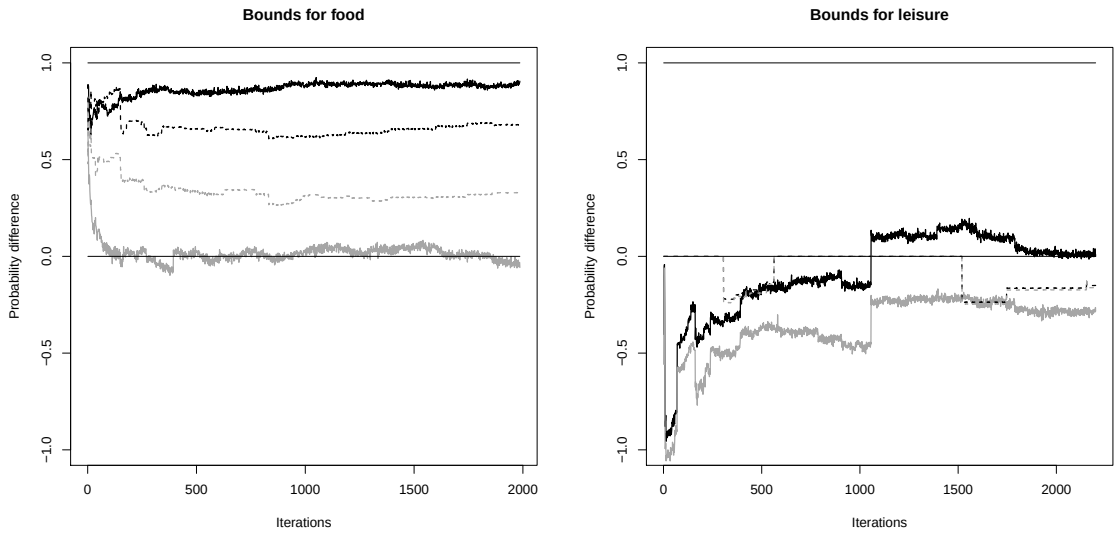

Figure 5: Convergence of upper (black) and lower (gray) bounds on $F_{Y(X=0.75)}(0.25) - F_{Y(X=0.25)}(0.25)$ for Y being the relative spending on food (left panel) and leisure (right panel). The solid lines indicate the bounds under no additional assumptions on the model, the dashed lines indicate the bounds under the additional assumption that Y is decreasing (food) or increasing (leisure) in overall expenditure X and that overall expenditure X is increasing in income of the head of the household Z. The penalty term is $\lambda = 1$ for an approximation of the unit interval by 17 points. At each iteration the method samples 25 new paths, of which 6 satisfy the respective monotonicity requirement.

	Food				Leisure			
y^*	Lower	Lower monotone	Upper monotone	Upper	Lower	Lower monotone	Upper monotone	Upper
0.15	0.26	0.50	0.78	0.84	−0.31	0	0	−0.032
0.25	0.008	0.31	0.67	0.88	−0.28	−0.18	−0.17	0.028
0.5	−0.33	0.0063	0.072	0.34	−0.34	−0.22	−0.19	0.09
0.75	−0.021	0	0	0.0075	−0.072	−0.046	0	0.0029

Table 2: Upper- and lower bounds for $F_{Y(X=0.75)}(y^*) - F_{Y(X=0.25)}(y^*)$ for different values y^*.

variable approach to estimate the production function h; Blundell, Chen & Kristensen (2007) estimate Engel curves semi-nonparametrically, imposing monotonicity in the second stage, and obtaining similar results. de Nadai & Lewbel (2016) work with an additively separable first stage and allow for the outcome Y to be measured with error. This is more general than what the proposed method can handle, which can only encompass measurement error in the dependent variable, but not the outcome. Nonetheless, their results are similar to the ones obtained here. In this sense these qualitative results are a "robustness check" for other non- or semiparametric approaches.

Moreover, this method makes it possible to gauge the identificatory content of monotonicity assumptions in the current model. In all cases is this content rather high. Imposing monotonicity *between the observables* makes the results much more clear-cut and in turn leads to rather strong implications in the cases considered. In this setting, monotonicity is a plausible assumption based on economic theory, but it is important to be aware of the strength of this assumption in other settings.

8 Conclusion

This article introduces a novel method for estimating bounds on functionals of the counterfactual distribution in instrumental variable models with general heterogeneity. Its main distinguishing feature is its applicability in practice, even for the most general models with continuous endogenous variables. The idea is to write the respective instrumental variable model as a system of counterfactual stochastic processes and to solve for an optimal probability measure on the paths of these processes subject to the constraint that the law of the joint processes induced by this probability measure replicates the observable distribution. The resulting optimization problem takes the form of an infinite dimensional (often linear) program on path spaces.

The main contribution of this article is to introduce the "sampling-of-paths" approach to solve these types of infinite dimensional programs. The underlying idea is to reduce the infinite dimensional program to a semi-infinite program (Anderson & Nash 1987) by only sampling a subset of the paths over which the program optimizes. Then an approximation of the (finite dimensional) state-space of the random variables leads to a finite dimensional program which can be solved efficiently. The main idea for reducing the infinite dimensional program to a semi-infinite dimensional one is to explicitly introduce randomness by sampling paths. This, in conjunction with large deviation results (Vapnik 1998) allows to obtain probabilistic approximation guarantees of the semi-infinite program to the infinite dimensional program. In particular, these guarantees imply a lower bound on the number of paths required for achieving a good approximation with high probability.

The focus of this article is on estimation, but large sample results are derived. In fact, the value functions corresponding to the counterfactual bounds are shown to be directional Hadamard differentiable (Bonnans & Shapiro 2013, Shapiro 1991). The directional Hadamard differentiability allows one to use the recently established subsampling results in the statistical literature (Dümbgen 1993, Fang & Santos 2018, Hong & Li 2018) to perform inference on each bound separable in practice. Since the bounds are univariate, one can then use established methods for obtaining confidence sets which cover the whole partially identified interval (Imbens & Manski 2004, Stoye 2009). Resampling methods might be too computationally expensive for general complex problems, in which case one should rely on the analytical expressions derived. In this regard, the proposed estimation method fits perfectly into the already established theory on inference in partially identified models. Together, they enable researchers to perform causal inference in the most general instrumental variable models.

A remaining challenge is to obtain an efficient data-driven method for choosing an appropriate penalty term λ for the practical optimization routine. This is a similar challenge to finding good penalty terms in high-dimensional regularized regression estimators, but more general, as the setting here is infinite dimensional in a counterfactual path space. Some heuristic guidelines can be given: one should choose the largest λ such that the "solution paths" of the sampling method converge to a fixed value after a "burn-in" period. If the solution path is "too erratic", then one should lower the value of λ. Formally establishing what "convergence", "burn-in period", and "too erratic" mean would not only solve this issue, but would open up potentially novel approaches for data-driven validation approaches in counterfactual settings. In particular, an analogue to the data-driven method for $\ell 1$-regularization in high-dimensional regression models as put forward in Belloni & Chernozhukov (2011) could be valuable.

The current practical implementation of the program works for univariate variables. Moreover, the only currently implemented additional nonparametric restriction which can be placed on the model is monotonicity. The program can straightforwardly be extended to higher dimensional settings, but runs into the curse of dimensionality as the stochastic processes become high-dimensional random fields. One standard way to

circumvent the curse of dimensionality is to introduce sparsity- and factor assumptions on the stochastic processes in a higher-dimensional setting. Furthermore, it is also imperative to allow for a wide variety of additional (non-) parametric assumptions in the model, like convexity, bounds, reflection processes, Slutsky-type conditions, first-passage times, martingale properties, etc. Furthermore, the current "sampling-and discarding" approaches are not efficient. From a computational perspective, it would be interesting to obtain more efficient resampling methods. One promising approach is to use ideas from sequential Monte-Carlo approaches (see e.g. Schweizer 2012), which need to be extended to the infinite dimensional path spaces considered here. Another idea is to directly use diffusion- or Levy-processes in the generation of the paths and use the Karhunen-Loève transform to generate a basis for the paths. An extension of the current program which already accommodates some of these additions is in the works.

Finally, the method allows researchers to compare the identificatory content of different frequently upheld (non-) parametric assumptions in instrumental variable models. For instance, one can run the method on a data set while only assuming continuity of the respective production functions. In general, this will lead to rather large counterfactual bounds. Then, one can run the same method again, but requiring the production of the first- or second stage to be monotone or convex, or bounded, etc. The bounds will then be rather different and give an indication of how much identificatory content the respective assumption has for the given model. In this sense, the method provides a general setting for evaluating the strength of different (non-) parametric assumptions on a model. A researcher can compare different structural assumptions and even dynamic assumptions within the same setting. This feature might potentially be also relevant for sensitivity arguments in structural models (e.g. Andrews, Gentzkow & Shapiro 2017, Bonhomme & Weidner 2018, Christensen & Connault 2019). In particular, it might be interesting to think about the strength of the instrument: intuitively, one would think that weaker instruments will in general lead to wider bounds.

References

Advani, A., Kitagawa, T. & Słoczyński, T. (2019), 'Mostly harmless simulations? using Monte Carlo studies for estimator selection', *Journal of Applied Econometrics* . forthcoming.

Aguiar, V., Allen, R. & Kashaev, N. (2019), Prices, Profits, and Production: Identification and counterfactuals, Technical report, University of Western Ontario.

Aliprantis, C. D. & Border, K. (2006), *Infinite Dimensional Analysis: a hitchhiker's guide*, Springer Science & Business Media.

Anastassiou, G. & Yu, X. (1992*a*), 'Convex and coconvex-probabilistic wavelet approximation', *Stochastic Analysis and Applications* **10**(5), 507–521.

Anastassiou, G. & Yu, X. (1992*b*), 'Monotone and probabilistic wavelet approximation', *Stochastic Analysis and Applications* **10**(3), 251–264.

Anderson, E., Lewis, A. & Wu, S.-Y. (1989), 'The capacity problem', *Optimization* **20**(6), 725–742.

Anderson, E. & Nash, P. (1987), *Linear programming in infinite dimensional spaces: Theory and applications*, Wiley.

Andrews, I., Gentzkow, M. & Shapiro, J. M. (2017), 'Measuring the sensitivity of parameter estimates to estimation moments', *The Quarterly Journal of Economics* **132**(4), 1553–1592.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**(434), 444–455.

Balke, A. & Pearl, J. (1994), Counterfactual probabilities: Computational methods, bounds and applications, *in* 'Proceedings of the Tenth international conference on Uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 46–54.

Balke, A. & Pearl, J. (1997), 'Bounds on treatment effects from studies with imperfect compliance', *Journal of the American Statistical Association* **92**(439), 1171–1176.

Bauer, H. (1996), *Probability Theory*, De Gruyter studies in Mathematics.

Belloni, A. & Chernozhukov, V. (2011), 'ℓ1-penalized quantile regression in high-dimensional sparse models', *The Annals of Statistics* **39**(1), 82–130.

Beresteanu, A., Molchanov, I. & Molinari, F. (2012), 'Partial identification using random set theory', *Journal of Econometrics* **166**(1), 17–32.

Billingsley, P. (1999), *Convergence of probability measures*, John Wiley & Sons.

Blundell, R., Chen, X. & Kristensen, D. (2007), 'Semi-nonparametric IV estimation of shape-invariant Engel curves', *Econometrica* **75**(6), 1613–1669.

Blundell, R. & Matzkin, R. L. (2014), 'Control functions in nonseparable simultaneous equations models', *Quantitative Economics* **5**(2), 271–295.

Bogachev, V. I. (2007), *Measure theory*, Vol. 2, Springer Science & Business Media.

Bonhomme, S. & Weidner, M. (2018), 'Minimizing sensitivity to model misspecification', *arXiv preprint arXiv:1807.02161* .

Bonnans, J. F. & Shapiro, A. (2013), *Perturbation analysis of optimization problems*, Springer Science & Business Media.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends® in Machine learning* **3**(1), 1–122.

Chen, X. (2007), 'Large sample sieve estimation of semi-nonparametric models', *Handbook of econometrics* **6**, 5549–5632.

Chernozhukov, V. & Hansen, C. (2005), 'An IV model of quantile treatment effects', *Econometrica* **73**(1), 245–261.

Chernozhukov, V., Lee, S. & Rosen, A. M. (2013), 'Intersection bounds: estimation and inference', *Econometrica* **81**(2), 667–737.

Chesher, A. (2003), 'Identification in nonseparable models', *Econometrica* **71**(5), 1405–1441.

Chesher, A. & Rosen, A. M. (2017), 'Generalized instrumental variable models', *Econometrica* **85**(3), 959–989.

Chiburis, R. C. (2010), 'Semiparametric bounds on treatment effects', *Journal of Econometrics* **159**(2), 267–275.

Choquet, G. (1954), Theory of capacities, *in* 'Annales de l'institut Fourier', Vol. 5, pp. 131–295.

Christensen, T. & Connault, B. (2019), 'Counterfactual sensitivity and robustness', *arXiv preprint arXiv:1904.00989* .

de Nadai, M. & Lewbel, A. (2016), 'Nonparametric errors in variables models with measurement errors on both sides of the equation', *Journal of Econometrics* **191**(1), 19–32.

Demuynck, T. (2015), 'Bounding average treatment effects: A linear programming approach', *Economics Letters* **137**, 75–77.

Dette, H., Hoderlein, S. & Neumeyer, N. (2016), 'Testing multivariate economic restrictions using quantiles: the example of Slutsky negative semidefiniteness', *Journal of Econometrics* **191**(1), 129–144.

d'Haultfœuille, X. & Février, P. (2015), 'Identification of nonseparable triangular models with discrete instruments', *Econometrica* **83**(3), 1199–1210.

Dümbgen, L. (1993), 'On nondifferentiable functions and the bootstrap', *Probability Theory and Related Fields* **95**(1), 125–140.

Fan, Y., Guerre, E. & Zhu, D. (2017), 'Partial identification of functionals of the joint distribution of "potential outcomes"', *Journal of econometrics* **197**(1), 42–59.

Fang, Z. & Santos, A. (2018), 'Inference on directionally differentiable functions', *The Review of Economic Studies* **86**(1), 377–412.

Florens, J.-P., Heckman, J. J., Meghir, C. & Vytlacil, E. (2008), 'Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects', *Econometrica* **76**(5), 1191–1206.

Folland, G. B. (2013), *Real analysis: modern techniques and their applications*, John Wiley & Sons.

Galichon, A. & Henry, M. (2011), 'Set identification in models with multiple equilibria', *The Review of Economic Studies* **78**(4), 1264–1298.

Garen, J. (1984), 'The returns to schooling: A selectivity bias approach with a continuous choice variable', *Econometrica* **52**(5), 1199–1218.

Giné, E. & Nickl, R. (2008), 'Uniform central limit theorems for kernel density estimators', *Probability Theory and Related Fields* **141**(3-4), 333–387.

Girosi, F. (1995), Approximation error bounds that use VC-bounds, *in* 'Proc. International Conference on Artificial Neural Networks, F. Fogelman-Soulie and P. Gallinari, editors', Vol. 1, pp. 295–302.

Gunsilius, F. (2019), Testability of instrument validity under continuous treatments, Technical report, MIT.

Hansen, L. P., Heaton, J. & Luttmer, E. G. (1995), 'Econometric evaluation of asset pricing models', *The Review of Financial Studies* **8**(2), 237–274.

Hausman, J. A. & Newey, W. K. (2016), 'Individual heterogeneity and average welfare', *Econometrica* **84**(3), 1225–1248.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).

Heckman, J. J. & Pinto, R. (2018), 'Unordered monotonicity', *Econometrica* **86**(1), 1–35.

Hess, H.-U. (1982), A Kuratowski approach to Wiener measure, *in* 'Measure Theory Oberwolfach 1981', pp. 336–346.

Hong, H. & Li, J. (2018), 'The numerical delta method', *Journal of Econometrics* **206**(2), 379–394.

Honoré, B. E. & Lleras-Muney, A. (2006), 'Bounds in competing risks models and the war on cancer', *Econometrica* **74**(6), 1675–1698.

Honoré, B. E. & Tamer, E. (2006), 'Bounds on parameters in panel dynamic discrete choice models', *Econometrica* **74**(3), 611–629.

Hu, K. Y. (1988), 'A generalization of Kolmogorov's extension theorem and an application to the construction of stochastic processes with random time domains', *The Annals of Probability* **16**(1), 222–230.

Imbens, G. (2019), Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics, Technical report, National Bureau of Economic Research.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475.

Imbens, G. W. & Manski, C. F. (2004), 'Confidence intervals for partially identified parameters', *Econometrica* **72**(6), 1845–1857.

Imbens, G. W. & Newey, W. K. (2009), 'Identification and estimation of triangular simultaneous equations models without additivity', *Econometrica* **77**(5), 1481–1512.

Kamat, V. (2017), 'Identification with latent choice sets: The case of the head start impact study', *arXiv:1711.02048* .

Kappen, H. J. (2007), An introduction to stochastic control theory, path integrals and reinforcement learning, *in* 'AIP conference proceedings', Vol. 887, AIP, pp. 149–181.

Karatzas, I. & Shreve, S. E. (1998), *Brownian motion and stochastic calculus*, Springer.

Kitagawa, T. (2009), Identification region of the potential outcome distributions under instrument independence, Technical report, UCL.

Kitamura, Y. & Stoye, J. (2018), 'Nonparametric analysis of random utility models', *Econometrica, forthcoming* .

Kluvánek, I. (1981), 'Remarks on bimeasures', *Proceedings of the American Mathematical Society* pp. 233–239.

Kuratowski, K. (1934), 'Sur une généralisation de la notion d'homéomorphie', *Fundamenta Mathematicae* **22**, 206–220.

Lafférs, L. (2015), 'Bounding average treatment effects using linear programming', *Empirical Economics* pp. 1–41.

Lai, H. & Wu, S.-Y. (1992), 'Extremal points and optimal solutions for general capacity problems', *Mathematical Programming* **54**(1), 87–113.

Manski, C. F. (1990), 'Nonparametric bounds on treatment effects', *The American Economic Review* **80**(2), 319–323.

Manski, C. F. (1997), 'Monotone treatment response', *Econometrica: Journal of the Econometric Society* pp. 1311–1334.

Manski, C. F. (2007), 'Partial identification of counterfactual choice probabilities', *International Economic Review* **48**(4), 1393–1410.

Manski, C. F. (2014), 'Identification of income–leisure preferences and evaluation of income tax policy', *Quantitative Economics* **5**(1), 145–174.

Matzkin, R. L. (2003), 'Nonparametric estimation of nonadditive random functions', *Econometrica* **71**(5), 1339–1375.

Mogstad, M., Santos, A. & Torgovitsky, A. (2018), 'Using instrumental variables for inference about policy relevant treatment effects', *Econometrica, forthcoming* .

Molinari, F. (2008), 'Partial identification of probability distributions with misclassified data', *Journal of Econometrics* **144**(1), 81–117.

Norets, A. & Tang, X. (2013), 'Semiparametric inference in dynamic binary choice models', *Review of Economic Studies* **81**(3), 1229–1262.

Parikh, N. & Boyd, S. (2014), 'Proximal algorithms', *Foundations and Trends® in Optimization* **1**(3), 127–239.

Pearl, J. (1995*a*), 'Causal diagrams for empirical research', *Biometrika* **82**(4), 669–688.

Pearl, J. (1995*b*), On the testability of causal models with latent and instrumental variables, *in* 'Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence', pp. 435–443.

Pucci de Farias, D. & Van Roy, B. (2004), 'On constraint sampling in the linear programming approach to approximate dynamic programming', *Mathematics of Operations Research* **29**(3), 462–478.

Robins, J. M. (1989), 'The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies', *Health service research methodology: a focus on AIDS* **113**, 159.

Rockafellar, R. T. (1974), *Conjugate duality and optimization*, SIAM.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of Educational Psychology* **66**(5), 688.

Russell, T. (2017), Sharp bounds on functionals of the joint distribution in the analysis of treatment effects, Technical report, University of Toronto.

Schennach, S. M. (2014), 'Entropic latent variable integration via simulation', *Econometrica* **82**(1), 345–385.

Schweizer, N. (2012), Non-asymptotic error bounds for sequential MCMC methods, PhD thesis, Rheinische Friedrich-Wilhelms Universität Bonn.

Shaikh, A. M. & Vytlacil, E. J. (2011), 'Partial identification in triangular systems of equations with binary dependent variables', *Econometrica* **79**(3), 949–955.

Shapiro, A. (1991), 'Asymptotic analysis of stochastic programs', *Annals of Operations Research* **30**(1), 169–186.

Song, S. (2018), Nonseparable triangular models with errors in endogenous variables, Technical report, University of Iowa.

Stoye, J. (2009), 'More on confidence intervals for partially identified parameters', *Econometrica* **77**(4), 1299–1315.

Tebaldi, P., Torgovitsky, A. & Yang, H. (2019), Nonparametric estimates of demand in the california health insurance exchange, Technical report, National Bureau of Economic Research.

Torgovitsky, A. (2015), 'Identification of nonseparable models using instruments with small support', *Econometrica* **83**(3), 1185–1197.

Torgovitsky, A. (2016), 'Nonparametric inference on state dependence with applications to employment dynamics'. University of Chicago working paper.

van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3, Cambridge university press.

van der Vaart, A. & Wellner, J. (2013), *Weak convergence and empirical processes: with applications to statistics*, Springer Science & Business Media.

van der Vaart, A. & Wellner, J. A. (2011), 'A local maximal inequality under uniform entropy', *Electronic Journal of Statistics* **5**(2011), 192.

Vapnik, V. (1992), Principles of risk minimization for learning theory, *in* 'Advances in neural information processing systems', pp. 831–838.

Vapnik, V. (1998), *Statistical learning theory. 1998*, Wiley, New York.

Wu, S.-Y., Fang, S.-C. & Lin, C.-J. (2001), 'Solving general capacity problem by relaxed cutting plane approach', *Annals of Operations Research* **103**(1), 193–211.

A Proofs

A.1 Proof of Proposition 1

Proof. The construction of the counterfactual processes $Y_x(w)$ and $X_z(w)$ on $[0,1]^{[0,1]}$ with laws $P_{Y|X=x}$ and $P_{X|Z=z}$ follows immediately from Kolmogorov's extension theorem (Karatzas & Shreve 1998, Theorem 2.2.2).

Together, the two laws $P_{Y(x)}$ and $P_{X(z)}$ generate the joint law $P_{[Y,X]^*(z)}$ as

$$P_{[Y,X]^*(z)}(A_y, A_x) = \int_{A_x} P_{Y(x)}(A_y) P_{X(z)}(dx), \tag{11}$$

which follows from the exclusion restriction: $P_{Y(x)}$ does not depend on Z. The fact that this joint law corresponds to a stochastic process $[Y,X]^*_z$ on $([0,1]\times[0,1])^{[0,1]}$ again follows from Kolmogorov's extension theorem. Intuitively, $[Y,X]^*_z$ is constructed as (Y_{X_z}, X_z), i.e. as the Cartesian product of the composed process Y_{X_z} and the process of the first stage X_z.

The next thing to show is that under Assumption 1, the processes $Y_x(w)$, $X_z(w)$, and $[Y,X]^*_z(w)$ are measurable as stochastic processes on the smaller space $D([0,1])$. Note that the measure space $(\mathcal{W}, \mathscr{B}_{\mathcal{W}})$ is large enough to accommodate all paths $Y_x(w)$, $X_z(w)$ satisfying Assumption 1. This follows from the fact that finite powers of the cardinality of the continuum are of the cardinality of the continuum themselves. Consider $X_z(w)$. Since $\mathcal{Z}$ is fixed, under Assumption 1 this stochastic process is progressively measurable with respect to its natural filtration $\{\mathscr{F}^X_z\}$, which is the smallest σ-algebra with respect to which X_z for non-random z is measurable. Indeed, by definition of the natural filtration, X_z is adapted to it. Moreover, under Assumption 1 every sample path is right-continuous. Then under Assumption 1 by Proposition 1.1.13 in Karatzas & Shreve (1998), it holds that X_z is also progressively measurable with respect to its natural filtration $\{\mathscr{F}^X_z\}$. The same argument shows measurability of $Y_x(w)$ with respect to the σ-field $\{\mathscr{F}^Y_x\}$ under Assumption 1.

Now focus on the joint process $[Y,X]^*_z$. It is right-continuous under Assumption 1 on its codomain $[0,1]^2$, as the Cartesian product of two continuous functions on the real line is continuous by a projection argument, so that its paths lie in the respective Skorokhod space $D([0,1])$ with codomain $[0,1]^2$. Therefore, and since z is nonrandom, Proposition 1.1.13 in Karatzas & Shreve (1998) implies that X_z is also progressively measurable with respect to its natural filtration $\{\mathscr{F}^{[Y,X]}_z\}$. Now the key for showing measurability of $[Y,X]^*_z$ is to show that the random process $X_z(w)$ is a stopping time for the process $Y_{X_z}(w)$, as this is how part of $[Y,X]^*_z$ is constructed under the exclusion restriction: $[Y,X]^*_z \equiv (Y_{X_z}, X_z)$. But this fact follows from the exclusion restriction: the path $Y_{X_z}(w)$ only depends on the current position x of $X_z(w)$ for each fixed z and not on the actual path $X_z(w)$, which means that Y_x does not depend on future values of x and x does not depend on future values of Y_x, or more formally, it holds that $\{X_z \leq x\} \in \mathscr{F}^Y_x$. Therefore, it follows again from Proposition 1.2.18 in Karatzas & Shreve (1998) that $Y_{X_z}(w)$ is measurable with respect to the σ-field $\mathscr{F}^{[Y,X]}_z$ of all events prior to z.

Finally, the independence restriction $Z \perp\!\!\!\perp W$ implies that one can compare the properties of the stochastic process $[Y,X]_z$ induced by the observable distribution $P_{Y,X|Z=z}$ to the stochastic process $[Y,X]^*_z(w)$ corresponding to $P_{[Y,X]^*(z)}$. □

A.2 Proof of Theorem 1

The proof of the theorem requires the following lemma, which bounds the approximation of the paths Y_x and X_z by the wavelet basis.

Lemma 1. *The wavelet operators W_κ acting on a càdlàg function $f \in L^2(\mathbb{R})$ through*

$$W_\kappa(f)(x) \coloneqq \sum_{j=-\infty}^{\infty} \langle f, \varphi_{\kappa j}\rangle \varphi_{\kappa j}(x),$$

where

$$\langle f, \varphi_{\kappa j}\rangle \coloneqq \int_{-\infty}^{\infty} f(t)\varphi_{\kappa j}(t)dt$$

and φ is the hat-function wavelet basis defined in the main text, satisfy

$$|W_\kappa(f)(x) - f(x)| \leq \omega'_f(2^{-\kappa+1}) \qquad \text{for all } x \in \mathbb{R} \text{ and } k \in \mathbb{Z},$$

where ω'_f is the extended modulus of continuity.

Proof. The proof is analogous to the proof of Theorem 1 in Anastassiou & Yu (1992*b*), but for càdlàg functions instead of continuous functions. The hat-function wavelet basis satisfies $\sum_{j=-\infty}^{\infty} \varphi(x-j) = 1$ on $\mathbb{R}$ (Anastassiou & Yu 1992*b*), so that for a square integrable function $f(x)$ we have

$$W_\kappa(f)(x) - f(x) = 2^{\kappa/2} \sum_{j=-\infty}^{\infty} \left[\langle f, \varphi_{\kappa j}\rangle - 2^{-\kappa/2} f(x)\right] \varphi(2^\kappa x - j).$$

The hat-function wavelet basis also satisfies $\int_{-\infty}^{\infty} \varphi(u-j)du = 1$, $j \in \mathbb{Z}$ (Anastassiou & Yu 1992*b*). Based on this, and by a change of variables, we have

$$\begin{aligned}
\langle f, \varphi_{\kappa j}\rangle - 2^{-\kappa/2} f(x) &= 2^{\kappa/2} \int_{-\infty}^{\infty} f(t)\varphi(2^\kappa t - j)dt - 2^{-\kappa/2} f(x) \\
&= 2^{-\kappa/2} \int_{-\infty}^{\infty} f(2^{-\kappa}u)\varphi(u-j)du - 2^{-\kappa/2} f(x) \\
&= 2^{-\kappa/2} \int_{-\infty}^{\infty} [f(2^{-\kappa}u) - f(x)]\varphi(u-j)du.
\end{aligned}$$

Since the support of φ is $[-1,1]$ and $\varphi \geq 0$, we have for $2^{-\kappa}(-1+j) \leq x \leq 2^{-\kappa}(1+j)$

$$|\langle f, \varphi_{\kappa j}\rangle - 2^{-\kappa/2} f(x)| = 2^{-\kappa/2} \left| \int_{-1+j}^{1+j} [f(2^{-\kappa}u) - f(x)]\varphi(u-j)du \right| \leq 2^{-\kappa/2} \int_{-1+j}^{1+j} |f(2^{-\kappa}u) - f(x)|\varphi(u-j)du.$$

The integral on the right is a Riemann integral, so that we can write it in terms of the upper Darboux integral of the function as

$$2^{-\kappa/2} \int_{-1+j}^{1+j} [f(2^{-\kappa}u) - f(x)]\varphi(u-j)du = 2^{-\kappa/2} \inf_{\Pi[2^{-\kappa+1}]} \sum_{i=1}^{\eta} (u_i - u_{i-1}) \sup_{s \in [u_{i-1}, u_i)} |f(2^{-\kappa}s - f(x)|\varphi(s-j),$$

where $\Pi[2^{k-1}]$ denotes a partition $2^{-\kappa}(-1+j) = u_0 < u_1 < \ldots < u_{\eta-1} < u_\eta = 2^{-\kappa}(1+j)$ of the interval $[2^{-\kappa}(-1+j), 2^{-\kappa}(1+j)]$ of length $2^{\kappa-1}$, and where the infimum is taken over all partitions $\Pi[2^{k-1}]$ of

arbitrary η.

Then we can bound the upper Darboux integral by

$$\begin{aligned}
&2^{-\kappa/2} \inf_{\Pi[2^{-\kappa+1}]} \sum_{i=1}^{\eta} (u_i - u_{i-1}) \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)|\varphi(s-j) \\
\leq& 2^{-\kappa/2} \inf_{\Pi[2^{-\kappa+1}]} \sum_{i=1}^{\eta} (u_i - u_{i-1}) \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)| \sup_{s\in[u_{i-1},u_i)} \varphi(s-j) \\
\leq& 2^{-\kappa/2} \inf_{\Pi[2^{-\kappa+1}]} \sum_{i=1}^{\eta} (u_i - u_{i-1}) \left\{ \max_{1\leq i\leq\eta} \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)| \right\} \sup_{s\in[u_{i-1},u_i)} \varphi(s-j) \\
\leq& 2^{-\kappa/2} \inf_{\Pi[2^{-\kappa+1}]} \max_{1\leq i\leq\eta} \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)| \sum_{i=1}^{\eta} (u_i - u_{i-1}) \sup_{s\in[u_{i-1},u_i)} \varphi(s-j).
\end{aligned}$$

The first inequality follows from the fact that $\varphi \geq 0$, the second inequality follows by choosing the largest difference over all intervals for a given partition, and the third by the fact that $\max_{1\leq i\leq\eta} \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)|$ does not depend on i anymore.

The following is the crucial argument for bounding the integral. By the fact f is càdlàg, it only has finitely many jumps that exceed any $\sigma > 0$, so that in order to minimize the expression, any partition $\Pi[2^{-\kappa+1}]$ has to partition the interval in such a way that the intervals $[u_{i-1}, u_i)$ line up with the points x where $f(x)$ jumps, i.e. is only continuous on the right with limit on the left. That is, if x is a point of discontinuity of f in the considered interval, then any partition that minimizes $\max_{1\leq i\leq\eta} \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)|$, needs to have $x = u_i$ for some i. Now since $\varphi(\cdot - j)$ is continuous on the given interval, the upper Darboux integral converges to the Riemann integral as the partition becomes finer. This still holds if we only consider partitions which have $x = u_i$ for points of discontinuities of f, since the points of discontinuity are countable and hence of (Lebesgue-) measure zero—note that Lebesgue-and Riemann integral coincide. Therefore, the infimum over all partitions $\Pi[2^{\kappa-1}]$ for the upper Darboux integral

$$\sum_{i=1}^{\eta} (u_i - u_{i-1}) \sup_{s\in[u_{i-1},u_i)} \varphi(s-j)$$

coincides with the infimum over all partitions $\Pi[2^{-\kappa}]$ which minimize

$$\max_{1\leq i\leq\eta} \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)|,$$

so that we can write

$$\begin{aligned}
&2^{-\kappa/2} \inf_{\Pi[2^{-\kappa+1}]} \max_{1\leq i\leq\eta} \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)| \sum_{i=1}^{\eta} (u_i - u_{i-1}) \sup_{s\in[u_{i-1},u_i)} \varphi(s-j) \\
=& 2^{-\kappa/2} \inf_{\Pi[2^{-\kappa+1}]} \max_{1\leq i\leq\eta} \sup_{s\in[u_{i-1},u_i)} |f(2^{-\kappa}s - f(x)| \inf_{\Pi[2^{-\kappa+1}]} \sum_{i=1}^{\eta} (u_i - u_{i-1}) \sup_{s\in[u_{i-1},u_i)} \varphi(s-j) \\
=& 2^{-\kappa/2} \omega_f'(2^{-\kappa+1}) \int_{-1+j}^{1+j} (u_i - u_{i-1})\varphi(u-j) \\
=& 2^{-\kappa/2} \omega_f'(2^{-\kappa+1}) \int_{-\infty}^{\infty} (u_i - u_{i-1})\varphi(u-j)
\end{aligned}$$

$$= 2^{-\kappa/2}\omega_f'(2^{-\kappa+1})$$

by the definition of $\omega_f'(2^{-\kappa+1})$, the fact that $\varphi(\cdot - j)$ is Riemann and hence Darboux integrable, and the fact that φ integrates to 1 as argued above.

By the fact that the support of φ is $[-1, 1]$, it holds that

$$\begin{aligned}
|W_\kappa(f)(x) - f(x)| &\leq 2^{\kappa/2} \sum_{j:2^\kappa x - j \in [-1,1]} \left| \langle f, \varphi_{\kappa j} \rangle - 2^{-\kappa/2} \right| \varphi(2^\kappa x - j) \\
&= \omega_f'(2^{-\kappa+1}) \sum_{j=-\infty}^{\infty} \varphi(2^\kappa x - j) \\
&= \omega_f'(2^{-\kappa+1}).
\end{aligned}$$

□

Proof of Theorem 1. The proof is split into two parts. The first part uses Lemma 1 to reduce the complexity of the paths Y_x and X_z via their wavelet approximations $\tilde{Y}_x$ and $\tilde{X}_z$, which will make the second part easier to handle. In fact, the second part consists of considering an approximated version of the programs (3) as an M-estimator in the sense of van der Vaart & Wellner (2013, chapter 2.14) and derives a lower bound on the number of paths l to be sampled by using the associated large deviation results from empirical process theory.

Part 1:

Focus on the constraint first. Using the notation from (4), one can write it as

$$\begin{aligned}
&\left\| F_{Y,X|Z=z}(\cdot,\cdot) - \int \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) P_W(dw) \right\|^2_{L^2([0,1]^2)} \\
= &\left\| \int [\mathbb{1}_{[0,\cdot]\times[0,\cdot]}(Y_x(w), X_z(w)) - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w))] \frac{dP_W}{dP_0}(w) \right\|^2_{L^2([0,1]^2)},
\end{aligned}$$

where the second line follows from the population constraint in (3) and the $L^2([0,1]^2)$ is taken with respect to Lebesgue measure over y, x. The existence of the Radon-Nikodym derivative is guaranteed by Assumption 2. The goal is to bound this distance, so that one can work with the wavelet basis in part 2. So write this difference as

$$\begin{aligned}
&\left\| \int \left[\mathbb{1}_{[0,\cdot]\times[0,\cdot]}(Y_x(w), X_z(w)) - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \right] P_W(dw) \right\|^2_{L^2([0,1]^2)} \\
= &\left\| \int \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(Y_x(w), X_z(w)) P_W(dw) - \mathcal{S}(Y_x(w)), \cdot, \alpha) \mathcal{S}(X_z(w)), \cdot, \alpha) P_W(dw) \right\|^2_{L^2([0,1]^2)} \\
&+ \left\| \int \mathcal{S}(Y_x(w)), \cdot, \alpha) \mathcal{S}(X_z(w)), \cdot, \alpha) P_W(dw) - \int \mathcal{S}(\tilde{Y}_x^\kappa(w)), \cdot, \alpha) \mathcal{S}(\tilde{X}_z^\kappa(w)), \cdot, \alpha) P_W(dw) \right\|^2_{L^2([0,1]^2)} \\
&+ \left\| \int \left[\mathcal{S}(\tilde{Y}_x^\kappa(w)), \cdot, \alpha) \mathcal{S}(\tilde{X}_z^\kappa(w)), \cdot, \alpha) - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \right] P_W(dw) \right\|^2_{L^2([0,1]^2)},
\end{aligned}$$

Here, $\mathcal{S}(Y_x(w)), y, \alpha)$ is the logistic approximation of $\mathbb{1}_{[0,y]}(Y_x(w))$ on $[0,1]$ of the form

$$\mathcal{S}(Y_x(w), y, \alpha) \coloneqq \frac{1}{1+\exp\left(-\alpha\left(y - Y_x(w) + \alpha^{-1/2}\right)\right)},$$

and analogously for $\mathcal{S}(X_z(w), x, \alpha)$. The reason for this simple form is that $Y_x(w) \in [0,1]$ for all x and $X_z(w) \in [0,1]$ for all z and all w, so that one only needs to take care of the upper bounds y and x, respectively. $\mathcal{S}(Y_x(w), y, \alpha) \to \mathbb{1}_{[0,y]}(Y_x(w))$ as $\alpha \to \infty$, so that the first and third term are of the order $o(1)$.

$\mathcal{S}(Y_x(w)), \cdot, \alpha)\mathcal{S}(X_z(w)), \cdot, \alpha)$ is infinitely often differentiable, so that by the mean-value theorem it holds for every $y, x \in [0,1]$ and $w \in [0,1]$ that

$$\left|\mathcal{S}(Y_x(w)), y, \alpha)\mathcal{S}(X_z(w)), x, \alpha) - \mathcal{S}(\tilde{Y}_x^\kappa(w)), y, \alpha)\mathcal{S}(\tilde{X}_z^\kappa(w)), x, \alpha)\right| \leq C\left|Y_x(w) - \tilde{Y}_x^\kappa(w)\right|\left|X_z(w) - \tilde{X}_z^\kappa(w)\right|,$$

where $C < +\infty$ is a positive constant. By Lemma 1 it holds

$$\left|Y_x(w) - \tilde{Y}_x^\kappa(w)\right| \leq \omega'_{Y_x(w)}(2^{\kappa+1}), \qquad \left|X_z(w) - \tilde{X}_z^\kappa(w)\right| \leq \omega'_{X_z(w)}(2^{\kappa+1}) \qquad \text{for all } w \in [0,1],$$

so that

$$\left|\mathcal{S}(Y_x(w)), y, \alpha)\mathcal{S}(X_z(w)), x, \alpha) - \mathcal{S}(\tilde{Y}_x^\kappa(w)), y, \alpha)\mathcal{S}(\tilde{X}_z^\kappa(w)), x, \alpha)\right| \leq C\omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1})$$

for all $w \in [0,1]$. By Hölder's inequality, Assumption 1, and the fact that P_W is a probability measure

$$\int C\omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1})P_W(dw) \leq C \sup_{w\in[0,1]} \omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1}).$$

Therefore, integrating both sides of the above inequality gives

$$\int \left|\mathcal{S}(Y_x(w)), y, \alpha)\mathcal{S}(X_z(w)), x, \alpha) - \mathcal{S}(\tilde{Y}_x^\kappa(w)), y, \alpha)\mathcal{S}(\tilde{X}_z^\kappa(w)), x, \alpha)\right| P_W(dw) \\ \leq C \sup_{w\in[0,1]} \omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1}).$$

The right hand side of the last inequality does not depend on x and y anymore, so that taking the L^2-norm with respect to Lebesgue measure on $[0,1]^2$ on both sides gives

$$\left\|\int \mathcal{S}(Y_x(w)), \cdot, \alpha)\mathcal{S}(X_z(w)), \cdot, \alpha)P_W(dw) - \int \mathcal{S}(\tilde{Y}_x^\kappa(w)), \cdot, \alpha)\mathcal{S}(\tilde{X}_z^\kappa(w)), \cdot, \alpha)P_W(dw)\right\|_{L^2([0,1]^2)} \\ \leq C \sup_{w\in[0,1]} \omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1}). \quad (12)$$

The first term can hence be bounded by

$$\left\|\int \left[\mathbb{1}_{[0,\cdot]\times[0,\cdot]}(Y_x(w), X_z(w)) - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w))\right] P_W(dw)\right\|_{L^2([0,1]^2)} \\ \leq C \sup_{w\in[0,1]} \omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1}) + o(1), \quad (13)$$

where the $o(1)$ terms captures the logistic approximation of the indicator function.

The same decomposition can be achieved for the objective function. Fix A_y and A_x throughout and for some arbitrary $P_W \in \mathscr{P}^*(\mathcal{W})$ consider

$$\left\| \int \left[K(Y_x(w), A_y, A_x) - \tilde{K}(\tilde{Y}_x^\kappa(w), A_y, A_x) \right] P_W(dw) \right\|^2_{L^\infty([0,1])}$$

A bound on this follows exactly as the bound for the constraint for each $P_W \in \mathscr{P}^*(\mathcal{W})$:

$$\left\| \int \left[K(Y_x(w), A_y, A_x) - \tilde{K}(\tilde{Y}_x^\kappa(w), A_y, A_x) \right] P_W(dw) \right\|^2_{L^\infty([0,1])} \leq C \sup_{w\in[0,1]} \omega'_{Y_x(w)}(2^{\kappa+1}) + o(1),$$

where the $o(1)$ appears by a potential approximation of K by $\tilde{K}$ in the case where K is an indicator function.

Part 2:

Therefore, from now on, the idea is to consider $\frac{dP_W}{dP_0}$ as an M-estimator of the problem

$$\begin{aligned} &\underset{\substack{\frac{dP_w}{dP_0} \\ P_0, P_W \in \mathscr{P}^*(\mathcal{W})}}{\text{minimize/maximize}} \quad \int \tilde{K}(\tilde{Y}_x^\kappa(w), A_y, A_x) \frac{dP_W}{dP_0}(w) P_0(dw) \\ &\text{s.t.} \quad \left\| F_{Y,X|Z=z} - \int \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) P_0(dw) \right\|^2_{L^2([0,1]^2)} \leq \varepsilon^* \end{aligned} \tag{14}$$

for some fixed κ and $\varepsilon^* > 0$ and to obtain a large deviations result which will result in a lower bound on the number of samples l needed for an approximation of the value function of the infinite programs (3) by the value function of the semi-infinite programs (4).

Rewrite the programs (14) as penalized programs. For this, the key is bounding the constraint.

$$\begin{aligned} &\left\| F_{Y,X|Z=z} - \int \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) P_0(dw) \right\|^2_{L^2([0,1]^2)} \leq \varepsilon^* \\ \iff &\left\| F_{Y,X|Z=z} - \int \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) P_0(dw) \right\|_{L^2([0,1]^2)} \leq \sqrt{\varepsilon^*}, \end{aligned}$$

so that one can work with the L^2-norm instead of the squared L^2-norm. It holds that

$$\begin{aligned} &\left\| F_{Y,X|Z=z} - \int \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) P_0(dw) \right\|_{L^2([0,1]^2)} \\ &= \left(\int_0^1 \int_0^1 \left[F_{Y,X|Z=z}(y,x) - \int \mathbb{1}_{[0,y]\times[0,x]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) P_0(dw) \right]^2 dydx \right)^{1/2} \\ &= \left(\int_0^1 \int_0^1 \left[\int \left\{ F_{Y,X|Z=z}(y,x) - \mathbb{1}_{[0,y]\times[0,x]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) \right\} P_0(dw) \right]^2 dydx \right)^{1/2} \\ &\leq \int \left(\int_0^1 \int_0^1 \left[F_{Y,X|Z=z}(y,x) - \mathbb{1}_{[0,y]\times[0,x]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) \right]^2 dydx \right)^{1/2} P_0(dw) \end{aligned}$$

$$= \int \left\| F_{Y,X|Z=z} - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) \right\|_{L^2([0,1]^2)} P_0(dw),$$

where the second to last line follows by Minkowski's inequality for integrals (Folland 2013, Theorem 6.19). Therefore, there must exist some $\varepsilon_0 > 0$ such that if

$$\int \left\| F_{Y,X|Z=z} - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) \right\|_{L^2([0,1]^2)} P_0(dw) \leq \sqrt{\varepsilon_0},$$

then the original constraint is less than or equal to ε^*.

So from now on, the original constraint will be replaced by the last expression for ε_0. This is done to write the optimization problem (14) in the form of an empirical process with respect to P_0 in order to apply the large deviation results derived in van der Vaart & Wellner (2013, chapter 2.14). So rewrite (14) in Lagrangian form (Rockafellar 1974, chapter 4) as

$$\begin{aligned} \min_{\substack{\frac{dP_W}{dP_0} \\ P_0, P_W \in \mathscr{P}^*(\mathcal{W})}} & \int \tilde{K}(\tilde{Y}_x^\kappa(w), A_y, A_x) \frac{dP_W}{dP_0}(w) P_0(dw) \\ & + \lambda_0 \int \left\| F_{Y,X|Z=z} - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) \frac{dP_W}{dP_0}(w) \right\|_{L^2([0,1]^2)} P_0(dw), \end{aligned}$$

where λ_0 is the penalty term corresponding to $\sqrt{\varepsilon_0}$. One can write the objective function in empirical process form used in van der Vaart & Wellner (2013) as

$$\begin{aligned} & P_0 m_h \coloneqq \int m_{h(w)} P_0(dw) \\ \equiv & \int \left[\tilde{K}(\tilde{Y}_x^\kappa(w), A_y, A_x) h(w) + \lambda_0 \left\| F_{Y,X|Z=z} - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w)) h(w) \right\|_{L^2([0,1]^2)} \right] P_0(dw), \end{aligned} \tag{15}$$

where $h(w) \coloneqq \frac{dP_W}{dP_0}(w)$ is a simpler notation for the Radon-Nikodym density. The same reasoning as above holds for the empirical counterpart, i.e. the programs (4). One can rewrite them in the same way as above in order to obtain the empirical counterpart of (15) as

$$\begin{aligned} & \mathbb{P}_{0,l} m_h \coloneqq \frac{1}{l} \sum_{i=1}^{l} m_{h(i)} \\ \equiv & \frac{1}{l} \sum_{i=1}^{l} \left[\tilde{K}(\tilde{Y}_x^\kappa(i), A_y, A_x) h(i) + \lambda_0 \left\| F_{Y,X|Z=z} - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(i), \tilde{X}_z^\kappa(i)) h(i) \right\|_{L^2([0,1]^2)} \right], \end{aligned} \tag{16}$$

where $\mathbb{P}_{0,l}$ denotes the empirical law for l samples of paths.

So the key now is to bound the uniform entropy numbers[31] of the set of functions $\mathcal{M}_h$, i.e. the set of all admissible functions m_h. This is achieved by splitting this function into its subparts.

[31] The uniform entropy number relative to the L^2-norm of a class $\mathcal{F}$ of functions is defined as $\sup_Q \log N(\varepsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))$, where the supremum runs over all probability measures with finite support, and where $N(\varepsilon, \mathcal{F}, L^2(Q))$ denotes the ϵ-covering number of the class $\mathcal{F}$ with respect to the L^2-norm of Q, see van der Vaart & Wellner (2013, p. 84).

Consider

$$\Lambda(w) \coloneqq \mathbb{1}_{[0,y]\times[0,x]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w))\frac{dP_W}{dP_0}(w).$$

To bound its uniform entropy number, note first that $\mathbb{1}_{[0,x]\times[0,y]}$ are of VC dimension 3 since Y, X, and Z are defined on $[0,1]$. Furthermore, recall that $\tilde{Y}_x^\kappa(w)$ and $\tilde{X}_z^\kappa(w)$ are composed of finitely many transformations (shifted and scaled) of the hat-function φ, so that *for fixed* κ the sets of all approximations $\tilde{Y}_x^\kappa$ and X_z^κ have finite VC dimension. Lemma 2.6.18 (vii) in van der Vaart & Wellner (2013) then implies that the set $\mathcal{C}_\mathbb{1}(\kappa)$ of functions of the form $\mathbb{1}_{[0,y]\times[0,x]}(\tilde{Y}_x^\kappa(i), \tilde{X}_z^\kappa(i))$ has finite VC dimension. In particular, the VC dimension is a function of κ, so that

$$d_{VC}(\mathcal{C}_\mathbb{1}(\kappa)) = V(\kappa) < +\infty.$$

Theorem 2.6.4 in van der Vaart & Wellner (2013) then allows to bound the covering number by the VC dimension, so that

$$N(\varepsilon, \mathcal{C}_\mathbb{1}(\kappa), L^2(Q)) \leq CV(\kappa)(4e)^{V(\kappa)}\left(\frac{1}{\varepsilon}\right)^{2(V(\kappa)-1)}$$

for every probability measure Q on $[0,1]$ and a universal constant C.

Under Assumption 2, the Radon-Nikodym densities are β-Hölder continuous and bounded by $C_{RN} < +\infty$. Then the monotonicity of the logarithmic function in combination with Corollary 2.7.2 in van der Vaart & Wellner (2013) implies that there exists a constant C such that

$$\log N(\varepsilon, C^\beta_{C_{RN}}([0,1]), L^2(Q)) \leq \log N_{[]}(2\varepsilon, C^\beta_{C_{RN}}([0,1]), L^2(Q)) \leq C\left(\frac{1}{\varepsilon}\right)^{1/\beta}$$

for every $\varepsilon > 0$ and probability measure Q on $[0,1]$. $C^\beta_{C_{RN}}$ denotes the space of all β-Hölder continuous functions bounded by C_{RN} and $N_{[]}$ denotes the bracketing number, see van der Vaart & Wellner (2013, p. 84).

By the fact that multiplication of functions is Lipschitz continuous on compact subsets of the real line (van der Vaart & Wellner 2013, Example 2.10.8), it follows from Theorem 2.10.20 in van der Vaart & Wellner (2013) that

$$\sup_Q \log N(\varepsilon C_{RN}, \Lambda, L^2(Q)) \leq \sup_Q \log N(\varepsilon, \mathcal{C}_\mathbb{1}(\kappa), L^2(Q)) + \sup_Q \log N(\varepsilon C_{RN}, C^\beta_{C_{RN}}([0,1]), L^2(Q)),$$

as an envelope function in the term on the left and the second term on the right is the function which is $C_{RN}\mathbb{1}_{[0,1]}$ on $[0,1]$ and an envelope function for the first term on the right hand side is $1_{[0,1]}$. Since $-\log(\varepsilon) = o(\varepsilon^{-1/\beta})$ for any $\beta > 0$, it follows that

$$\sup_Q \log N(\varepsilon C_{RN}, \Lambda, L^2(Q)) \leq C \sup_Q \log N(\varepsilon C_{RN}, C^\beta_{C_{RN}}([0,1]), L^2(Q))$$

for some constant C that also depends on C_{RN}.

From the reverse triangle inequality it follows that the L^2-norm is Lipschitz continuous with Lipschitz constant 1. Therefore, for fixed $F_{Y,X|Z=z}$ and λ_0, the set of functions

$$\overline{\Lambda} \coloneqq \lambda_0 \left\| F_{Y,X|Z=z} - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^\kappa(w), \tilde{X}_z^\kappa(w))h(w) \right\|_{L^2([0,1]^2)}$$

also has a bounded entropy number. In fact, by Theorem 2.10.20 in van der Vaart & Wellner (2013) it holds

that

$$\sup_Q \log N(\varepsilon C_{RN}, \overline{\Lambda}, L^2(Q)) \leq C \sup_Q \log N(\varepsilon C_{RN}, C^\beta_{C_{RN}}([0,1]), L^2(Q)),$$

where C is a uniform constant that also depends on κ, β, C_{RN}, $F_{Y,X|Z=z}$, and λ_0.

The same type of approximation holds for

$$\Lambda_K(w) \coloneqq \tilde{K}(\tilde{Y}^\kappa_x(w), A_y, A_x)\frac{dP_W}{dP_0}(w).$$

To see this, note that by Assumption 3 the kernel $\tilde{K}$ is either an indicator function, in which case it has finite VC dimension, or is Lipschitz continuous. The case of an indicator function is perfectly analogous to the case above, so focus on the case where $\tilde{K}$ is Lipschitz. Then, as before, for fixed κ, the set of functions $\tilde{Y}^\kappa_x$ has finite VC dimension. Since $\tilde{K}$ is Lipschitz continuous with Lipschitz constant $L < +\infty$, it holds by Theorem 2.10.20 in van der Vaart & Wellner (2013) for the class $\tilde{\mathcal{K}}$ containing these Lipschitz continuous objective functions that

$$\sup_Q \log N(\varepsilon L, \tilde{\mathcal{K}}, L^2(Q)) \leq \sup_Q \log N(\varepsilon, \tilde{\mathcal{Y}}^\kappa_x, L^2(Q)),$$

because an envelope function for $\tilde{\mathcal{Y}}^\kappa_x$ is $\mathbb{1}_{[0,1]}$, as all paths $\tilde{Y}^\kappa_x$ have codimension $[0,1]$.

Applying Theorem 2.10.20 in van der Vaart & Wellner (2013) once again while noting that the multiplication operator is Lipschitz continuous on compact intervals gives

$$\sup_Q \log N(\varepsilon C_{RN}, \Lambda_K, L^2(Q)) \leq \sup_Q \log N(\varepsilon L, \tilde{\mathcal{K}}, L^2(Q)) + \sup_Q \log N(\varepsilon C_{RN}, C^\beta_{C_{RN}}([0,1]), L^2(Q)),$$

and so

$$\sup_Q \log N(\varepsilon C_{RN}, \Lambda_K, L^2(Q)) \leq C_K \sup_Q \log N(\varepsilon C_{RN}, C^\beta_{C_{RN}}([0,1]), L^2(Q))$$

by the same reasoning as before; C_K is a uniform constant.

Therefore, putting everything together and using Theorem 2.10.20 in van der Vaart & Wellner (2013) once more, it holds that

$$\sup_Q \log N(\varepsilon C_{RN}, \mathcal{M}_h, L^2(Q)) \leq C \sup_Q \log N(\varepsilon C_{RN}, C^\beta_{C_{RN}}([0,1]), L^2(Q)) \leq C\left(\frac{1}{\varepsilon}\right)^{1/\beta}, \tag{17}$$

where C is some constant that depends on $\kappa, \beta, \lambda, C_{RN}$, and $F_{Y,X|Z=z}$.

Now the goal is to bound $P^*(\sup_{m_h \in \mathcal{M}_h} |\mathbb{P}_{l,0} m_h - P_0 m_h| > t)$ for small $t > 0$, where P^* denotes outer probability to avoid measurability issues. It holds

$$\begin{aligned} &P^*\left(\sup_{m_h \in \mathcal{M}_h} |\mathbb{P}_{l,0} m_h - P_0 m_h| > t\right) \\ =&P^*\left(\sqrt{l}\sup_{m_h \in \mathcal{M}_h} |\mathbb{P}_{l,0} m_h - P_0 m_h| > \sqrt{l}t\right). \end{aligned}$$

The tail bound derived in Theorem 2.14.10 of van der Vaart & Wellner (2013) implies that for every $\delta > 0$ and $t > 0$

$$P^*\left(\sqrt{l}\sup_{m_h \in \mathcal{M}_h} |\mathbb{P}_{l,0} m_h - P_0 m_h| > \sqrt{l}t\right) \leq L \exp\left(D(\sqrt{l}t)^{U+\delta}\right)\exp(-2lt^2), \tag{18}$$

where $U \coloneqq \frac{6\beta-1}{\beta(2\beta+1)}$ and the constants L and D depends on the constant C, β, and δ. Since $\beta > \frac{1}{2}$, there exists a small enough $\delta > 0$ such that $U + \delta < 2$. Therefore, one can bound the term further by

$$L \exp\left(D(\sqrt{l}t)^{U+\delta}\right) \exp(-2lt^2) \leq \bar{C} \exp(-\bar{D}lt^2),$$

for constants $0 < \bar{C} < +\infty$ and (potentially small) $0 < \bar{D} < 2$.

Now apply inversion. It follows that

$$P^* \left(\sup_{m_h \in \mathcal{M}_h} |\mathbb{P}_{l,0} m_h - P_0 m_h| > \sqrt{\frac{\log\left(\frac{\bar{C}}{\rho}\right)}{\bar{D}l}} \right) \leq \rho \qquad \text{for } \rho \in (0,1),$$

and so

$$P^* \left(\sup_{m_h \in \mathcal{M}_h} |\mathbb{P}_{l,0} m_h - P_0 m_h| \leq \sqrt{\frac{\log\left(\frac{\bar{C}}{\rho}\right)}{\bar{D}l}} \right) \geq 1 - \rho \qquad \text{for } \rho \in (0,1).$$

Therefore, with probability at least $1 - \rho$ it holds that

$$\sup_{m_h \in \mathcal{M}_h} |\mathbb{P}_{l,0} m_h - P_0 m_h| \leq \sqrt{\frac{\log\left(\frac{\bar{C}}{\rho}\right)}{\bar{D}l}}.$$

In order to put parts 1 and 2 together write (3) in penalized form as

$$\min_{\substack{\frac{dP_W}{dP_0} \\ P_0, P_W \in \mathscr{P}^*(\mathcal{W})}} \int K(Y_x(w), A_y, A_x) \frac{dP_W}{dP_0}(w) P_0(dw) + \lambda_0 \int \left\| F_{Y,X|Z=z} - \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(Y_x(w), X_z(w)) \frac{dP_W}{dP_0}(w) \right\|_{L^2([0,1]^2)} P_0(dw).$$

Denote by V^* and V_* the value function of the maximization and minimization of (3), respectively, and by $\tilde{V}^*_\kappa$ and $\tilde{V}_{*,\kappa}$ their analogues for (15). Then by part 1 and the fact that the bound is independent of P_W it holds that

$$\max\{|V^* - \tilde{V}^*_\kappa|, |V_* - \tilde{V}_{*,\kappa}|\} \leq C \sup_{w \in [0,1]} \omega'_{Y_x(w)}(2^{\kappa+1}) \omega'_{X_z(w)}(2^{\kappa+1}) + o(1).$$

Denote by $\tilde{V}^*_{l,\kappa}$ and $\tilde{V}_{*,l,\kappa}$ the value functions of (16). Then by part 2 it holds that

$$\max\{|\tilde{V}^*_\kappa - \tilde{V}^*_{l,\kappa}|, |\tilde{V}_{*,\kappa} - \tilde{V}_{*,l,\kappa}|\} \leq \sqrt{\frac{\log\left(\frac{\bar{C}}{\rho}\right)}{\bar{D}l}}$$

with probability of at least $1 - \rho$. Therefore, by the triangle inequality it holds that

$$\max\{|V^* - \tilde{V}^*_{l,\kappa}|, |V_* - \tilde{V}_{*,l\kappa}|\} \leq \max\{|V^* - \tilde{V}^*_\kappa|, |V_* - \tilde{V}_{*,\kappa}|\} + \max\{|\tilde{V}^*_\kappa - \tilde{V}^*_{l,\kappa}|, |\tilde{V}_{*,\kappa} - \tilde{V}_{*,l,\kappa}|\}$$

$$\leq C \sup_{w\in[0,1]} \omega'_{Y_x(w)}(2^{\kappa+1})\omega'_{X_z(w)}(2^{\kappa+1}) + o(1) + \sqrt{\frac{\log\left(\frac{\bar{C}}{\rho}\right)}{\bar{D}l}}$$

with probability of at least $1-\rho$. □

A.3 Proof of Proposition 2

Proof. Define the constraint correspondence $\tilde{\mathcal{C}} : \mathcal{F}_{Y,X|Z=z} \times (0,1) \to \hat{\mathscr{P}}^*_{\mathcal{W}}$

$$\tilde{\mathcal{C}}(F_{Y,X,Z},\varepsilon) \coloneqq$$
$$\left\{ \hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W}) : \left\| F_{Y,X|Z=z}(y,x) - \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{[0,\cdot]\times[0,\cdot]}(\tilde{Y}^\kappa_x(i), \tilde{X}^\kappa_z(i)) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^2)} \leq \varepsilon \right\}$$

as in the main text, which is non-empty by Assumption 4. Note that under Assumption 2 $\hat{\mathscr{P}}^*(\mathcal{W})$ is compact in the weak topology since the Radon-Nikodym derivatives are uniformly bounded by $C_{RN} < +\infty$. To see this, note that the set of all probability measures $\hat{P}_W$ on the finite index set $\{i\}_{i=1,\dots,l}$ is tight. Hence, by Prokhorov's Theorem (Karatzas & Shreve 1998, Theorem 2.4.7), it is relatively compact in the weak topology. Under Assumption 2 the set $\hat{\mathscr{P}}^*(\mathcal{W})$ is a closed subset of this set in the weak topology, as all Radon-Nikodym derivatives are uniformly bounded by a weak inequality.

Also note that $\tilde{\mathcal{C}}(F_{Y,X|Z},\varepsilon)$ is continuous. Its upper hemicontinuity can be shown by a sequencing argument as $\hat{\mathscr{P}}^*(\mathcal{W})$ is compact in the weak topology (Aliprantis & Border 2006, Theorem 17.20): let $\{\hat{F}_{Y,X|Z;n}\}_{n\in\mathbb{N}} \in \mathcal{F}_{Y,X|Z=z}$ be a sequence such that $\|\hat{F}_{Y,X|Z=z;n} - F_{Y,X|Z=z}\|_{L^\infty([0,1]^2)} \to 0$ for all $z \in [0,1]$ as $n \to \infty$ and $\{\hat{P}_{W;n}\}_{n\in\mathbb{N}} \in \hat{\mathscr{P}}^*(\mathcal{W})$ be a sequence satisfying $\hat{P}_{W;n} \in \tilde{\mathcal{C}}(F_{Y,X|Z;n},\varepsilon)$. Under Assumption 4, the continuity of the L^2-norm, and the weak inequality in $\tilde{\mathcal{C}}$, it holds that $\{\hat{P}_{W;n}\}_{n\in\mathbb{N}}$ must converge to an element $\hat{P}^*_W \in \hat{\mathscr{P}}^*(\mathcal{W})$ which satisfies the constraint, so that this sequence has a limit point in $\tilde{\mathcal{C}}(F_{Y,X|Z},\varepsilon)$.

As for lower hemicontinuity, fix some sequence $\{\hat{F}_{Y,X|Z;n}\}_{n\in\mathbb{N}}$. Pick some arbitrary $\hat{P}_W \in \tilde{\mathcal{C}}(F_{Y,X|Z},\varepsilon)$. Under Assumption 4, there exists a ball $\mathcal{B}_r$ around $F_{Y,X|Z}$ such that the constraint correspondence is not empty there. If there is only one element in the constraint correspondence for $F_{Y,X|Z}$, then lower hemicontinuity is trivially fulfilled. So suppose there are at least two solutions $\hat{P}_W$ and $\hat{P}'_W$ in $\tilde{\mathcal{C}}(F_{Y,X|Z},\varepsilon)$. Since $\mathcal{B}_r$ is convex, the L^2-norm is continuous, and the inequality in $\tilde{\mathcal{C}}$ is weak, it holds that for large enough $n \in \mathbb{N}$ there exist also two solutions $\hat{P}_{W,n}, \hat{P}'_{W,n} \in \tilde{\mathcal{C}}(\hat{F}_{Y,X|Z;n},\varepsilon)$ which are close to $\hat{P}_W$ and $\hat{P}'_W$, respectively, so that lower hemicontinuity is fulfilled.

The objective function defined by the kernel K is continuous under Assumption 3 on the graph of $\tilde{\mathcal{C}}(F_{Y,X|Z},\varepsilon)$ for fixed $\varepsilon > 0$. Therefore, the Berge Maximum Theorem (Aliprantis & Border 2006, Theorem 17.31) implies that the value functions $\tilde{V}^*_{l,\kappa}(F_{Y,X|Z})$ and $\tilde{V}_{*,l,\kappa}(F_{Y,X|Z})$ are continuous. Now under the Glivenko-Cantelli Theorem (van der Vaart 2000, Theorem 19.1), it holds that $\|\hat{F}_{Y,X|Z=z;n} - F_{Y,X|Z=z}\|_{L^\infty([0,1]^2)} \to 0$ for all $z \in [0,1]$ as $n \to \infty$ almost surely. Now by the continuity of the value functions, one can apply the Continuous Mapping Theorem (van der Vaart 2000, Theorem 18.11), which implies that the value functions converge almost surely. □

A.4 Proof of Proposition 3

Proof. The goal is to apply the functional delta method (Shapiro 1991, Theorem 2.1), for which the key is to prove directional Hadamard differentiability of the value functions $\tilde{V}_{*,l,\kappa}(F_{Y,X|Z})$ and $\tilde{V}^*_{l,\kappa}(F_{Y,X|Z})$. Focus

on the minimization problem in (4), as maximization is perfectly analogous. It is convenient to write (4) in terms of the notation in Bonnans & Shapiro (2013):

$$\begin{aligned} &\min_{\hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W})} K(\hat{P}_W) \\ &\text{subject to} \qquad G(\hat{P}_W, F_{Y,X|Z=z}) \in [0, \varepsilon], \end{aligned} \tag{19}$$

where

$$G(\hat{P}_W, F_{Y,X|Z=z}) \coloneqq \left\| F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{[0,\cdot]\times[0,\cdot]\times[0,\cdot]}(\tilde{Y}_x^{\kappa}(i), \tilde{X}_z^{\kappa}(i)) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^2)}.$$

Theorem 4.26 in Bonnans & Shapiro (2013) then provides conditions for directional Hadamard differentiability of $\tilde{V}_*$. The main condition is a restricted version of Robinson's constraint qualification:

$$0 \in \text{int}\{G(\hat{P}_W, F_{Y,X|Z=z}) + \delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X|Z=z})\}, \tag{20}$$

where $F_{Y,X|Z=z}$ is the data-generating process in the population, $\hat{P}_W$ is an argument which minimizes (19), $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X|Z=z})$ is the directional derivative of $G(\cdot,\cdot)$ in its first argument at this optimal $\hat{P}_W$ with respect to any direction in $\hat{\mathscr{P}}^*(\mathcal{W})$, and "int" denotes the topological interior of a set. Since $G(\cdot,\cdot)$ is a functional mapping to $[0,\varepsilon]$, (20) reduces to showing that $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X|Z=z})$ can take negative and positive values of size less than or equal to $-\varepsilon$ for some direction $\hat{P}'_W \in \hat{\mathscr{P}}^*(\mathcal{W})$.

For this, calculate $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X|Z=z})$ as[32]

$$\begin{aligned} \delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X|Z=z}) &= \lim_{t\to 0} t^{-1}\left[\left\| F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W + t\hat{P}_1}{d\hat{P}_0}\right\|^2_{L^2} - \left\| F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}\right\|^2_{L^2}\right] \\ &= \lim_{t\to 0}\int t^{-1}\left[\left(F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W + td\hat{P}_1}{d\hat{P}_0}\right)^2 - \left(F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}\right)^2\right] dydx \end{aligned}$$

Multiplying out the square terms and simplifying (using the linearity of the Radon-Nikodym derivative $\frac{d\hat{P}_w + td\hat{P}_1}{d\hat{P}_0} = \frac{d\hat{P}_w}{d\hat{P}_0} + t\frac{d\hat{P}_1}{d\hat{P}_0}$) gives

$$\begin{aligned} &\lim_{t\to 0}\int t^{-1}\left[\left(F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W + td\hat{P}_1}{d\hat{P}_0}\right)^2 - \left(F_{Y,X|Z=z} - \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}\right)^2\right] dydx \\ &= \lim_{t\to 0}\int t^{-1}\left[-2F_{Y,X|Z=z}\frac{t}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0} + 2t\left(\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}\frac{1}{l}\sum_{i}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}\right) + \frac{t^2}{l^2}\left(\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}\right)^2\right] dydx. \end{aligned}$$

Since $F_{Y,X|Z=z}$, $\frac{t}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}$, and $\frac{t}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}$ take values in the unit interval, Lebesgue's Dominated

[32] In the following derivation, the shorthand notation $\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}$, $\|\cdot\|_{L^2}$ and $\int F_{Y,X|Z=z}dydx$ is used to denote $\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}_{[0,y]\times[0,x]}(\tilde{Y}_x^{\kappa}(i), \tilde{X}_z^{\kappa}(i))\frac{d\hat{P}_W}{d\hat{P}_0}(i)$, $\|\cdot\|_{L^2([0,1]^2)}$, and $\int_{[0,1]^2} F_{Y,X|Z=z}(y,x)dydx$, respectively.

Convergence Theorem permits interchanging the limit and the integral, which after simplification gives

$$\lim_{t\to 0}\int t^{-1}\left[-2F_{Y,X|Z=z}\frac{t}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}+2t\left(\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}\frac{1}{l}\sum_{i}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}\right)+\frac{t^2}{l^2}\left(\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}\right)^2\right]dydx$$
$$=\int 2\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}\left[F_{Y,X|Z=z}-\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_W}{d\hat{P}_0}\right]dydx.$$

To see that (20) can be satisfied, ignore the "weight term" $2\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}\frac{d\hat{P}_1}{d\hat{P}_0}$ in the last expression for a second and note that the term in brackets is the constraint in (4), which can be rewritten as $F_{Y,X|Z=z}-\hat{P}_W(\tilde{Y}_x^\kappa\in[0,y],\tilde{X}_z^\kappa\in[0,x])$. Therefore, without the "weight term", which is always non-negative, the expression for $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}),F_{Y,X|Z=z})$ takes the form of a comparison between these two probabilities in terms of second-order stochastic dominance. Therefore, if either the cumulative distribution function induced by the probability $\hat{P}_W(\tilde{Y}_x^\kappa\in[0,y],\tilde{X}_z^\kappa\in[0,x])$ second-order stochastically dominates $F_{Y,X|Z=z}$ by some large enough value $\zeta>0$—which makes the above integral negative—or vice versa, then the requirement for Robinson's condition is fulfilled. This can happen in programs (4) under Assumption 4. Therefore, the restricted version of Robinson's constraint qualification is satisfied by the program (4).

Furthermore, note that under Assumption 4 the program (4) admits an optimal solution for $F_{Y,X|Z=z}+tF'_{Y,X|Z=z}$ where $F_{Y,X|Z=z},F'_{Y,X|Z=z}\in\mathcal{F}_{Y,X|Z=z}$ when t is small enough such that $F'_{Y,X|Z=z}\in\mathcal{B}_r\subset\mathcal{F}_{Y,X|Z=z}$, which is the second requirement needed in order to apply Theorem 4.26 in Bonnans & Shapiro (2013). Therefore, Theorem 4.26 in combination with Proposition 4.47 in Bonnans & Shapiro (2013) implies that the value function $\tilde{V}_{*,l,\kappa}(F_{Y,X|Z=z})$ is directionally Hadamard differentiable in every direction $F'_{Y,X|Z=z}\in\mathcal{F}_{Y,X|Z=z}$ and that the Hadamard derivative $d\tilde{V}_{*,F_{Y,X|Z=z}}(F'_{Y,X|Z=z})$ takes the form

$$d\tilde{V}_{*,F_{Y,X|Z=z}}(F'_{Y,X|Z=z})=\delta_{F_{Y,X|Z=z}}L(\hat{P}_1,\lambda(\hat{P}_W),F_{Y,X|Z=z})(F'_{Y,X|Z=z}),$$

where $L(\cdot,\cdot,\cdot)$ denotes the Lagrangian of the program (4), $\lambda(\hat{P}_1)$ denotes the respective Lagrange multiplier (which is unique for given $\hat{P}_1$ by Proposition 4.47 in Bonnans & Shapiro 2013), and $\delta_{F_{Y,X|Z=z}}$ is the directional derivative of L in its third argument in direction $F'_{Y,X|Z=z}$ at $F_{Y,X|Z=z}$.

Now by Donsker's theorem (van der Vaart 2000, Theorem 19.3), it holds that $\sqrt{n}(\hat{F}_{Y,X|Z=z;n}-F_{Y,X|Z=z})\Rightarrow\mathbb{G}_{F_{Y,X|Z=z}}$, where $\mathbb{G}_{F_{Y,X|Z=z}}$ is a Brownian bridge with covariance function

$$\mathrm{Cov}_{\mathbb{G}_{F_{Y,X|Z=z}}}=F_{Y,X|Z=z}(\min\{y,y'\},\min\{x,x'\},\min\{z,z'\})-F_{Y,X|Z=z}(y,x)F_{Y,X|Z=z'}(y',x')$$

for all $(y,x,z),(y',x',z')\in[0,1]^3$. Therefore, applying the functional delta method (Shapiro 1991, Theorem 2.1) directly yields that

$$\sqrt{n}(\hat{\tilde{V}}_*(\hat{F}_{Y,X|Z=z;n})-\tilde{V}_{*,l,\kappa}(F_{Y,X|Z=z}))\Rightarrow d\tilde{V}_{*,F_{Y,X|Z=z}}(\mathbb{G}_{F_{Y,X|Z=z}}).$$

The same argument holds for $\hat{\tilde{V}}^*$. □