Bounds in continuous instrumental variable models

Florian F Gunsilius*

MIT

September 11, 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 negate these undeniable advantages—particularly in general instrumental variable models with continuous variables. Therefore, 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. This new 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, it shows that 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.

*I want to thank Susanne Schennach, Toru Kitagawa, Francesca Molinari, and Whitney Newey for very helpful discussions. I also thank the audiences at Boston College, Brandeis, Brown, Cornell, Georgetown, Johns Hopkins, MIT, the University of Michigan, the University of Pennsylvania as well as the 3rd Econometrics workshop at Notre Dame for very helpful comments. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University. This work supersedes part of my Brown University job market paper titled "Testability and bounds in continuous instrumental variable models". All errors are mine.

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 applications which limits a broader use in general models, as noted by 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 many practical settings of interest, in particular when the endogenous variable has many points in its support or even is continuous. 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 sharp upper and lower 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 in the form of response profiles of agents. This also permits the researcher to include assumptions about particular individual behavior into the model, improving transparency of the required assumptions. The key for this is a novel representation of the general 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 efficiently 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 by using large deviation theory (Vapnik 1998) for the purpose of function approximation (Girosi 1995). In particular, a researcher can use these results to gauge how good the approximation of the semi-infinite dimensional- to the infinite dimensional program is for a given sample of basis functions. This approximation result could potentially be also of use in other settings where general unobserved heterogeneity is to be approximated (Bonhomme, Lamadon & Manresa 2017). Furthermore, the sampling approach allows researchers to gauge the identification content of (nonparametric) structural assumptions

in their model, as one can sample from the basis under additional shape 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 potential scope of 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 shrink drastically and become much 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 gauge the identificatory strength of frequently made (non-) parametric assumptions on the model.

The focus in this article lies on (practical) 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 in closed-form. 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.

[1]This application also serves as a sanity check in place of an impossible Monte-Carlo simulations exercise in a "counterfactual space". 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 that the method is faulty. This application is actually more challenging than a hypothetical 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 something like this should be possible with *any* approach. The fact that this method does provide informative bounds is a testament to its potential usefulness.

[2]One challenge is the computational burden of resampling methods in this setting, which can be substantial for general problems. It is therefore imperative to derive the proper large sample results for the bounds in these settings and use the analytical results.

[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 (their $\mathcal{V} \subset \mathbb{R}^d$ for some dimension d). In contrast, the proposed optimization method is defined over an infinite dimensional path space. It would be interesting to examine the connection between these settings further.

2 Existing literature

The setting for the method introduced in this article is that of a nonseparable triangular model, the most general form of an instrumental variable 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 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 assumptions or other structural assumptions besides continuity and does not rely on a control function approach.[4] Through the setting of nonseparable triangular models, 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 welfare analysis (Dette, Hoderlein & Neumeyer 2016, Hausman & Newey 2016).[5]

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 which 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); in contrast, Chesher & Rosen (2017) also estimate the the model parameters such as the production functions (an infinite dimensional quantity).[6]

This article is not the first to propose an optimization approach for obtaining bounds in partial identification models (Chiburis 2010, Demuynck 2015, Honoré & Tamer 2006, Honoré & Lleras-Muney 2006, Manski 2007, Molinari 2008, Norets & Tang 2013, Lafférs 2015, Kamat 2017, Torgovitsky 2016, Kitamura & Stoye 2018, Mogstad, Santos & Torgovitsky 2018), but it

[4] Even the continuity requirement can be relaxed by working in Skorokhod space instead of Hölder spaces, see the discussion in the main section.

[5] 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.

[6] This fact was pointed out by Francesca Molinari.

is the first to present an infinite dimensional optimization problem for doing so. The infinite dimensionality is key in order to solve the most complex models—the ones with a continuous endogenous variable—theoretically as well as practically. Furthermore, in combination with the novel "sampling of paths"- approach it allows researchers to solve complex but finite models by considering random variables with many points 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) and Manski (1990) 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. 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 is also convenient as it 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 this setting, in the case where endogenous variable is discrete. Manski (1997) derives sharp bounds on causal effects in general models under monotonicity assumptions.

3 The theoretical method and its properties

This section introduces the theoretical method and provides intuitive explanations for the introduced concepts. Randomized controlled trials with imperfect compliance will serve as a running example, as the introduced concepts turn out to have a natural interpretation in this setting.

3.1 Intuitive explanation of the estimation problem and explanation of the stochastic process representation

The proposed method deals with nonseparable triangular models of the following form:

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

Here, Y is the outcome variable, X is the endogenous treatment in the sense that it depends on the unobservable variable W, and Z is an instrument satisfying the relevance condition $Z \not\perp\!\!\!\perp X$ as well as the validity restriction $Z \perp\!\!\!\perp W$:[7]

Assumption 1 (Instrument validity and relevance)**.** *In model* (1) *the instrument Z is valid and relevant, i.e. it holds that (i) $Z \perp\!\!\!\perp W$ (validity) and that $Z \not\perp\!\!\!\perp X$ (relevance).*

W can be of arbitrary and even of infinite dimension, just like Y, X, and Z. The production functions h and g are unknown.[8]

[7] $Z \perp\!\!\!\perp W$ means that Z is independent of W, i.e. $P_{Z,W} = P_Z P_W$. Note that 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. Also, this article neither addresses the important issue of weak identification (e.g. Kleibergen 2002, Stock, Wright & Yogo 2002) nor the testability of instrument validity (e.g. Pearl 1995, Kitagawa 2015). The instrument must be valid and relevant for the proposed method to work. For a recent work which obtains bounds under a relaxed independence assumption in the binary treatment setting, consider Masten & Poirier (2018).

[8] Unless stated otherwise, the assumption is that all variables have support in a Polish space, i.e. a separable, complete, and metrizable space. In particular, Euclidean spaces are Polish. The support $\mathcal{S}$ of a measure μ is

3.1.1 The endogeneity problem

The main quantity of interest in this model is the counterfactual law $P_{Y(X)}$.[9] Note that this law in general does not coincide with the observable law $P_{Y|X}$, because of the endogeneity problem. To see this, write

$$P_{Y(X)} = \int_{\mathcal{W}} P_{Y|X,W=w} P_W(dw) \tag{2}$$

using model (1), where $\mathcal{W}$ is the support of the law P_W and where $P_{Y|X,W=w}$ is the conditional law of Y given X while setting $W = w$. Now, since W is unobservable, the only distribution in the data which provides correct information on the data-generating process is $P_{Y,X|Z}$ (respectively: $P_{Y,X,Z}$). To see this, write $P_{Y,X|Z}$ as

$$\begin{aligned} P_{Y,X|Z} &= \int_{\mathcal{W}} P_{Y|X,Z,W=w} P_{X|Z,W=w} P_W(dw) \\ &= \int_{\mathcal{W}} P_{Y|X,W=w} P_{X|Z,W=w} P_W(dw) \end{aligned}$$

because $Z \perp\!\!\!\perp Y|(X,W)$ by model (1) (Balke & Pearl 1994, p. 50).[10]

Now, if X were actually exogenous, i.e. $X \perp\!\!\!\perp W$, it would hold that $P_{X|Z,W} = P_{X|Z}$, so that $P_{Y(X)}$ would coincide with the observable $P_{Y|X}$, because

$$\begin{aligned} P_{Y|X} = \frac{P_{Y,X|Z}}{P_{X|Z}} &= \frac{\int_{\mathcal{W}} P_{Y|X,W=w} P_{X|Z,W=w} P_W(dw)}{P_{X|Z}} \\ &= \frac{P_{X|Z} \int_{\mathcal{W}} P_{Y|X,W=w} P_W(dw)}{P_{X|Z}} = P_{Y(X)}. \end{aligned}$$

As soon as X is endogenous, i.e. $X \not\perp\!\!\!\perp W$, the above line of reasoning does not work anymore, so that $P_{Y(X)} \neq P_{Y|X}$ in general. Therefore, without any structural assumptions on the model one will in general be only able to identify bounds on functionals of $P_{Y(X)}$.

---

defined by two properties: first, $\mu(\mathcal{S}^c) = 0$, where $\mathcal{S}^c$ denotes the complement of $\mathcal{S}$. Second, for any open set G with $G \cap \mathcal{S} \neq \emptyset$ it holds $\mu(G \cap \mathcal{S}) > 0$, see Aliprantis & Border (2006, p. 441). Throughout, the support of a random variable will be denoted by scripture notation, i.e. $\mathcal{X}$ is the support of X and $\mathcal{X}_z$ denotes the support of the conditional probability measure $P_{X|Z=z}$ for $z \in \mathcal{Z}$.

[9] This notation is analogous to the counterfactual notation in causal inference introduced in (Rubin 1974). Note that X can be arbitrary and need not be binary. Moreover, the notation $P_{Y(X=x)}$ will mean that the treatment has *exogenously* fixed to take the value $x \in \mathcal{X}$. Analogously for all other counterfactual distributions.

[10] To see this, note that (1) incorporates the *exclusion restriction* that Z has no direct influence on Y, which follows from the fact that h is not a function of Z. This, in combination with the independence restriction $Z \perp\!\!\!\perp W$ implies that Z is independent of Y once one has conditioned on the direct effects of X and W on Y.

3.1.2 The key idea: Representing the instrumental variable model as a system of stochastic processes

This endogeneity problem has sparked a large literature in partial identification in the setting where X is binary, as pointed out in the previous section. Unfortunately, the binary- and even discrete cases are much simpler to handle in terms of their complexity than the general continuous case. In fact, in the binary case, i.e. where Y, X, Z take values in $\{0, 1\}$, the problem of bounding the counterfactual probability $P_{Y(X)}$ can be solved by finding the solution to a simple linear program using the response variables W, as in Balke & Pearl (1994) and Balke & Pearl (1997).

One realization in the binary case is that W *indexes* the respective production functions $g(Z, \cdot)$ and $h(X, \cdot)$ from model (1). In the binary case, W possess only eight realizations as there are four possible functions mapping Z to X and X to Y in each case. For instance, w_1^1 corresponds to the function g mapping the realization $Z = 0$ to $X = 0$ and the realization $Z = 1$ to $X = 0$ (the never takers), w_2^1 corresponds to the function g mapping $Z = 0$ to $X = 0$ and $Z = 1$ to $X = 1$ (the compliers), w_3^1 corresponds to the function g mapping $Z = 0$ to $X = 1$ and $Z = 1$ to $X = 0$ (the defiers), and w_4^1 corresponds to the function g mapping $Z = 0$ to $X = 1$ and $Z = 1$ to $X = 1$ (the always takers). An analogous set-up holds for $w_1^2, \ldots, w_4^2$ in terms of realizations of Y and X.[11]

Generalizing this idea, it is not hard to see that the cardinality of the support $\mathcal{W}$ of W will necessarily be of the rate $n^m + q^n$, where m is the number of points in $\mathcal{Z}$, n is the number of points in $\mathcal{X}$, and q is the number of points in $\mathcal{Y}$. Therefore, already a simple generalization to the case where all variables can take on 3 values each will, without further assumptions, lead to $2 \cdot 3^3 = 54$ values for W, composed of 27 values indexing the functions in the first- and second stage, respectively. Cheng & Small (2006) provide an identification result in this setting by circumventing this issue: they make monotonicity assumptions on the production functions g and h in order to bring down the cardinality from 27 to 4 in each stage of the model again. The setting of continuous Y, X, and Z exacerbates the problem even more.

The key realization in the continuous setting therefore is that $g(Z, U)$ and $h(X, V)$ induce stochastic processes $Y_X(v)$ and $X_Z(u)$ with random domains X and Z described by the counterfactual measures $P_{(X,Z)}$ and $P_{(Y,X)}$, respectively. This is captured in the following

Proposition 1 (Stochastic process representation)**.** *Model* (1) *under the independence restriction is equivalent to a system of counterfactual stochastic processes* Y_X *and* X_Z *on random domains*

[11] The superscript w_1^1 serves to distinguish the values $w \in W$ which correspond to the first stage (in which case the superscript is 1) of the nonseparable triangular model from those corresponding to the second (in which case the superscript is 2). In the alternative notation of model (1) where the unobservable in the first stage is usually called U and the unobservable in the second stage is V, the values w_i^1 would correspond to values u_i while w_i^2 would correspond to some v_i.

$\mathcal{X}, \mathcal{Z}$ on measure spaces $(\mathcal{X}, \mathscr{B}_X), (\mathcal{Z}, \mathscr{B}_Z)$ and with corresponding laws

$$P_{(Y,X)} = \int P_{Y(X=x)} P_X^*(dx) \qquad \text{and} \qquad P_{(X,Z)} = \int P_{X(Z=z)} P_Z^*(dz),$$

respectively. P_X^* and P_Z^* are the counterfactual measure for exogenous X and Z and in general need not coincide with the observable P_X and P_Z. However, the independence restriction $Z \perp\!\!\!\perp W$ implies $P_Z = P_Z^*$. Furthermore, the exclusion restriction implies that these laws induce the counterfactual law

$$\begin{aligned} P_{(Y,X,Z)}(A_y, A_x, A_z) &= \int_{A_x} P_{Y(X=x)}(A_y) P_{(X,Z)}(dx, A_z) \\ &= \int_{A_z} \int_{A_x} P_{Y(X=x)}(A_y) P_{X(Z=z)}(dx) P_Z^*(dz), \end{aligned}$$

for Borel sets A_y, A_x, A_z, which corresponds to a joint counterfactual process $[Y, X]_Z^*$ with random domain $\mathcal{Z}$. The fact that $P_Z = P_Z^*$ allows to compare the counterfactual process $[Y, X]_Z^*$ to the stochastic process $[Y, X]_Z$ induced by the observable joint law $P_{Y,X,Z}$.[12]

Intuitively, Proposition 1 states that the second stage of model (1) corresponds to a stochastic process Y_X where the randomness stems from some general probability space $(\Omega, \mathcal{B}_\Omega, P)$, where $\mathscr{B}_\Omega$ denotes the Borel σ-algebra on the space Ω. 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 $(\mathcal{W}, \mathscr{B}_\mathcal{W}, P_W)$.[13] Then elements $w \in \mathcal{W}$ index paths $Y_X(w)$ of the respective stochastic process. Similarly, the first stage of model (1) corresponds to a stochastic process X_Z where the randomness also stems from the unobservable confounder W. Elements $w \in \mathcal{W}$ index paths $X_Z(w)$ of the respective stochastic process. Not surprisingly, these counterfactual stochastic processes turn out to have a natural interpretation in terms of randomized controlled trials with imperfect compliance, generalizing the defier-, complier-, always taker-, never taker distinction from Angrist, Imbens & Rubin (1996), as explained in the following

Running example (1). *Suppose data from a (fictive) randomized controlled trial estimating the efficacy of a (continuous) treatment X on some (continuous) outcome Y is available. The researchers conducting this trial are worried about imperfect compliance among the participants as they could not perfectly enforce that the participants actually took the randomly assigned amount.*

[12] Notice the use of the *joint* laws $P_{(Y,X)}$, $P_{(X,Z)}$, $P_{(Y,X,Z)}$, and $P_{Y,X,Z}$ in this construction. Throughout, the focus will be on these and not the corresponding conditional laws $P_{Y(X)}$, $P_{X(Z)}$, $P_{(Y,X)(Z)}$, $P_{Y,X|Z}$, etc. which simplifies the mathematical notation in the proofs. The approach in this article works with either, the conditional- and the unconditional laws. The randomness of X and Z is highlighted through the notation of the stochastic processes Y_Z, X_Z, $[Y, X]_Z$, etc. where a capital Z is chosen to denote the randomness in Z. In this respect, the notation $Y_{X=x}$ means that the stochastic process is conditioned on the value of the point x.

[13] 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).

As a robustness check to their standard results they want to use instrumental variable estimation methods, in which they use the initially assigned amount as the (continuous) instrument Z. Due to the fact that too high amounts of the treatment might actually be detrimental for different reasons, the researchers do not want to uphold a monotonicity assumption, neither in the relation between Y and X, nor in the relation between X and Z, but they are comfortable with assuming continuity in the relation between Y and X as well as X and Z, i.e. that participants do not deviate to far from the assigned amount and that the treatment has a continuous effect on the outcome.

The counterfactual processes Y_X and X_Z in this setting have a natural interpretation as the response profiles of a hypothetical participant*. This is the generalization of the never taker-, always taker-, complier-, defier- distinction from the binary case. In a continuous setting there is an uncountable number of such response profiles, not just four. Each of these profiles is specified by one path of the stochastic processes X_Z and Y_X of the first and second stage of the instrumental variable model. One path, i.e. response profile, $X_Z(w)$ in the first stage prescribes the hypothetical participant which amount of the treatment X she takes for* every *possible initially assigned amount $Z = z$. For instance, the path $X_Z(w)$ corresponding to the classical never taker is the one which is always zero for each value of Z. The path $X_Z(w)$ corresponding to the classical complier would coincide with the $45°$ line, since this response profile consists of taking exactly the amount initially assigned. Analogously, one path, i.e. response profile, $Y_X(w)$ describes the reaction Y the hypothetical participant would have for every possible amount $X = x$ of the treatment.*

The continuity assumption is actually not necessary for the approach to work. In fact, in many randomized trials subjects are dropping out completely. In this case, the researchers conducting these experiments would want to allow paths with jumps, in particular paths which jump to zero and stay there (in case of a drop-out). The proposed method can incorporate these assumptions by allowing for more general stochastic processes, modeling this behavior. △

3.2 The theoretical result: Constructing the infinite dimensional linear programs

Based on the representation of the instrumental variable model (1), the proposed method intuitively proceeds by finding an optimal composition (i.e. measure on the paths of Y_X and X_Z *jointly*) which *replicates the observable distribution $F_{Y,X,Z}$* in the data while maximizing or minimizing the respective functional of interest over the paths of the counterfactual process Y_X.

3.2.1 Statement of the theorem and required assumptions

In addition to Assumption 1, the method requires the following assumptions.

Assumption 2 (Normalization)**.** *Y, X, and Z take values in $[0, 1]$.*

Assumption 3 (Continuity)**.** *For some fixed real numbers α, β, γ, $\delta > 0$ and fixed constants $K_x, K_y < +\infty$ the counterfactual processes Y_X and X_Z satisfy*

$$E(|Y_{X=x_1} - Y_{X=x_2}|^{\alpha}) \leq K_y|x_1 - x_2|^{1+\beta} \quad \text{and} \quad E(|X_{Z=z_1} - X_{Z=z_2}|^{\gamma}) \leq K_x|z_1 - z_2|^{1+\delta} \tag{3}$$

for all $x_1, x_2 \in \mathcal{X}$ and $z_1, z_2 \in \mathcal{Z}$.

Assumption 3 implies that almost all paths $Y_{X=x}(w)$ are Hölder continuous[14] of the order $\frac{\beta}{\alpha}$ and almost all paths $X_{Z=z}(w)$ are Hölder continuous of the order $\frac{\delta}{\gamma}$ by an application of the Kolmogorv-Chentsov theorem (Karatzas & Shreve 1998, Theorem 2.2.8) to stochastic processes in random time.

Based on these assumption, the formal method can find sharp upper- and lower bounds on functionals of the counterfactual law $P_{(Y,X)}$. For this, it is convenient to introduce operators $K : (\mathcal{W}, \mathscr{B}_{\mathcal{W}}) \to (\mathcal{Y} \times \mathcal{X}, \mathscr{B}_{\mathcal{Y}} \otimes \mathscr{B}_{\mathcal{X}})$. In this article, the focus lies on operators which take the form of kernel operators as

$$\begin{aligned} KP_W(A_y, A_x) &\coloneqq E_{P_W}[K(Y_X, A_y, A_x)] \\ &= \int K(Y_X(w), A_y, A_x) P_W(dw), \qquad A_y \in \mathscr{B}_{\mathcal{Y}}, A_x \in \mathscr{B}_{\mathcal{X}}. \end{aligned} \tag{4}$$

Assumption 4 (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 continuous or (ii) takes the form of an indicator function[15] in its first argument.*[16]

Important special cases for the kernel are (i) $K(Y_X(w), A_y, A_x) \coloneqq \mathbb{1}_{A_y}\{Y_{X \in A_x}(w)\}$, and (ii) $K(Y_X(w), A_y, A_x) \coloneqq K(Y_X(w), A_x) \coloneqq Y_{X \in A_x}(w)$. Optimization in case (i) for given A_y and A_x results in bounds on the counterfactual probabilities specified by the events A_y and A_x, while optimization in case (ii) will produce bounds on the counterfactual average of Y and the probability that $x \in A_x$. Note in particular that the operator defined in Assumption (4) is *linear*, as its domain is the space of all (probability) measures on the unit interval $(W, \mathscr{B}_{\mathcal{W}})$.[17] In fact, by the classical Riesz representation result (Aliprantis & Border 2006, chapter 14), the space of all Borel measures on $(W, \mathscr{B}_{\mathcal{W}})$ is the dual space of all bounded and continuous functions $\mathcal{W}$. Based on these assumptions, the following is the main theoretical result of this article.

[14] Hölder continuity is a refinement of continuity. A function $f : \mathcal{X} \to \mathcal{Y}$ is Hölder continuous of order $\alpha > 0$ if there exists a constant $0 < c < +\infty$ such that $\|f(x_1) - f(x_2)\|_Y \leq c\|x_1 - x_2\|_X^{\alpha}$ for some norms $\|\cdot\|_X, \|\cdot\|_Y$ on $\mathcal{X}$ and $\mathcal{Y}$. In particular, a Hölder continuous function of order 1 is Lipschitz continuous.

[15] The indicator function is defined as $\mathbb{1}_G(x) = 1$ if $x \in G$ and $\mathbb{1}_G(x) = 0$ if $x \notin G$.

[16] For *fixed* A_y and A_x, the kernel K maps $K : [0, 1] \to \mathbb{R}$.

[17] The approach in this article can handle nonlinear functionals $K(P_W, A_y, A_x)$ equally well. In this case, the practical solution approaches derived further down would require nonlinear techniques like stochastic gradient descent. The focus is on linear objective functions as they cover a wide range of problems of interest and allow for straightforward and efficient solution concepts in practice.

Theorem 1 (Infinite dimensional linear program for bounds on functionals of $P_{(Y,X)}$)**.** *Let Assumptions 1 – 4 hold. Then (lower/upper) bounds on functionals of the counterfactual distribution $P_{(Y,X)}$ for fixed A_y and A_x can be obtained as solutions to the following infinite dimensional linear programs:*

$$\begin{aligned} &\text{minimize/maximize}_{P_W \in \mathscr{P}^*(\mathcal{W})} \quad KP_W(A_y, A_x) \\ &s.t.\ F_{Y,X,Z}(y,x,z) = P_W(Y_X \in [0,y], X_Z \in [0,x], Z \in [0,z]) \end{aligned} \tag{5}$$

for all $(y,x,z) \in [0,1]^3$. $F_{Y,X,Z}$ is the distribution function corresponding to the observable law $P_{Y,X,Z}$. $\mathscr{P}^*(\mathcal{W})$ denotes the set of all measures on $(\mathcal{W}, \mathscr{B}_{\mathcal{W}})$ which induce stochastic processes satisfying Assumption 3.

3.2.2 Intuition of the theorem and explanation of assumptions

Theorem 1 is the main theoretical result of this article. The intuitive idea for its construction is as follows.

The goal is to obtain functionals on the counterfactual law $P_{(Y,X)}$, which does not coincide with the observable $P_{Y,X}$ because of the endogeneity problem described earlier. The endogeneity problem does not affect the *joint* distribution of $P_{Y,X,Z}$, however, so that the latter gives correct information on the problem. Moreover, in light of Proposition 1, one can view the counterfactual laws $P_{(Y,X)}$ and $P_{(X,Z)}$ as the laws corresponding to the counterfactual stochastic processes Y_X and X_Z, which together induce a joint counterfactual process $[Y,X]^*_Z$ under the exclusion- and the independence restriction of model (1). Now in order to obtain bounds on the respective functional on $P_{(Y,X)}$ one wants to find the optimal combination of counterfactual processes Y_X and X_Z whose induced joint process $[Y,X]^*_Z$ has joint law $P_{(Y,X,Z)}$ which *coincides* with the observable joint law. "Optimal combination of counterfactual processes" means finding an optimal measure P_W on the set $(\mathcal{W}, \mathscr{B}_{\mathcal{W}})$, which induces a joint process $[Y,X]^*_Z$ that maximizes (for an upper bound) or minimizes (for a lower bound) the functional of $P_{(Y,X)}$ of interest.

The setting of a randomized controlled trial with imperfect compliance provides a natural explanation of the idea.

Running example (2). *Recall that a path $X_Z(w)$ of the counterfactual process X_Z describes one response profile of a hypothetical participant for all possible states $z \in \mathcal{Z}$ this participant could potentially face. Similarly for a path $Y_X(w)$. The idea for Theorem (1) is then to find an optimal mixture of response profiles for this hypothetical participant that maximizes (for an upper bound) or minimizes (for a lower bound) the objective function of interest* subject to the constraint that *this mixture of response profiles perfectly replicates the* empirically observed *mixture of response profiles of all actual participants, which is captured by $F_{Y,X,Z}$.* △

Assumption 2 is a normalization of the observable random variables. It implies that all variables are univariate. This assumption is in fact not necessary for Theorem 1 to hold, but it

simplifies the exposition and mathematical derivations. One can extend Theorem 1 straightforwardly by allowing for multivariate Y, X, and Z, 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 current practical solution approach suffers from the curse of dimensionality as no sparsity or factor-model assumptions are placed on the stochastic processes.[18]

The smoothness requirements of Assumption 3 are the only structural assumption placed on the counterfactual processes. Again, this assumption is in fact not necessary for Theorem 1 to hold. Theoretically, one could define the paths $Y_X(w)$ and $X_Z(w)$ on the space $\mathbb{R}^{[0,1]}$ of all functions from the unit interval mapping to the reals. However, this space is too big in the sense that many sets of paths of interest, like the space $C([0,1])$ of continuous functions on $[0,1]$, are not measurable in it (Bauer 1996, Corollary 38.5). A "small enough" space which is still more general than the space of continuous functions is the Skorokhod space $D([0,1])$ under the Skorokhod metric (Karatzas & Shreve 1998, p. 409), which allows for the paths to have finitely many jumps. For all practical intents and purposes, the Skorohod space is large enough as finitely many jumps of varying intensity are able to capture any reasonable practical response profile in general.

The smoothness requirements on the paths enter Theorem 1 via the set $\mathscr{P}^*(\mathcal{W})$. This set restricts the possible measures P_W on $(\mathcal{W}, \mathscr{B}_\mathcal{W})$ to measures which only put positive probability on continuous paths $Y_X(w)$ and $X_Z(w)$. In fact, $\mathscr{P}^*(\mathcal{W})$ is the key to introducing any other form of assumption into the optimization problem, by restricting it to only "allowed" probability measures. For instance, one could require the paths $X_Z(w)$ to be strictly increasing in addition to being continuous. Then $\mathscr{P}^*(\mathcal{W})$ would only comprise those measure P_W which do not put positive probabilities on decreasing paths. This way, one can introduce any parametric or nonparametric assumption on the model in principle, by restricting $\mathcal{P}^*(\mathcal{W})$ to only contain probability measures P_W which do not put positive probabilities on unwanted paths. In the practical implementation of Theorem 1 below, these assumptions will be enforced by sampling paths $Y_X(w)$ and $X_Z(w)$ using one representative measure $P_W \in \mathcal{P}^*(\mathcal{W})$. The practical analogue to Theorem 1 then naturally only optimizes over probability measures which do not put positive probability on other paths than those, because no other paths than the existing ones are sampled.

Also notice that there is a need for optimization in Theorem 1 in general, even though there is an uncountable number of equality constraints in the model. To see this, recall the production functions $h(X,W)$ and $g(Z,W)$ are not injective in W. This translates to the stochastic processes $Y_X(w)$ and $X_Z(w)$ each possessing paths which intersect in general. Now, intuitively, if the paths $Y_X(w_1)$ and $Y_X(w_2)$ intersect at a point, it is not clear which path corresponds to what part of the conditional distribution $P_{Y(X=x)}$, which is just another way of stating that knowing the joint distribution $P_{Y,X,Z}$ does not pin down the laws $P_{(Y,X)}$ and $P_{(X,Z)}$ uniquely in general, as the processes are dependent: Y_X depends on the position of X_z.[19]

[18] One can circumvent this curse of dimensionality by placing more assumptions on the geometries of the counterfactual random fields Y_X and X_Z, like sparsity, factor structures, more smoothness, etc.

[19] In this respect, it is interesting to consider point-identification results from the literature on nonseparable

In this respect, notice that the optimizations (5) need not produce a solution as the constraint correspondence

$$\mathcal{C}(F_{Y,X,Z}) \coloneqq \{P_W \in \mathscr{P}^*(\mathcal{W}) : P_W(Y_X \in [0,y], X_Z \in [0,x], Z \in [0,Z]) = F_{Y,X,Z}(y,x,z)\}$$

can be empty for some $F_{Y,X,Z}$. This happens for instance when the observable distribution is equivalent to a joint process $[Y,X]_Z$ with jumps, which cannot be replicated by a counterfactual joint process $[Y,X]^*_Z$ as a combination of processes Y_X and X_Z under the continuity Assumption 3. This is not an issue for the approach, but an actual feature, as a situation like this implies testability of model (1). For more details, see Gunsilius (2019*b*).

3.2.3 Implications and corollaries of the theorem

It is imperative to notice that A_y and A_X in the statement of Theorem 1 need to be fixed by the researcher, which is why the objective function becomes a functional under the specification (4). The solution to the optimization problems (5) then each produces a number, not a function. For instance, solving the optimizations (5) for given A_y and A_x under specification (i) of the kernel gives bounds on the joint probability that the outcome lies in A_y and the treatment is in A_x. From this joint counterfactual distribution one can obtain bounds on the conditional probability $P_{Y(X\in A_x)}(A_y)$ that the outcome lies in A_y given that the (now exogenous) treatment lies in A_x.

In general, handling operators mapping into σ-algebras is more burdensome than operators mapping into the real numbers. Under Assumption 2, one can change the codimension of these operators by introducing the standard filtrations on the sets $\mathcal{X}$ and $\mathcal{Z}$ defined by

$$\mathscr{F}_x \coloneqq \sigma(Y_{X=t}, 0 \leq t \leq x), \qquad \text{and} \qquad \mathscr{F}_z \coloneqq \sigma(Y_{X=t}, 0 \leq t \leq x),$$

where $\sigma(X)$ is the σ-algebra induced by the random variable X (Karatzas & Shreve 1998). Then one can define the kernel operators $K : (\mathcal{W}, \mathscr{B}_{\mathcal{W}}) \to \mathcal{Y} \times \mathcal{X}$ of the form

$$\begin{aligned} KP_W(y,x) &\coloneqq E_{P_W}[K(Y_X, y, x)] \\ &= \int K(Y_X(w), y, x) P_W(dw), \qquad y \in \mathcal{Y}, x \in \mathcal{X}, \end{aligned} \tag{6}$$

triangular models, such as Imbens & Newey (2009). In fact, they require the production function $g(z,\cdot)$ to be strictly increasing and continuous in W. This assumption makes g injective in W. Injectivity of $g(Z,W)$ in W means that the paths $X_Z(w)$ do not intersect almost surely, so that for each (y,x) there is an almost surely unique w inducing this combination. Under some additional structural assumptions on the second stage, this would induce identification of the latent distributions $P_{(Y,X)}$ and $P_{(X,Z)}$. Point-identification therefore is often achieved by ruling out crossings of the paths of the counterfactual stochastic processes. Also, note that an additive separability assumption in the unobservable error term W in the instrumental variable model (Allen & Rehbeck 2019) transforms into other structural assumptions on the paths of stochastic processes. Consider the first stage as an example. If the distribution of W is independent of Z and the function g in model (1) is additively separable in W, then this introduces a stochastic process whose paths can be modeled by an additive relation. It could be interesting to link the stochastic process representation to these types of random utility models.

as the process Y_X is progressively measurable (Karatzas & Shreve 1998, Proposition 1.13). With this set-up, the optimization of the programs (5) under the specification (i) of the kernel then gives bound on the counterfactual probability $F_{(Y,X)}(y,x)$ that the outcome lies in the interval $[0,y]$ and the treatment lies in the interval $[0,x]$.[20]

Different choices of kernel functions in Theorem 1 lead to different causal effects. The following are two examples.

Corollary 1 (Bounds on the ATE)**.** *Theorem 1 provides upper and lower bounds on* $E[Y\mathbb{1}_{A_x}\{X\}]-E[Y\mathbb{1}_{A'_x}\{X\}]$ *for two events* A_x *and* A'_x *by changing the kernel operator in specification (ii) to*

$$K(Y_X(w),A_y,A_x)-K(Y_X(w),A_y,A'_x) \coloneqq Y_{X\in A_x}(w)-Y_{X\in A'_x}(w).$$

Based on this, upper and lower bounds on the ATE

$$E[Y|X\in A_x]-E[Y|X\in A'_x]$$

can be derived by conditioning.

Corollary 2 (Bounds on distributional effects)**.** *Theorem 1 provides upper and lower bounds on* $P(Y\in[0,y],X\in[0,x])-P(Y\in[0,y],X\in[0,x'])$ *by changing the kernel operator in specification (i) to*

$$K(Y_X(w),y,x)-K(Y_X(w),y,x') \coloneqq \mathbb{1}_{[0,y]}\{Y_{X\in[0,x]}(w)\}-\mathbb{1}_{[0,y]}\{Y_{X\in[0,x']}(w)\}.$$

Based on this, upper and lower bounds on the distributional effect

$$P(Y\leq y|X\leq x)-P(Y\leq y|X\leq x')$$

can be derived via conditioning.

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 (5) under the specification of the kernel proposed in Corollary 2 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 $F_{(Y,X)}(y,x_0)$. Respective bounds on the quantile distributions can then be obtained by inverting these envelopes.

[20] Recall that under Assumption 2 the codomains for Y, X, and Z are $[0,1]$. Also, here and throughout the article, $F_{(Y,X)}$ denotes the counterfactual distribution corresponding to the counterfactual measure $P_{(Y,X)}$.

4 Implementing the method in practice

The programs (5) in Theorem 1 are infinite dimensional—as one is optimizing over probability measures on paths of stochastic processes—with a continuum of equality constraints. Finding optimal measures over a path space is a general mathematical and computational problem which shows up in many guises. Examples are general stochastic control- and reinforcement learning problems with a continuous action space as well as optimization routines via Feynman's path integral formulation (Kappen 2007). Existing practical solution approaches usually exploit a Markovian- or martingale structure in the problem and apply dynamic programming ideas. In contrast, the programs (5) do not rely on a dynamic structure as these stochastic processes are defined in a "counterfactual space". This precludes the use of existing approaches for solving them.

This section therefore introduces a novel probabilistic approach to solve the problems (5) approximately in practice with probabilistic finite sample guarantees for the validity of the approximation.[21] The method proceeds by sampling shape preserving bases of the path space and solving semi-infinite dimensional analogues of (5) over this sample of basis functions. The introduction of randomness via the sampling of basis functions is what provides finite sample guarantees for the method to approximate the solution to the infinite dimensional problem. The reason for this is twofold. First, by using shape preserving bases to approximate the respective functions, one can obtain bounds on how accurate the approximation is for *a given sample* of basis functions (Anastassiou & Yu 1992*b*). Second, the randomness introduced by sampling permits the use of standard large deviation theory (Vapnik 1998, Chapter 5) which provides probabilistic guarantees for the approximation of (5) via a random sample.[22]

Note in this respect that standard estimation approaches—for example via the classical method of sieves (Chen 2007)—most likely cannot be employed for solving (5). The reason is that (5) are not classical estimation problems. In fact, the estimation problem which estimation procedures are designed to solve is only captured by the *constraint* in problems (5). On top of this, there is another optimization to solve in a "counterfactual space" for which no data exist to perform estimation procedures. To expand on this, note that one could try and use standard regression approaches to approximate just the constraint by using a form of nonparametric regression. That is, one could relax the equality by letting both sides deviate a little in some functional norm like the L^2-norm, and then obtain a measure P_W which solves the approximate constraint. This approach would result in a unique P_W, ignoring the actual counterfactual optimization problem. Even though standard estimation approaches do not work, the intuition for using basis functions

[21] One can in fact only approximate the solutions to (5) (Kappen 2007, p. 152) since one cannot solve the problem over all possible paths on a space.

[22] The sampled basis need not be shape-preserving. Any standard basis used in the method of Sieves, like (trigonometric) polynomials, splines, wavelets, or artificial neural networks work (Chen 2007). The focus is on shape-preserving wavelets, because they allow the introduction of further assumptions besides continuity like monotonicity or convexity.

as in Anastassiou & Yu (1992*b*) and Chen (2007) to solve the problem proves useful.

4.1 Construction of the semi-infinite dimensional approximations

The idea is to approximate the infinite dimensional problems (5) by the semi-infinite dimensional analogues

$$\begin{aligned} &min/max_{\hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W})} \quad \frac{1}{l}\sum_{i=1}^{l} \tilde{K}(\tilde{Y}_{\tilde{X}}(i), A_y, A_x)\frac{d\hat{P}_W}{d\hat{P}_0}(i) \\ &\text{s.t.} \quad \left\| F_{Y,X,Z} - \frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), \cdot, \cdot, \cdot, \eta)\frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)} \leq \varepsilon \end{aligned} \tag{7}$$

for some small $\varepsilon > 0$ which should be chosen based on finite sample properties.[23]

The idea underlying the approximations (7) is to sample a number l of basis functions which approximate the path space of the processes Y_X and X_Z. Of particular convenience are basis functions used in Sieves estimation such as polynomials, trigonometric polynomials, splines, wavelets, and neural networks (Chen 2007). Results for how well these basis functions can approximate a given function have been derived (Chen 2007, p. 5573), which are important for establishing the probabilistic approximation guarantees in Theorem 2 later on. 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 mother wavelet $\varphi : \mathbb{R} \to [0,1]$ 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 fixed dilation* κ is

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

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

[23] Rewriting the programs (7) 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.

Anastassiou & Yu (1992*a*).[24] This property is convenient as it permits to introduce additional structural assumptions on the model in a straightforward way by sampling only wavelets which lead to a certain shape of the paths. Furthermore, one can derive, for any given dilation κ, the degree of approximation of the wavelet basis to some function with a given modulus of continuity[25] (Anastassiou & Yu 1992*b*, Theorem 1). In principle, summing over all $\kappa \in \mathbb{Z}$ will replicate any possible function in L^2, but the approximation guarantees for fixed κ already imply that choosing a large enough dilation can give good approximations to "regular" functions, where regularity is defined via their modulus of continuity. In the following, a fixed κ is chosen and the dependence of $\tilde{Y}_{\tilde{X}=x}(i)$ and $\tilde{X}_{\tilde{Z}=z}(i)$ on κ is made implicit.[26]

The dependence of $\tilde{Y}_{\tilde{X}=x}(i)$ and $\tilde{X}_{\tilde{Z}=z}(i)$ on the index i, which indexes the elements in the sample of basis function shows that the problems (7) 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 the proceeds 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 probabilities 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 an additional assumption on the *theoretical sampling law* P_0, for which $\hat{P}_0$ is its "finite sample" approximation:

Assumption 5 (Representative P_0)**.** *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 Radon-Nikodym derivative satisfying $0 < c_{RN} \leq \frac{dP_W}{dP_0} \leq C_{RN} < +\infty$ for fixed constants c_{RN} and C_{RN}.*

Assumption 5 is the theoretical analogue to the fact that all $\hat{P}_W$ are absolutely continuous with respect to $\hat{P}_0$ by construction.[27] For the theoretical data-generating process, this is not necessarily

[24] By choosing appropriate coefficients $\alpha_j(i)$ and $\beta_j(i)$, one can potentially introduce different shape assumptions beyond monotonicity and convexity.

[25] If a function f satisfies $|f(x) - f(x')| \leq \omega_f(|x - x'|)$ for all x, x' in its support and some function $\omega_f : [0, +\infty) \to [0, +\infty)$, then this ω_f is the *modulus of continuity of* f.

[26] In practice, one can sum over several levels κ, which would only improve the approximation. The subsequent proofs work with just one fixed κ, so that they give a "worst case" approximation for a given κ. Furthermore, note that one can use any sieve basis to generate paths. The programs will then automatically adjust to these sampled paths. The proofs in this article work for any other of these Sieve bases, as long as results about their approximation properties exists as in Chen (2007, p. 5573).

[27] In fact, it requires something slightly stronger, namely that all Radon-Nikodym derivatives in this set are uniformly bounded. This assumption, in contrast to the mere absolute continuity assumption, will yield "hard" finite sample approximation results, compared to only only soft convergence results.

the case, even under Assumptions 3. However, since by the finite dimensional construction all measures $\hat{P}_W$ are automatically 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 5 is non-testable. It is an assumption on $\mathscr{P}^*(\mathcal{W})$ which necessarily follows from the approximation argument.

The function $\Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta)$ can either be the product of indicator functions in the constraint of (5), i.e.

$$\Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta) = \mathbb{1}_{[0,y]}(\tilde{Y}_{\tilde{X}}(i)))\mathbb{1}_{[0,x]}(\tilde{X}_{\tilde{Z}}(i))\mathbb{1}_{[0,z]}(\tilde{Z}(i))$$

in

$$P_W(Y_X \in [0,y], X_Z \in [0,x], Z \in [0,z]) = \int \mathbb{1}_{[0,y]}(Y_X(w))\mathbb{1}_{[0,x]}(X_Z(w))\mathbb{1}_{[0,z]}(Z(w))P_W(dw)$$

or an approximation of this product of indicator functions by some sigmoid functions, like the logistic function $S(x) = \frac{\exp(x)}{\exp(x)+1}$. For example, one can approximate $\mathbb{1}_{[0,y]}\{Y_X(w)\}$ arbitrarily well by

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

as $\eta \to +\infty$. One can use this approximation for all three indicator functions and take the product of these three logistic functions. For notational purposes, $\Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta)$ denotes this product, where the same $\eta > 0$ is chosen for all three approximations.[28] The kernel $\tilde{K}(\tilde{Y}_{\tilde{X}}(i), A_y, A_x)$ coincides with the kernel $K(\tilde{Y}_{\tilde{X}}(i), A_y, A_x)$ when the latter is bounded and continuous (for instance when $K(\tilde{Y}_{\tilde{X}}(i), A_y, A_x) = \tilde{Y}_{\tilde{X}\in A_x}(i)$) and is a smooth approximation to $K(\tilde{Y}_{\tilde{X}}(i), A_y, A_x)$ when the latter takes an indicator function form. In that case, the variable $\eta > 0$ which controls the degree of approximation, will be made implicit to save on notation.

Approximating the constraint in (5) by a relaxed version via the L^2-norm is convenient for three main reasons. First, the probabilistic approximation results derived momentarily allow control of the approximation of the theoretical constraint in (5) in L^2-norm, which makes it convenient to connect the value $\varepsilon > 0$ to the sample l and the modulus of continuity of the observable distribution $F_{Y,X,Z}$. Second, for estimation of (7) one replaces the theoretical data-generating process $F_{Y,X,Z}$ by an estimator $\hat{F}_{Y,X,Z}$, which could lead to the constraints not being satisfied with respect to $\hat{F}_{Y,X,Z}$, even though they might be with respect to $F_{Y,X,Z}$. A slight relaxation of

[28] In the practical implementation of the method in the next section, where the objective is also linear, the constraint uses the indicator functions and not their approximations, as the problem is easier to handle from a computational perspective in this case: the use of indicator functions introduces sparse matrices in the program which can be handled efficiently. However, more complicated and possibly nonlinear objective functions might necessitate more general tools like stochastic gradient descent approaches to solve these programs, which most often require smooth constraint- and objective functions. In such a case it would be imperative to approximate the indicator functions by sigmoid functions.

these constraints guards against this possibility. Thirdly, providing inference results on the finite dimensional version of problem (7) is fundamentally easier with the relaxed constraint.

One more assumption is needed before stating the probabilistic approximation theorem. In particular, one needs to strengthen Assumption 4.

Assumption 6 (Strengthening of Assumption 4)**.** *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 continuously differentiable in $Y_X(w)$ with finite VC dimension[29] *or*

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

Assumption 6 is a strengthening of Assumption 4 in that it requires general kernels to have finite VC dimension and to be continuously differentiable instead of merely continuous. Requiring a finite VC dimension is overly restrictive, in fact, as one could allow for weaker restrictions and still obtain the same probabilistic approximation guarantees[30]; the concept of VC dimension is well-understood in econometrics, however, which makes it a convenient assumption. In this respect, note that indicator functions of the form $\mathbb{1}_{[0,y]}(Y_X(w))$ have VC dimension equal to 2. Logistic approximations to indicator functions of the form $S(Y_X(w), y, \eta)$ have finite VC dimension as proved in Lemma 1 in the appendix, so that indicator functions as well as their logistic approximation of the form presented here will lead to probabilistic approximation guarantees. Furthermore, the set of kernels $K(Y_X(w), x) := Y_{X\in[0,x]}(w)$ has finite VC dimension (Vapnik 1998, p. 192), which shows that the now following probabilistic approximation result holds for the average treatment effect in particular.

Theorem 2 (Probabilistic approximation via the "sampling of paths"-approach)**.** *Under Assumptions 1 – 3, 5, and 6, there exists for every $\varepsilon > 0$ a sample size*

$$\mathbb{N} \ni l^* := l\left(\varepsilon, \eta, \kappa, \alpha, \beta, \gamma, c_{RN}, C_{RN}, \delta, d_{VC}(\Theta), d_{VC}(K), \rho\right)$$

such that with probability $1-\rho$,

$$\max\{|V^* - \tilde{V}^*|, |V_* - \tilde{V}_*|\} < \varepsilon,$$

where V^*, V_* are the maximal and minimal values of (5) and $\tilde{V}^*$ and $\tilde{V}_*$ are the maximal and minimal values of (7). Here, $\eta > 0$ is the degree of approximation of the logistic function

[29] The VC dimension $d_{VC}(\mathcal{S})$ of a set $\mathcal{S}$ of indicator functions is equal to the largest number h of vectors that can be separated into two different classes in all the 2^h possible ways using this set $\mathcal{S}$ of functions (Vapnik 1998, p. 147). The concept of VC dimension also makes sense for more general functions than indicator functions.

[30] As an example, one could only require that the *annealed entropy of the set of indicators* (Vapnik 1998, p. 191) of K has sublinear growth in l, which is a sufficient and necessary restriction for the probabilistic approximation to hold (Vapnik 1998, Theorem 15.1), but this assumption would require introducing a new concept compared to the well-known concept of VC dimension.

to the indicator function, κ is the dilation parameter of the shape-preserving wavelet, α, β, γ, and δ are the smoothness coefficients from Assumption 3, c_{RN} and C_{RN} are the bounds on the Radon-Nikodym derivatives in Assumption 5, and $d_{VC}(\Theta)$ and $d_{VC}(K)$ are the VC dimensions of $\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta)$ and $K(Y_X(w), A_y, A_x)$.

The actual requirement l^* must satisfy for this approximation is $\max\{\mathcal{D}(l^*), \mathcal{D}'(l^*)\} < \varepsilon$, where

$$\mathcal{D}(l) \coloneqq c(\eta) + C(\eta)^2 2^{2(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}} + \frac{(C_{RN} - c_{RN})^2}{l}\left[d_{VC}(\Theta)\left(\ln\left(\frac{2l}{d_{VC}(\Theta)}\right)+1\right) - \ln(\rho/4) + 1\right]$$

and

$$\mathcal{D}'(l) \coloneqq c'(\eta) + C'(\eta)^2 2^{2(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}} + \frac{(C_{RN} - c_{RN})^2}{l}\left[d_{VC}(K)\left(\ln\left(\frac{2l}{d_{VC}(K)}\right)+1\right) - \ln(\rho/4) + 1\right]$$

are the approximation bounds for the constraint and the objective function, respectively.

Theorem 2 provides a lower bound on the number of sampled paths l for the semi-infinite program to provide a good approximation to the infinite dimensional program with "high probability". This number l in general depends on all parameters introduced to approximate the infinite programs (η, κ, $d_{VC}(\Theta)$, $d_{VC}(K)$), the assumed smoothness of the paths (α, β, γ, δ), as well as the probability ρ and the accuracy ε of the approximation. This is the most general form of this approximation result, as it works with the logistic approximations to the indicator functions and hence relies on the VC dimension of these functions. In the case where constraint and objective function kernel are indicator functions, $d_{VC}(\Theta) = d_{VC}(K) = 2$. This also shows that one can use any approximation to the indicator function with finite VC dimension. Lastly, κ is introduced by using the above mentioned shape preserving wavelets. When using a different sieve basis for which approximation results are known (e.g. Chen 2007, p. 5573), l^* will depend on a different approximation parameter than κ.

The basic idea underlying Theorem 2 seems new in the mathematical literature on approximating infinite dimensional program by semi-infinite dimensional programs. 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 2 seems to be the first in using these arguments for solving optimization problems on infinite dimensional path space, especially without the introduction of dynamic assumptions (Kappen 2007). For these types of problems, the randomness introduced is actually *desired*, 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. In contrast, a sampling approach appears to be the only fruitful approach towards solving problems on path spaces of similar generality to (5) at all in practice.

4.2 Inference results

The actual *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.[31] This subsection introduces large sample results which enable the researcher to perform standard inference on the solution of the programs (7). These results are only derived for each bound separately. 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.

Another point worth mentioning: even though the programs (7) have relaxed constraints, it could potentially still be the case that for very small $\varepsilon > 0$ there exist data-generating processes $F_{Y,X,Z}$ for which no $\hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W})$ exists which satisfies the constraint. Still, as mentioned before, a $F_{Y,X,Z}$ which cannot be replicated by some $\hat{P}_W$ directly introduces testable assumptions on the model (Gunsilius 2019*b*). In this sense the programs (7) could potentially be used for deriving general tests for instrument validity in practice. 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, $\mathcal{F}_{Y,X,Z}$ denotes the set of all cumulative distribution functions on $[0,1]^3$ equipped with the $L^\infty([0,1]^3)$-norm.

Assumption 7 (Non-emptiness of the constraint set)**.** *For given $F_{Y,X,Z} \in \mathcal{F}_{Y,X,Z}$ there exists a neighborhood $\mathcal{U} \in \mathcal{F}_{Y,X,Z}$ such that the constraint set $\tilde{\mathcal{C}} :=$*

$$\left\{ \hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W}) : \left\| F_{Y,X,Z}(y,x,z) - \frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)} \leq \varepsilon \right\}$$

is non-empty for some small $\varepsilon > 0$ and all $F'_{Y,X,Z} \in \mathcal{U}$.

Theorem in Gunsilius (2019*b*) contains a smoothness assumption on $F_{Y,X,Z}$ which guarantees that the theoretical constraint in (5)—and therefore the constraint in (7) for every $\varepsilon > 0$ and large enough sample l—is satisfied. Since this is just one possible sufficient condition for the constraint correspondence to be non-empty among many, and since emptiness of the constraint correspondence for some $F_{Y,X,Z}$ would generate new knowledge on the model which would be beneficial in practical purposes, it seems to be warranted to write Assumption 7 as a high-level assumption.

In the following, $\hat{F}_{Y,X,Z;n}$ denotes the standard empirical distribution function and $\hat{\underline{V}}_*(\hat{F}_{Y,X,Z})$ and $\hat{\overline{V}}^*(F_{Y,X,Z})$ denote the minimal and maximal value functions of (7). The first result concerns

[31] The theoretical results in this section are derived for the standard empirical cumulative distribution function 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.

the consistency of these value functions. The idea is to use the standard 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 $L^\infty([0,1]^3)$-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 – 3 and 5 – 7 it holds that*

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

almost surely as $n \to \infty$*.*

The derivation of the large sample distribution of $\hat{\tilde{V}}^*$ and $\hat{\tilde{V}}_*$ follows form the classical Donsker 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

Proposition 3 (Large sample distribution)**.** *Under Assumptions 1 – 3 and 5 – 7 it holds that*

$$\begin{aligned}
&\sqrt{n}(\hat{\tilde{V}}_*(\hat{F}_{Y,X,Z;n}) - \hat{\tilde{V}}_*(F_{Y,X,Z})) \rightsquigarrow d\hat{\tilde{V}}_{*,F_{Y,X,Z}}(\mathbb{G}_{F_{Y,X,Z}}) \qquad \text{and} \\
&\sqrt{n}(\hat{\tilde{V}}^*(\hat{F}_{Y,X,Z;n}) - \hat{\tilde{V}}^*(F_{Y,X,Z})) \rightsquigarrow d\hat{\tilde{V}}^*_{F_{Y,X,Z}}(\mathbb{G}_{F_{Y,X,Z}}),
\end{aligned}$$

as $n \to \infty$*.*

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

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

for all $(y,x,z),(y',x',z') \in [0,1]^3$, and "$\rightsquigarrow$" denotes weak convergence.

The directional Hadamard derivative takes the form

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

where $L(\cdot,\cdot,\cdot)$ denotes the Lagrangian of the program (7), $\lambda(\hat{P}_1)$ denotes the respective Lagrange multiplier (which is unique for given $\hat{P}_1$ by Proposition 4.47 in Bonnans & Shapiro 2013), and $\delta_{F_{Y,X,Z}} L$ is the directional derivative of L in its third argument in direction $F'_{Y,X,Z}$ at $F_{Y,X,Z}$. Due to the generality of Assumption 6, one cannot say more in general about the form of the Lagrangian. In the special—but frequent—case where the kernel K induces a convex functional, the directional Hadamard derivative takes the form of an inner product (Bonnans & Shapiro 2013, Theorem 4.24).[32]

[32] Proposition 3 works equally well with a smoothed estimator $\hat{F}_{Y,X,Z;h_n}$ of $F_{Y,X,Z}$ and bandwidth h_n. The only

Proposition 3 implies that the rate of convergence is parametric, which should not be surprising 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 rather standard distribution from a purely statistical perspective, as it takes the form of the first-order directional Hadamard derivative of the value function taken at $F_{Y,X,Z}$ in the direction of $\mathbb{G}_{F_{Y,X,Z}}$. When the optimization program is linear, these Hadamard derivatives can be calculated analytically in the respective setting as they take the form of an optimization over an inner product in this setting (Bonnans & Shapiro 2013, section 4.3). In addition, there are many 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 (7). 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 2 and Proposition 3 give a complete picture of the approximative behavior of the programs (7). 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.

5 Proof of concept: Bounds on expenditures

As a demonstration of its capabilities, the method is applied to estimate bounds on expenditure differences using the 1995/1996 UK family expenditure survey. This problem is well suited for demonstrating the method's practical capabilities 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.

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

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.

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.

The only structural assumption upheld 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. Beyond this, no other structural assumptions are introduced. 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 current 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 very tight, showing the identificatory strength of these assumptions in this setting. All of these results are contained in the following two subsections.

5.1 Computational approach

The programs (7) 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 $\hat{F}_{Y,X,Z;n}$ on the values taken by the sample $(Y_i, X_i, Z_i)_{i=1,\dots,n}$. The second is to evaluate $\hat{F}_{Y,X,Z;n}$ 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. 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 in two ways from the theoretical approach. First, it uses a smoothed variant $\hat{F}_{Y,X,Z;h_n}$ of the empirical cumulative distribution function, where the bandwidth is determined via cross-validation. Heuristically, it seems as the introduced smoothness gives slightly more robust results compared to the standard empirical cumulative distribution function. Second, the practical implementation reduces the statistical randomness in the programs (5) by conditioning on Z. That is, instead of replicating the distribution $\hat{F}_{Y,X,Z;h_n}$ in the constraint of (5), the programs replicate the conditional cumulative distribution function $\hat{F}_{Y,X|Z=z;h_n}$ by the measure $P_W(Y_X \in [0,y], X_{Z=z} \in [0,x])$ for a grid of Z-values determined by the approximation ι of the unit interval. This reduces the computational burden of the program slightly (because one

does not need to introduce another set of indicator functions for Z) while giving the same results as the general version.[33]

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[34], $\vec{1}$ 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.[35] $\tilde{\Theta}$ 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 $L^2([0,1]^3)$-norm from (7) 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) can be rewritten as

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

where λ_{min} and λ_{max} are allowed to be different in principle. 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

[33] 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.

[34] 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 $\vec{1}$.

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

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 (5) and *a fortiori* (7) 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.[36] This sparseness is a blessing 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, and (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.[37]

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 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 intuition for this is that for a small sample l_s in practice the constraints in (10) will be overdetermined and will not admit a solution in general. So 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 ignores the objective function. On the other hand, if λ is too low, the program ignores the constraint, which will always result in trivial bounds. This implies, however, that there exists a range of λ-values

[36] In the case where the indicator function in the constraint (5) is approximated by a logistic function, the matrix $\tilde{\Theta}$ is no longer binary sparse.

[37] Discarding elements ex-post in optimization routines is not new. For instance Wu, Fang & Lin (2001) use discards in solving a related optimization to the proposed one, the general capacity problem on Euclidean state space.

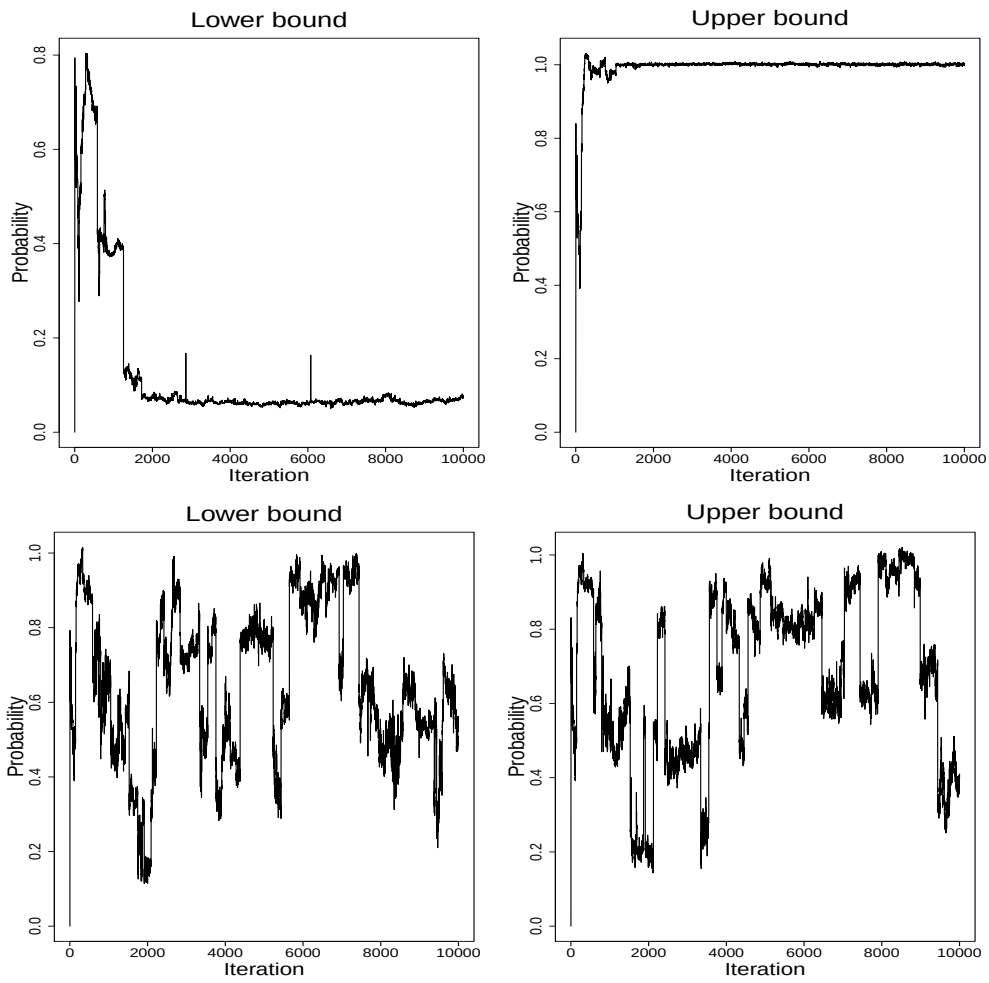

Figure 1: Convergence of upper and lower bounds on the counterfactual probability $F_{Y(X)=0.75}(0.75)$ in the household expenditure example for a coarse approximation of the unit interval. Here, Y is the relative expenditure of a household on leisure and X is the overall expenditure of the household. The penalties are $\lambda = 100$ (top) and $\lambda = 600$ (bottom) for a coarse approximation ι 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 top panels show convergence of the algorithm, whereas the algorithm does not converge for the bottom panels. The value in the top left panel is nontrivial at 0.065.

for which constraint and objective function are balanced off, which provides the correct solutions. Figure 1 depicts the behavior of the solution paths of this "sampling trick" in a stylized setting of the household expenditure application where the coarseness of the approximation of the unit interval makes it possible to sample *all possible* paths.[38]

[38]The bounds are enormous, which is due to the coarse approximation of the unit interval, but surprisingly non-trivial, as the lower bound is roughly 0.065. This shows that the method can produce nontrivial bounds even in extremely coarse approximations of the unit interval. In the actual application below, a finer approximation is chosen which results in considerably tighter bounds.

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. In the following application a λ with value 1 seems to provide this.[39]

5.2 Estimating qualitative bounds on household expenditures

This application not only serves as a demonstration of the capabilities of the method, but also a sanity check. 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. Note that a sanity check is required as the method works in *counterfactual space*, so that it is not clear how to set-up Monte-Carlo simulations which allow to check the estimation of the counterfactual bounds. In fact, this application is actually more challenging than a hypothetical 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 something like this should be possible with *any* method. The fact that this method does provide informative bounds is a testament to its potential usefulness.

Therefore, in the following the focus will be on the outcomes food and leisure. Figure 2 depicts the solution paths for obtaining bounds on the counterfactual difference $F_{Y(X=0.75)}(0.25) - F_{Y(X=0.25)}(0.25)$ 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 parameters λ_{min} and λ_{max} are both equal to 1, because the grid is now much finer than before. Remarkably, the bounds in this setting are qualitatively informative for the problem. The left panel depicts the households' expenditures on leisure and the right depicts their expenditures on food. 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 500 iterations are 0.88 and 0.0080 for the general upper- and lower bound, and 0.67 and 0.31 for the corresponding upper-

[39] It is an intriguing open question how to determine an appropriate λ by data-driven methods. Note that standard cross-validation approaches are most-likely not of help as they would not take into account the optimization problem as simply focus on replicating the constraint. 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.

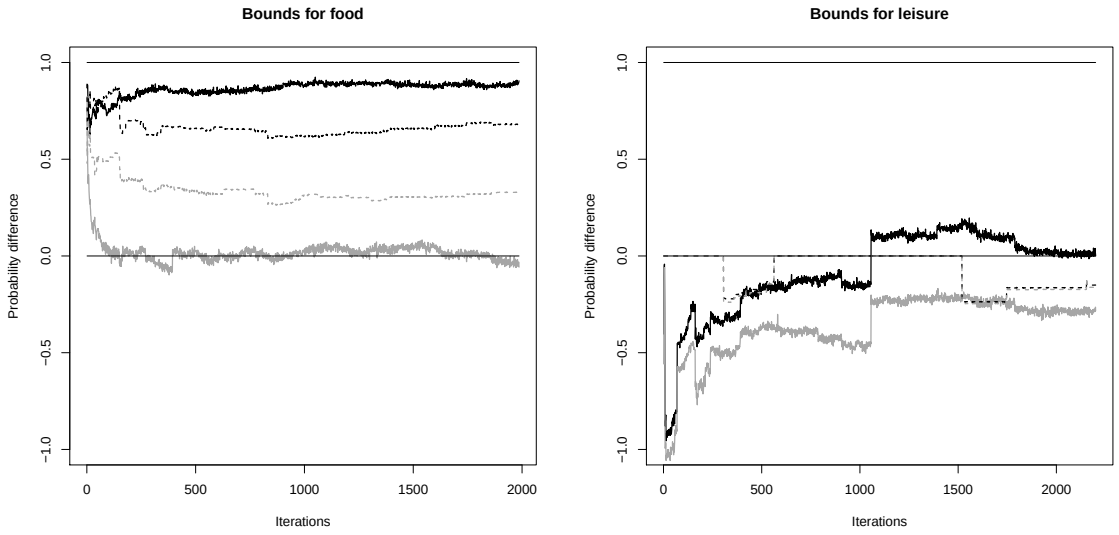

Figure 2: Convergence of upper (black) and lower (gray) bounds on $F_{Y(X=0.75)}(0.25) - F_{Y(X=0.25)}(0.25)$ for Y being the relative spending on food (left panel) and leisure (right panel). The solid lines indicate the bounds under no additional assumptions on the model, the dashed lines indicate the bounds under the additional assumption that Y is decreasing (food) or increasing (leisure) in overall expenditure X and that X is increasing in income of the head of the household Z. The penalty terms are $\lambda_{min} = \lambda_{max} = 1$ 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.

and lower bounds for the monotone model. All four bounds are positive[40], which indicates that families that spend a lot in general ($X = 0.75$) and spend up to a quarter on food ($Y \in [0, 0.25]$) 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 quartiles 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. This is expected from a necessity good at this given level $y^* = 0.25$. Still, 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 vastly shift up the lower bound, indicating that monotonicity has a strong identificatory content in this setting. In particular, they imply that vastly more families (between 30% and 67%) in the upper quartile of overall spending ($X = 0.75$) spend only up to a quarter 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 88% and as low as 0.8%, but still positive.

The results for expenditure on leisure for this given scenario are similarly compelling. For a

[40]Overall, it is striking that the lower general bound fluctuates around zero in a way that longer average are mostly positive.

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 a quarter of their overall expenditure on leisure ($Y \in [0, 0.25]$) 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 lie in the lower quartile for expenditure on leisure ($Y \in [0, 0.25]$) 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 these levels (i.e. for $y^* = 0.25$). In particular, the averages of the last 500 iterations of the general bounds are 0.030 and -0.28, implying that typically more families (up to 28%) in the lower quartile of overall spending ($X = 0.25$) spend only up to a quarter of their overall on food compared to families in the upper quartile of overall spending. It is true that the upper bound is slightly positive, but it is sufficiently close to zero to warrant this qualitative interpretation. What is more, the bounds for the monotonic model almost coincide and fluctuate, indicating that monotonicity has a very strong identificatory content in this setting and that the general bounds are actually more robust in this setting.

The value of $y^* = 0.25$ was chosen to compare families who do not spend more than a quarter of overall expenditures on the respective good overall. One can repeat this exercise for other values y^*, for instance 0.5. Figure 3 depicts this setting.

The results for leisure are similar to the ones for $y^* = 0.25$. In particular, the averages of the last 500 iterations of the general bounds are 0.09 and -0.34, respectively, indicating the same general result as previously. The bounds for the monotonic model are -0.19 and -0.22, showing that once again monotonicity has a strong identificatory content by making the implications much more pronounced. Here again, however, are the monotonic bounds rather erratic compared to the general bounds. As a side-note: the bounds for the monotonic model first lie outside of the general bounds, but converge to values in the interior of the latter. This follows from the fact that the computational approach only samples a quarter of paths to be monotone, so that it takes longer for the monotone bounds to converge.

The results for food are different to the ones for $y^* = 0.25$. In fact, the averages of the last 500 iterations of the general bounds are 0.34 and -0.33, indicating that there is no clear difference between families that spend much overall ($X = 0.75$) and families that do not ($X = 0.25$). This is a realistic result, however, as the values $y^* = 0.5$ compares families that spend up to half of their whole income on food. Surprisingly, however, the monotonicity assumption still give the same (albeit slightly weaker) clear results that food is a necessity good at this level. In fact, the monotonic bounds are 0.072 and 0.0063, indicating that typically more families (between 0.63% and 7.2%) in the upper quartile of overall spending spend up to a half on food compared to families in lower quartiles. Again, monotonicity assumptions have a rather strong identificatory content in this setting.

Finally, Figure **??** depicts these differences in counterfactual distributions for $y^* = 0.15$, where

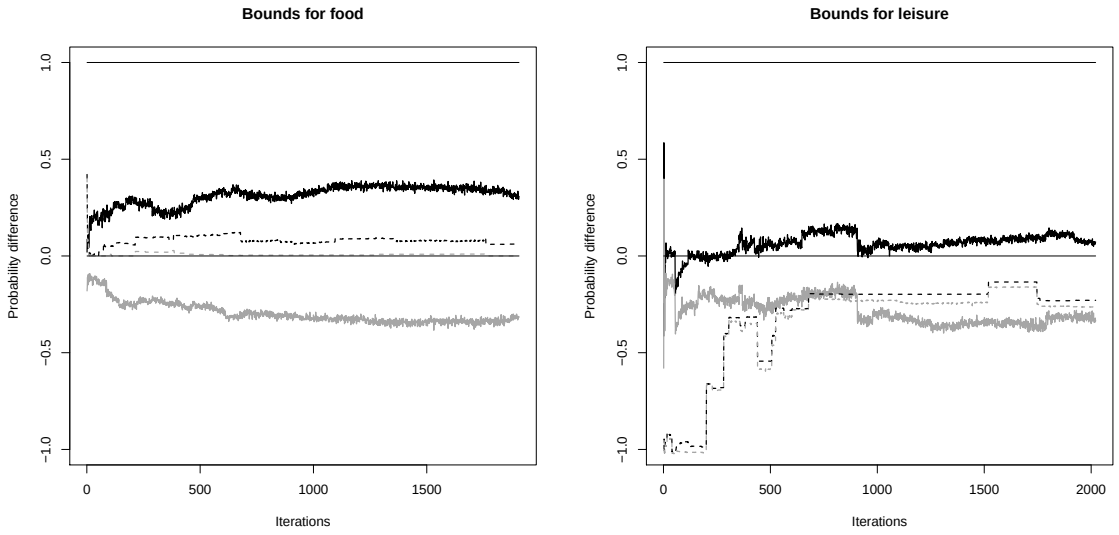

Figure 3: Convergence of upper (black) and lower (gray) bounds on $F_{Y(X=0.75)}(0.5) - F_{Y(X=0.25)}(0.5)$ for Y being the relative spending on food (left panel) and leisure (right panel). The solid lines indicate the bounds under no additional assumptions on the model, the dashed lines indicate the 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 terms are $\lambda_{min} = \lambda_{max} = 1$ 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.

the results are even more pronounced. This is not surprising, as one compares a much smaller set of individuals with larger differences in mass distributions as for the previous cases. It turns out, however, that the method sampled too few decreasing paths Y_X that went through the interval $[0, y^*] = [0, 0.15]$ at vales $X = 0.75$ and $X = 0.25$, yielding upper-and lower bounds that were constantly zero. In this case, this is not an issue, as the general bounds without monotonicity are already informative in this setting. One can circumvent this problem by forcing the program to sample paths that must go through the respective events of interest/

The following table provides an overview of the results at the levels $y^* = 0.15, 0.25, 0.5, 0.75$. Recall that the objective function in each case is to maximize or minimize the probability that $Y_X \in [0, y^*]$, so that the level $y^* = 0.75$ is not likely to be informative, which is exactly what the method finds.

Overall, these qualitative results are remarkably informative. 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

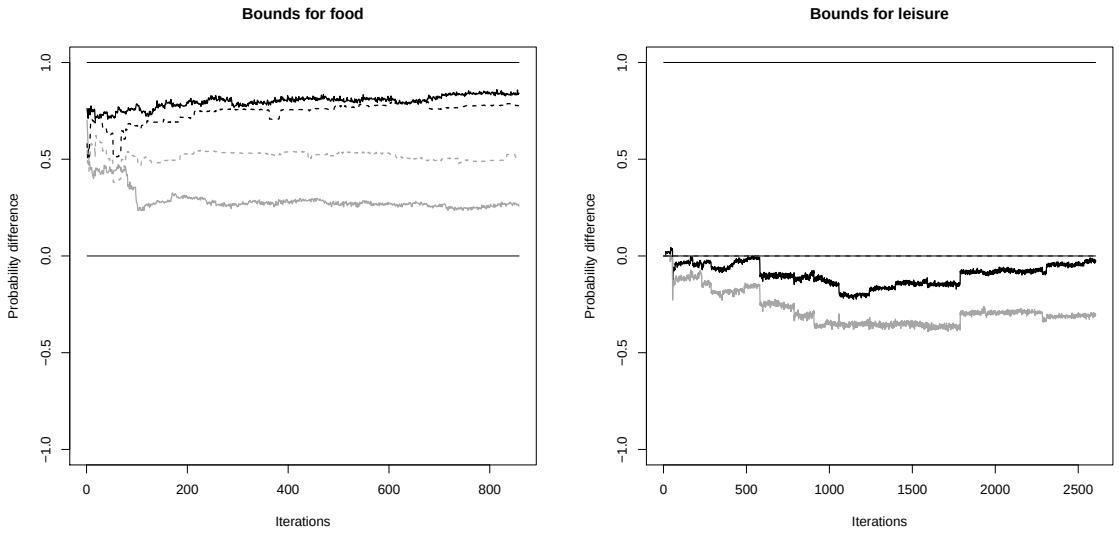

Figure 4: Convergence of upper (black) and lower (gray) bounds on $F_{Y(X=0.75)}(0.15) - F_{Y(X=0.25)}(0.15)$ for Y being the relative spending on food (left panel) and leisure (right panel). The solid lines indicate the bounds under no additional assumptions on the model, the dashed lines indicate the bounds under the additional assumption that Y is decreasing (food) or increasing (leisure) in overall expenditure X and that overall expenditure X is increasing in income of the head of the household Z. The penalty terms are $\lambda_{min} = \lambda_{max} = 1$ 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 1: Upper- and lower bounds for $F_{Y(X=0.75)}(y^*) - F_{Y(X=0.25)}(y^*)$ for different values y^*.

& 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 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, but 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 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, which can now be gauged for the first time.

6 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 treatments. The statistical problem it is designed to solve is not a standard estimation problem. In fact, the underlying 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.

Despite many advances in the mathematical literature for solving infinite dimensional (linear) programs (Anderson & Nash 1987), there have been no results available for solving infinite dimensional programs on path spaces without making dynamic assumptions (Kappen 2007). As a second main contribution, this article therefore introduces a novel "sampling-of-paths" approach to solve these types of 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. It is this introduced randomness for solving general programs on paths spaces which seems novel in the mathematical literature. In the literature on function approximation theory, there have been approaches using large deviation results (Girosi 1995), but they eventually arrive at deterministic approximation guarantees. One other article (Pucci de Farias & Van Roy 2004) also explicitly introduces randomness by sampling constraints in Markovian optimization problems. The differ-

ence between the proposed method and that article is that the proposed method works directly on path spaces and for more general problems than Markovian optimization problems[41]; in fact, it seems like a sampling approach is the only fruitful approach for solving these general problems as one cannot optimize over all possible paths without making dynamic assumptions. Also, the proposed method samples the whole *state space*, not just constraints as the method in Pucci de Farias & Van Roy (2004).

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) and analytical expressions for the large sample distributions are derived. 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).[42] 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 (Hastie, Tibshirani & Friedman 2009, chapter 18), 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

[41] The solution method proposed here can handle any type of dynamic assumption on the processes, like Markovianity or martingale assumptions, by sampling the respective paths of the process.

[42] Resampling methods might be too computationally expensive for general complex problems, in which case one should rely on the analytical expressions derived.

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, jumps in processes, Slutsky-type conditions, first-passage times, martingale properties, etc. An extension of the current program which already accommodates some of these additions is in the works.

Finally, the generality of this method allows researchers to compare the identificatory content of rather different, but nonetheless 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).

References

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

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

Allen, R. & Rehbeck, J. (2019), 'Identification with additively separable heterogeneity', *Econometrica* **87**(3), 1021–1054.

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.

Bartlett, P. L. & Maass, W. (2003), 'Vapnik-Chervonenkis dimension of neural nets', *The handbook of brain theory and neural networks* pp. 1188–1192.

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.

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., Lamadon, T. & Manresa, E. (2017), Discretizing unobserved heterogeneity. University of Chicago, Becker Friedman Institute for Economics Working Paper.

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.

Cheng, J. & Small, D. S. (2006), 'Bounds on causal effects in three-arm trials with non-compliance', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(5), 815–836.

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.

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*a*), Essays in nonparametric econometrics, PhD thesis, Brown University.

Gunsilius, F. (2019*b*), Testability of instrument validity under continuous treatments. unpublished manuscript.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics, Springer.

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. 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.

Karpinski, M. & Macintyre, A. (1997), 'Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks', *Journal of Computer and System Sciences* **54**(1), 169–176.

Kitagawa, T. (2009), Identification region of the potential outcome distributions under instrument independence. Cemmap Working paper.

Kitagawa, T. (2015), 'A test for instrument validity', *Econometrica* **83**(5), 2043–2063.

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

Kleibergen, F. (2002), 'Pivotal statistics for testing structural parameters in instrumental variables regression', *Econometrica* **70**(5), 1781–1803.

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.

Laffértes, 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.

Masten, M. A. & Poirier, A. (2018), 'Identification of treatment effects under conditional partial independence', *Econometrica* **86**(1), 317–351.

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), 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.

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. University of Toronto Working paper.

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

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. University of Iowa working paper.

Stock, J. H., Wright, J. H. & Yogo, M. (2002), 'A survey of weak instruments and weak identification in generalized method of moments', *Journal of Business & Economic Statistics* **20**(4), 518–529.

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

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.

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 omitted from the main text

A.1 Proof of Proposition 1

Proof. The construction of the counterfactual processes $Y_X(w)$ and $X_Z(w)$ follows from the generalization of Kolmogorov's extension theorem to random domains established in Theorem 1' of

Hu (1988). Focus on $X_Z(w)$ as the construction for $Y_X(w)$ is analogous. Let $P_{X(Z)}$ be the counterfactual conditional measure of X given exogenously fixed Z with *counterfactual* measure P_Z^*. Together they induce a counterfactual bimeasure (Kluvánek 1981, p. 234)

$$P_{(X,Z)}(A_x, A_z) = \int_{A_z} P_{X(Z=z)}(A_x) P_Z^*(dz). \tag{11}$$

Now consider the Borel σ-algebra on $\mathcal{Z} \subset \mathbb{R}^d$, which is induced by the semi-ring $\mathscr{S} \coloneqq \{(a, b] \subset \mathbb{R}^d : -\infty \leq a \leq b \leq +\infty\}$ of all rectangles which are open on the left and closed on the right. This σ-algebra satisfies condition (C1) in Hu (1988). To see this, let $F, F' \subset \mathbb{R}^d$ be finite subsets. Moreover, define by $\mathscr{D}_F, \mathscr{D}_{F'}$ the σ-algebras generated by all sets $D_F, D_{F'} \in \mathscr{S}$ such that $F \subset D_F$ and $F' \subset D_{F'}$. Then it follows that $D_F \cap D_{F'} \in \mathscr{D}_F \cap \mathscr{D}_{F'}$ for all those sets, because the respective σ-algebras are generated by the semi-ring $\mathscr{S}$.

Now focus on the other conditions of Theorem 1' in Hu (1988). Fix a finite subset $F \subset \mathbb{R}^d$ again. We define the bimeasure $P_{(X,Z),F}$ as

$$P_{(X,Z),F}(A_x, D) \coloneqq P_{(X,Z)}(A_x, D_F)\big|_{\mathscr{D}_F} \qquad \text{for } A_x \subset \mathscr{B}_X \text{ and } D_F \in \mathscr{D}_F,$$

where $b(A_x, \cdot)|_{\mathscr{D}_F}$ defines the restriction of the bimeasure $b(A_x, \cdot)$ to the σ-algebra $\mathscr{D}_F$. Then the set of bimeasures $\{P_{(X,Z),F} : F \subset \mathbb{R}^d \text{ finite}\}$ straightforwardly satisfies the consistency condition required in Theorem 1', because each σ-algebra $\mathscr{D}_F$ is a sub-σ-algebra of $\mathscr{B}_Z$ and $P_{(X,Z)}$ is a probability bimeasure defined on $\mathscr{B}_X \otimes \mathscr{B}_Z$. The conclusion of Theorem 1' in Hu (1988) now shows the existence of the stochastic process X_Z with random time domain Z so that $P(X_Z \in A) = \int P_{X(Z=z)}(A_z) P_Z^*(dz)$ for events $A \coloneqq \prod_{z \in \mathcal{Z}} A_z$.

Now consider the process $Y_X(v)$ for which a similar reasoning holds. In this case the bimeasure needed for Theorem 1' in Hu (1988) is

$$P_{(Y,X)}(A_y, A_x) = \int_{A_x} P_{Y(X=x)}(A_y) P_X^*(dx), \tag{12}$$

where P_X^* is *not* the observable measure, but a counterfactual one. Here again, this bimeasure satisfies the requirements if we define the same σ-algebra on $\mathcal{X}$ as we did before on $\mathcal{Z}$. Together, the two processes generate the joint counterfactual process

$$P_{(Y,X,Z)}(A_y, A_x, A_z) = \int_{A_x} P_{Y(X=x)}(A_y) P_{(X,Z)}(dx, A_z) = \int_{A_z} \int_{A_x} P_{Y(X=x)}(A_y) P_{X(Z=z)}(dx) P_Z^*(dz), \tag{13}$$

which follows from the exclusion restriction: $P_{Y(X)}$ does not depend on Z. The independence restriction $Z \perp\!\!\!\perp W$ implies that $P_Z^* = P_Z$. Therefore, one can test the validity of the instrument Z by comparing the properties of the stochastic process $[Y, X]_Z$ induced by the observable distribution $P_{Y,X,Z}$ to the stochastic process $[Y, X]_Z^*(w)$ corresponding to $P_{(Y,X,Z)}$. □

A.2 Proof of Theorem 1

Proof. It follows from the Kolmogorov-Chentsov theorem (Karatzas & Shreve 1998, Theorem 2.2.8) in conjunction with Theorem 2.4.10 in Karatzas & Shreve (1998) that $\mathscr{P}^*(\mathcal{W})$ is relatively compact in the weak topology. In fact, $\mathscr{P}^*(\mathcal{W})$ is actually compact in the weak topology since the constants K_y and K_x in Assumption 3 are fixed for all $P_W \in \mathscr{P}^*(\mathcal{W})$.[43] The kernel $K(Y_X(w), A_y, A_x)$ in the objective function is itself either bounded and continuous for fixed A_y, A_x or takes the form of an indicator function by Assumption 4, in which case it can be arbitrarily well approximated by smooth functions. Therefore, if there is a sequence $\{P_W^n\}_{n\in\mathbb{N}} \in \mathscr{P}^*(\mathcal{W})$ converging weakly to some $P_W \in \mathscr{P}^*(\mathcal{W})$, then it holds that

$$\int K(Y_X(w), A_y, A_x) P_W^n(dw) \to \int K(Y_X(w), A_y, A_x) P_W(dw)$$

by the definition of weak convergence of measures and the continuity of either K or its smooth approximation. This shows that the maximum and minimum of this optimization problem are well defined for these K in question. The fact that these optimization problems provide bounds on the counterfactual laws $P_{(Y,X)}$ and $P_{(X,Z)}$ follows from Proposition 1 and the argumentation before Theorem 1. □

A.3 Proof of Theorem 2

The proof is split into two lemmas and the main proof.

Lemma 1 (VC dimension of $S(Y_X(w), y, \eta)$ and $\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta)$)**.**
$d_{VC}(S(Y_X(w), y, \eta)) < +\infty$ *and* $d_{VC}(\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta) < +\infty) < +\infty$.

Proof. The idea lies in applying Theorem 8 in Bartlett & Maass (2003), which is a more user-friendly form of Theorem 2 in Karpinski & Macintyre (1997). First, since $S(Y_X(w), y, \eta)$ for given y and η is a real valued function, one can obtain its VC dimension by considering its *complete set of indicators* $\mathbb{1}_{[0,+\infty)}\{S(Y_X(w), y, \eta) - b\}$, where $b \in [0, 1]$ is a free parameter (Vapnik 1998, Section 5.2.3). Call this set of indicator functions $\mathcal{F} := \left\{w \mapsto f(\theta, w) : \theta \in \mathbb{R}^3\right\}$, where $\theta := (y, \eta, \beta)$ and $f(\theta, w) := \mathbb{1}_{[0,+\infty)}\{S(Y_X(w), y, \eta) - b\}$. The functional form of $S(Y_X(w), y, \eta)$ consists of only the arithmetic operations $+$, $-$,$\cdot$, $/$, the function $\exp(\cdot)$, as well as a jump conditional on $\geq$, defined over the set of real numbers if one considers the term $\eta^{-\frac{1}{2}}$ to be another parameter. In particular, an algorithm for an input w—which in turn defines the path $Y_X(w)$—only needs 16 steps to calculate the binary output of the function $\mathbb{1}_{[0,+\infty)}\{S(Y_X(w), y, \eta) - b\}$, in at most 4 parameters, if one considers $\eta^{-\frac{1}{2}}$ to be another parameter. Then Theorem 8 in Bartlett & Maass

[43] A sequence of measures $(P_n)_{n\in\mathbb{N}}$ converges weakly to a measure P if and only if

$$\int f(x) P_n(dx) \to \int f(x) P(dx)$$

for every bounded and continuous real function f on $\mathcal{X}$ (Karatzas & Shreve 1998, Definition 2.4.1).

(2003) implies the finiteness of $d_{VC}(\mathcal{F})$. By the definition of the VC dimension of real-valued functions (Vapnik 1998, p. 191), it follows that $d_{VC}(S(Y_X(w), y, \eta)) < +\infty$.

The same argument holds for the set of functions $\mathbb{1}_{[0,+\infty)}\{\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta) - b\}$, where $\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta)$ is a product of $S(Y_X(w), y, \eta)$, $S(X_Z(w), x, \eta)$, and $S(Z(w), z, \eta)$, which all need a finite number of computational steps as defined above. The computation of $\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta)$ then requires two more multiplications, so that the algorithm for $\mathbb{1}_{[0,+\infty)}\{\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta) - b\}$ also concludes in finitely many steps for finitely many parameters. Again, Theorem 8 in Bartlett & Maass (2003) now implies $d_{VC}(\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta)) < +\infty$. □

Lemma 2 (Approximation of the constraint)**.** *Under Assumptions 1 - 3, and 5*

$$\left\| F_{Y,X,Z}(y,x,z) - \frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)}$$
$$\leq c(\eta) + C(\eta)^2 2^{2(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}} + \frac{(C_{RN} - c_{RN})^2}{l}\left[d_{VC}(\Theta)\left(\ln\left(\frac{2l}{d_{VC}(\Theta)} \right) + 1 \right) - \ln(\rho/4) + 1 \right] < +\infty$$

for all $l \in \mathbb{N}$. Here, $0 < c(\eta), C(\eta) < +\infty$ are constants, $\alpha, \beta, \gamma, \delta$ are defined in Assumption 3, $\ln(\cdot)$ defines the natural logarithm, κ is the respective order of the dyadic approximation of the unit interval, c_{RN} and C_{RN} are the bounds on the Radon-Nikodym derivatives defined in Assumption 5 and $d_{VC}(\Theta)$ is a shorthand for the VC dimension $d_{VC}(\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta)$ of Θ.

Proof. Using the notation from (7), one can write the constraint as

$$\left\| F_{Y,X,Z}(\cdot,\cdot,\cdot) - \frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), \cdot, \cdot, \cdot, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)}$$
$$= \left\| \int \mathbb{1}_{[0,\cdot]}(Y_X(w)) \mathbb{1}_{[0,\cdot]}(X_Z(w)) \mathbb{1}_{[0,\cdot]}(Z(w)) P_W(dw) - \frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), \cdot, \cdot, \cdot, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)},$$

where the second line follows from the population constraint in (5) and the $L^2([0,1]^3)$ is taken with respect to Lebesgue measure over y, x, and z under Assumption 2. From the triangle inequality it follows that

$$\left\| \int \mathbb{1}_{[0,\cdot]}(Y_X(w)) \mathbb{1}_{[0,\cdot]}(X_Z(w)) \mathbb{1}_{[0,\cdot]}(Z(w)) P_W(dw) - \frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), \cdot, \cdot, \cdot, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)}$$
$$\leq \left\| \int \left[\mathbb{1}_{[0,\cdot]}(Y_X(w)) \mathbb{1}_{[0,\cdot]}(X_Z(w)) \mathbb{1}_{[0,\cdot]}(Z(w)) - \Theta(Y_X(w), X_Z(w), Z(w), \cdot, \cdot, \cdot, \eta) \right] P_W(dw) \right\|^2_{L^2([0,1]^3)}$$
$$+ \left\| \int \left[\Theta(Y_X(w), X_Z(w), Z(w), \cdot, \cdot, \cdot, \eta) - \Theta(\tilde{Y}_{\tilde{X}}(w), \tilde{X}_{\tilde{Z}}(w), \tilde{Z}(w), \cdot, \cdot, \cdot, \eta) \right] P_W(dw) \right\|^2_{L^2([0,1]^3)}$$

$$+\left\|\int \Theta(\tilde{Y}_{\tilde{X}}(w),\tilde{X}_{\tilde{Z}}(w),\tilde{Z}(w),\cdot,\cdot,\cdot,\eta)P_W(dw)-\frac{1}{l}\sum_{i=1}^{l}\Theta(\tilde{Y}_{\tilde{X}}(i),\tilde{X}_{\tilde{Z}}(i),\tilde{Z}(i),\cdot,\cdot,\cdot,\eta)\frac{d\hat{P}_W}{d\hat{P}_0}(i)\right\|^2_{L^2([0,1]^3)}.$$

The first term can be made smaller than any $\epsilon > 0$ by choosing $\eta(\epsilon)$ large enough. This follows from the approximation of the indicator function by the logistic function defined above. To obtain a bound on the second term, note that by the Kolmogorov-Chentsov Theorem (Karatzas & Shreve 1998, Theorem 2.2.8) in combination with Assumption 3, the modulus of continuity of P_W-almost each path $Y_{X=x}(w)$, viewed as a function on $[0,1]$, is $\omega_{Y_{X=x}(w)}(|x-x'|) = |x-x'|^{\frac{\beta}{\alpha}}$, and the modulus of continuity of P_W-almost each path $X_{Z=z}(w)$ is $\omega_{X_{Z=z}(w)}(|z-z'|) = |z-z'|^{\frac{\delta}{\gamma}}$. Furthermore, since Θ is a logistic approximation of the indicator function, it is real analytic in $Y_X(w)$, $X_Z(w)$, and $Z(w)$ and in particular Lipschitz continuous for any $\eta < +\infty$, so that it holds for all $y,x,z \in [0,1]^3$ that

$$\begin{aligned}
&\left|\Theta(Y_{X=x}(w),X_{Z=z}(w),Z(w),y,x,z,\eta)-\Theta(\tilde{Y}_{\tilde{X}=x}(w),\tilde{X}_{\tilde{Z}=z}(w),\tilde{Z}(w),y,x,z,\eta)\right| \\
\leq& C(\eta)\left|Y_{X=x}(w)-\tilde{Y}_{\tilde{X}=x}(w)\right|\left|X_{Z=z}(w)-\tilde{Z}_{\tilde{Z}=z}(w)\right| \\
\leq& C(\eta)\omega_{Y_{X=x}(w)}(2^{-\kappa+1})\omega_{X_{Z=z}(w)}(2^{-\kappa+1}) \\
\leq& C(\eta)2^{(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}}.
\end{aligned}$$

The first inequality follows from the multivariate analogue of the mean-value theorem in conjunction with the fact that the gradient of Θ in its first three argument is bounded above by some constant $0 < C(\eta) < +\infty$ for every $\eta < +\infty$, and the fact that Z takes values in the unit interval, so that $\sup_{z,z'\in[0,1]}|z-z'| \leq 1$.[44] The second inequality follows from Theorem 1 in Anastassiou & Yu (1992*b*), which gives a bound on the wavelet approximation in terms of the modulus of continuity of the function to be approximated.[45] The final inequality follows from Assumption 3 and the dyadic approximation of order κ. Since this inequality holds for every y,x,z, it also holds for the supremum, which implies that it also holds for the $L^2([0,1]^3)$ norm by Hölder's inequality combined with the fact that the unit interval is bounded. Therefore, the second term can be made smaller than $C(\eta)2^{(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}}$.

A bound on the third term can be found via the large deviation results in Vapnik (1998, Chapter 5) and relies on a similar idea to Girosi (1995). To do so, note that under Assumption 5

$$\int \Theta(\tilde{Y}_{\tilde{X}}(w),\tilde{X}_{\tilde{Z}}(w),\tilde{Z}(w),y,x,z,\eta)P_W(dw) = \int \Theta(\tilde{Y}_{\tilde{X}}(w),\tilde{X}_{\tilde{Z}}(w),\tilde{Z}(w),y,x,z,\eta)\frac{dP_W}{dP_0}(w)P_0(dw),$$

because the Radon-Nikodym derivative $\frac{dP_W}{dP_0}$ exists and is bounded for all $P_W \in \mathscr{P}^*(\mathcal{W})$ by

[44] Note that $C(\eta) \to \infty$ as $\eta \to \infty$.

[45] When using different bases, it is here where one changes the approximative property of this basis, which is the only difference to the current proof. Everything else stays the same.

Assumption 5, so that one can focus on approximating the term on the right side. The goal is to interpret

$$R(y,x,z) \coloneqq \int \Theta(\tilde{Y}_{\tilde{X}}(w), \tilde{X}_{\tilde{Z}}(w), \tilde{Z}(w), y, x, z, \eta) \frac{dP_W}{dP_0}(w) P_0(dw)$$

as a risk functional in the sense of Vapnik (1998). The empirical risk counterpart then is

$$R_{emp}(y,x,z) \coloneqq \frac{1}{l} \sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i).$$

Now since $\frac{dP_W}{dP_0}$ is bounded by Assumption 5, Theorem 5.1 in Vapnik (1998) implies that

$$\begin{aligned} & P\left\{ \sup_{(y,x,z)\in[0,1]^3} |R(y,x,z) - R_{emp}(y,x,z)| > \tau \right\} \\ \leq & 4 \exp\left\{ H_{\text{ann}}(2l) - l \left(\frac{\tau - l^{-1}}{C_{RN} - c_{RN}} \right)^2 \right\}. \end{aligned} \tag{14}$$

Here, $H_{\text{ann}}(l)$ is the annealed entropy of the function $\Theta(\tilde{Y}_{\tilde{X}}(w), \tilde{X}_{\tilde{Z}}(w), \tilde{Z}(w), y, x, z, \eta)$, defined in Vapnik (1998, section 3.8), $0 < c_{RN} \leq C_{RN} < +\infty$ are the bounds for the Radon-Nikodym derivatives, and $\tau > 0$ is some small value.

This result reduces to a simpler form in the current setting. First, equation (5.12) in Vapnik (1998) implies

$$H_{\text{ann}}(\lambda, l) \leq d_{VC}(\Theta) \left(\ln\left(\frac{l}{d_{VC}(\Theta)} \right) + 1 \right) < +\infty,$$

where $d_{VC}(\Theta)$ is a short-hand notation for $d_{VC}(\Theta(Y_X(w), X_Z(w), Z(w), y, x, z, \eta)$. The finiteness of the VC dimension follows from Lemma 1. Second, one can rewrite the result (14) in the following way (Vapnik 1998, p. 193): with probability $1 - \rho$ the inequality

$$\|R(y,x,z) - R_{emp}(y,x,z)\|_{L^\infty([0,1]^3)} \leq (C_{RN} - c_{RN}) \sqrt{\mathcal{E}(l)}$$

holds true, where

$$\mathcal{E}(l) \coloneqq l^{-1} \left[d_{VC}(\Theta) \left(\ln\left(\frac{2l}{d_{VC}(\Theta)} \right) + 1 \right) - \ln(\rho/4) + 1 \right],$$

where $\ln(x)$ denotes the natural logarithm of x. Note finally that the $L^\infty([0,1]^3)$ bound implies a $L^2([0,1]^3)$ bound by Hölder's inequality.

Putting everything together, it therefore follows that with probability $1 - \rho$

$$\left\| F_{Y,X,Z}(y,x,z) - \frac{1}{l} \sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)}$$

$$\leq c(\eta) + C(\eta)^2 2^{2(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}} + \frac{(C_{RN} - c_{RN})^2}{l}\left[d_{VC}(\Theta)\left(\ln\left(\frac{2l}{d_{VC}(\Theta)}\right)+1\right) - \ln(\rho/4)+1\right],$$

where the first $0 < c(\eta) < +\infty$ embodies the degree of approximation of Θ to the indicator functions. □

Proof of Theorem 2. Fix A_y and A_x throughout and for some arbitrary $P_W \in \mathscr{P}^*(\mathcal{W})$ consider

$$\left\| \int K(Y_X(w), A_y, A_x) P_W(dw) - \frac{1}{l}\sum_{i=1}^{l} \tilde{K}(\tilde{Y}_{\tilde{X}}(i), A_y, A_x) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|_{L^\infty([0,1]^3)},$$

which can be bounded in exactly the same way as the constraint in the proof of Lemma 2 to give with probability $1-\rho$

$$\begin{aligned}
&\left\| \int K(Y_X(w), A_y, A_x) P_W(dw) - \frac{1}{l}\sum_{i=1}^{l} \tilde{K}(\tilde{Y}_{\tilde{X}}(i), A_y, A_x) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|_{L^\infty([0,1]^3)} \\
\leq& c'(\eta) + C'(\eta)^2 2^{2(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}} + \frac{(C_{RN} - c_{RN})^2}{l}\left[d_{VC}(K)\left(\ln\left(\frac{2l}{d_{VC}(K)}\right)+1\right) - \ln(\rho/4)+1\right],
\end{aligned}$$

where $0 \leq c'(\eta) < +\infty$, $0 < C'(\eta) < +\infty$ are constants, $d_{VC}(K)$ is the shorthand for the VC dimension $d_{VC}(K(Y_X(w), A_y, A_x))$, and where η is introduced to account for the setting where K is an indicator variable and is approximated by a logistic function of the form $S(Y_X(w), y, \eta)$. In the case where K is differentiable already, $c'(\eta) = 0$ and $C'(\eta)$ does not depend on η.

This approximation holds for every $P_W \in \mathscr{P}^*(\mathcal{W})$, and since the latter is compact in the weak topology by Theorem 1, it also holds for the (not necessarily unique) P^*_W and P_{W_*} which induce the maximum and minimum values of the programs (5). Therefore, for every $\varepsilon > 0$ there exists an $l^* \in \mathbb{N}$ such that $\max\{\mathcal{D}(l^*), \mathcal{D}'(l^*)\} < \varepsilon$, where

$$\mathcal{D}(l) \coloneqq c(\eta) + C(\eta)^2 2^{2(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}} + \frac{(C_{RN} - c_{RN})^2}{l}\left[d_{VC}(\Theta)\left(\ln\left(\frac{2l}{d_{VC}(\Theta)}\right)+1\right) - \ln(\rho/4)+1\right]$$

and

$$\mathcal{D}'(l) \coloneqq c'(\eta) + C'(\eta)^2 2^{2(-\kappa+1)\frac{\beta\gamma+\delta\alpha}{\alpha+\gamma}} + \frac{(C_{RN} - c_{RN})^2}{l}\left[d_{VC}(K)\left(\ln\left(\frac{2l}{d_{VC}(K)}\right)+1\right) - \ln(\rho/4)+1\right]$$

are the approximation bounds for the constraint and the objective function, respectively. This implies that the maximum- and minimum values V^* and V_* of (5) and the maximum and minimum values $\tilde{V}^*$ and $\tilde{V}_*$ of (7) satisfy

$$\max\{|V^* - \tilde{V}^*|, |V_* - \tilde{V}_*|\} < \varepsilon.$$

□

A.4 Proof of Proposition 2

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

$$\tilde{\mathcal{C}}(F_{Y,X,Z}, \varepsilon) \coloneqq$$
$$\left\{ \hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W}) : \left\| F_{Y,X,Z}(y,x,z) - \frac{1}{l} \sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)} \leq \varepsilon \right\}$$

where $\mathcal{F}_{Y,X,Z}$ denotes the set of all cumulative distribution functions on $[0,1]^3$ which is equipped with the uniform norm $\|\cdot\|_{L^\infty([0,1]^3)}$. Note that under Assumption 5 $\hat{\mathscr{P}}^*(\mathcal{W})$ is compact in the weak topology since the Radon-Nikodym derivatives are uniformly bounded by $0 < c_{RN} \leq C_{RN} < +\infty$. 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 5 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. 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 $\{\hat{F}_{Y,X,Z;n}\}_{n\in\mathbb{N}} \in \mathcal{F}_{Y,X,Z}$ be a sequence such that $\|\hat{F}_{Y,X,Z;n} - F_{Y,X,Z}\|_{L^\infty([0,1]^3)} \to 0$ as $n \to \infty$ and $\{\hat{P}_{W;n}\}_{n\in\mathbb{N}} \in \hat{\mathscr{P}}^*(\mathcal{W})$ be a sequence satisfying $\hat{P}_{W;n} \in \tilde{\mathcal{C}}(F_{Y,X,Z;n}, \varepsilon)$. Since all functions involved in the inequality are continuous, and the inequality in the constraint correspondence is weak, 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 $\hat{P}_W \in \tilde{\mathcal{C}}(F_{Y,X,Z}, \varepsilon)$. Now note that $\hat{\mathscr{P}}^*(\mathcal{W})$ is convex under Assumption 5, because a convex combination of these two measures is always at most as large as C_{RN} and at least as small as c_{RN}. Based on this, the constraint correspondence is easily seen to be convex valued by the triangle inequality. Now under Assumption 7, there exists a neighborhood $\mathcal{U}$ 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 $\tilde{\mathcal{C}}(F_{Y,X,Z}, \varepsilon)$. By continuity of the $L^2([0,1]^3)$-norm and the weak inequality of the constraint 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. Moreover, since the constraint correspondence is also convex-valued, any convex combination between them is a solution as well, showing lower hemicontinuity.

The objective function defined by the kernel K is continuous under Assumption 6 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 $\tilde{V}^*(F_{Y,X,Z})$ and $\tilde{V}_*(F_{Y,X,Z})$ are

continuous. Now under the Glivenko-Cantelli Theorem (van der Vaart 2000, Theorem 19.1), it holds that $\|\hat{F}_{Y,X,Z;n} - F_{Y,X,Z}\|_{L^\infty([0,1]^3)} \to 0$ as $n \to \infty$ almost surely. Now by the continuity of the value functions, one can apply the Continuous Mapping Theorem (van der Vaart & Wellner 2013, Theorem 1.9.5), which implies that the value functions converge almost surely. □

A.5 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 $\hat{\check{V}}_*(F_{Y,X,Z})$ and $\hat{\check{V}}^*(F_{Y,X,Z})$. Focus on the minimization problem in (7), as maximization is perfectly analogous. It is convenient to write (7) in terms of the notation in Bonnans & Shapiro (2013):

$$\begin{aligned} &\min_{\hat{P}_W \in \hat{\mathscr{P}}^*(\mathcal{W})} \quad K(\hat{P}_W) \\ &\text{subject to} \qquad G(\hat{P}_W, F_{Y,X,Z}) \in [0, \varepsilon], \end{aligned} \tag{15}$$

where

$$G(\hat{P}_W, F_{Y,X,Z}) := \left\| F_{Y,X,Z} - \frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), \cdot, \cdot, \cdot, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i) \right\|^2_{L^2([0,1]^3)}.$$

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

$$0 \in \text{int}\{G(\hat{P}_W, F_{Y,X,Z}) + \delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X,Z})\}, \tag{16}$$

where $F_{Y,X,Z}$ is the data-generating process in the population, $\hat{P}_W$ is an argument which minimizes (15), $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X,Z})$ 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 $[0,\varepsilon]$, (16) reduces to showing that $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X,Z})$ 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 $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X,Z})$ as[46]

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

[46] In the following derivation, the shorthand notations $\frac{1}{l}\sum_{i=1}^{l} \Theta \frac{d\hat{P}_W}{d\hat{P}_0}$, $\|\cdot\|_{L^2}$ and $\int F_{Y,X,Z} dydxdz$ are used to denote $\frac{1}{l}\sum_{i=1}^{l} \Theta(\tilde{Y}_{\tilde{X}}(i), \tilde{X}_{\tilde{Z}}(i), \tilde{Z}(i), y, x, z, \eta) \frac{d\hat{P}_W}{d\hat{P}_0}(i)$, $\|\cdot\|_{L^2([0,1]^3)}$, and $\int_{[0,1]^3} F_{Y,X,Z}(y,x,z) dydxdz$, respectively.

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

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

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

Since $F_{Y,X,Z}$, $\frac{t}{l} \sum_{i=1}^{l} \Theta \frac{d\hat{P}_1}{d\hat{P}_0}$, and $\frac{t}{l} \sum_{i=1}^{l} \Theta \frac{d\hat{P}_W}{d\hat{P}_0}$ take values in the unit interval, Lebesgue's Dominated Convergence Theorem permits interchanging the limit and the integral, which after simplification gives

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

To see that (16) can be satisfied, ignore the "weight term" $2\frac{1}{l} \sum_{i=1}^{l} \Theta \frac{d\hat{P}_1}{d\hat{P}_0}$ in the last expression for a second and note that the term in brackets is the constraint in (7), which can be rewritten as $F_{Y,X,Z} - \hat{P}_W(\tilde{Y}_{\tilde{X}} \in [0, y], \tilde{X}_{\tilde{Z}} \in [0, x], \tilde{Z} \in [0, z])$. Therefore, without the "weight term", which is always non-negative, the expression for $\delta G_{\hat{P}_W}(\hat{\mathscr{P}}^*(\mathcal{W}), F_{Y,X,Z})$ takes the form of a comparison between these two probabilities in terms of second-order stochastic dominance. Therefore, it is clear to see that if either the cumulative distribution function induced by the probability $\hat{P}_W(\tilde{Y}_{\tilde{X}} \in [0, y], \tilde{X}_{\tilde{Z}} \in [0, x], \tilde{Z} \in [0, z])$ second-order stochastically dominates $F_{Y,X,Z}$ 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 of course happen in programs (7), which is why the constraint was implemented in the first place. Therefore, the restricted version of Robinson's constraint qualification is satisfied by the program (7).

Furthermore, note that under Assumption 7 the program (7) admits an optimal solution for $F_{Y,X,Z} + tF'_{Y,X,Z}$ where $F_{Y,X,Z}, F'_{Y,X,Z} \in \mathcal{F}_{Y,X,Z}$ when t is small enough such that $F'_{Y,X,Z} \in \mathcal{U} \subset \mathcal{F}_{Y,X,Z}$, 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 $\hat{\tilde{V}}_*(F_{Y,X,Z})$ is directionally Hadamard differentiable in every direction $F'_{Y,X,Z} \in \mathcal{F}_{Y,X,Z}$ and that the Hadamard derivative $d\hat{\tilde{V}}_{*,F_{Y,X,Z}}(F'_{Y,X,Z})$ takes the

form

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

where $L(\cdot,\cdot,\cdot)$ denotes the Lagrangian of the program (7), $\lambda(\hat{P}_1)$ denotes the respective Lagrange multiplier (which is unique for given $\hat{P}_1$ by Proposition 4.47 in Bonnans & Shapiro 2013), and $\delta_{F_{Y,X,Z}}$ is the directional derivative of L in its third argument in direction $F'_{Y,X,Z}$ at $F_{Y,X,Z}$.

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

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

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

$$\sqrt{n}(\hat{\tilde{V}}_*(\hat{F}_{Y,X,Z;n}) - \hat{\tilde{V}}_*(F_{Y,X,Z})) \rightsquigarrow d\hat{\tilde{V}}_{*,F_{Y,X,Z}}(\mathbb{G}_{F_{Y,X,Z}}).$$

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